TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Bioinformatik

Applying Data Integration And Knowledge Management Techniques To Analyze Systems Biology Data

Venkata Pardhasaradhi Satagopam

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Martin Bichler

Prüfer der Dissertation:

1. Prof. Dr. Burkhard Rost
2. Prof. Dr. Miguel Andrade, Johannes Gutenberg-Universität Mainz

Die Dissertation wurde am 11.01.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 23.04.2018 angenommen.

I would like to dedicate this thesis to my parents, teachers and professors for their teachings, continuous support and encouragement.

Acknowledgements

First and foremost, I owe my sincere gratitude to my supervisor Reinhard Schneider. It has been a great opportunity for me to work under his guidance. His valuable advices and constructive discussions have helped me grow professionally as well as personally. Reinhard has afforded me freedom and flexibility in my work and his continuous support and trust in me fuelled my exploration of novel research projects and make new scientific collaborations. Reinhard's guidance helped me in all the times of research and writing of this thesis.

It has been a great honour for me to be a part of Rostlab at TUM. I am very grateful to Burkhard Rost for guiding my doctoral research and giving me the opportunity to acquire the doctorate. Burkhard is a visionary scientist, his advices and continues support helped me throughout my dissertation work.

A part of this work would not have been possible without the support from our collaborators. Therefore, I would like to thank all collaborators for their trust in me, especially Karima Djabali and Jackleen Marji in Progeria project; David C Rubinsztein, Andrea Caricasole, Christian Blaschke, and Luis Gmez in TAMAHUD project; Stavros J. Hamodrakas, Margarita C. Theodoropoulou, and Nikolaos C. Papandreou in Human-gpDB project; John Briggs, Norman Davey, and Kevin O'Brien in HIV mutation browser project; Kiran Kumar Bali, and Rohini Kuner in inflammatory, neuropathic, cancer pain projects; Jean-Karim Heriché, and Jan Ellenberg in MitoCheck project; Samik Ghosh, and Hiroaki Kitano in Garuda and PD-map projects.

It has been a great pleasure for me to be a member of the Schneider group starting at European Molecular Biology Laboratory (EMBL) in Heidelberg and then moving to the Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg in Luxembourg. I would like to thank my friends and former colleagues at EMBL, Maria Secrier, Evangelos Pafilis, Bettina Roth, Georgios Pavlopoulos, Theodoros Soldatos, Salvador Santiago, Sean O'Donoghue,

Mechthilde Falkenhahn, Heiko Horn, Carlos Villacorta-Martin, Sven Haag and Afshin Khan for all the discussions, exchange of ideas, the work we did and the fun we had together. I also thank Andrés Lindau, Michael Wahlers and Rupert Lück for their IT support.

I would to thank Rostlab members at TUM, Tatyana Goldberg, Andrea Schafferhans, Guy Yachdav, Yana Bromberg, Edda Kloppmann, Maximilian Hecht, Juan Miguel Cejuela, Timothy Karl, Jonas Reeb, Esmeralda Vicedo, Laszlo Kajan, Marco Punta and Lothar Richter for scientific discussions. I am indebted to Marlena Drabik, Inga Weise from Rostlab, Manuela Fischer from the dean's office and Ute Stinzel from doctoral office for their help and continuous guidance with the administrative matters.

I would like to thank my colleagues at LCSB, Wei Gu, Christophe Trefois, Adriano Barbosa da Silva, Marek Ostaszewski, Roland Krause, Valentin Groues, Yohan Jarosz, Sascha Herzinger, Kirsten Roomp, Marie-Laure Magnani, Peter Banda, Patrick May, Piotr Gawron, Maria Biryukov, Enrico Glaab, Stephan Gebel, Sarah Peter, Lars Geffers, Jane Murray, Cyrille Thinnes, Françoise Meisch, Aishwarya Alex Namasivayam, Dheeraj Reddy Bobbili, Kavita Rege, Vasco Verissimo, Ganna Androsova, Maharshi Vyas, Noua Toukourou, David Hoksza, Dietlind Gerloff, Pinar Alper, Joachim Kutzera and Jacek Lebioda for our many insightful scientific discussions and great times we had together. I also thank my colleagues at ITTM S.A. Andreas Kremer, Ayoung Oh, Nils Christian, Serge Eifes and Yana Malevanova for their support.

I am very thankful to my present and past professional mentors Rudi Balling, Regina Becker, Rejko Krüger, George Casari, David Jackson, Martin Stein, Günther Kurapkat and mentor in my personal life Wilfried Günzl for their visionary discussions, advices and encouragement.

I am grateful to my family and friends, especially my parents, my brothers, sister and my in-laws for their support and encouragement throughout my research and my life in general. I am thankful to my two wonderful boys Srihari and Sriram; and my dear wife Gurpreet for their love, support, motivation as well as being a source of inspiration and strength to me to pursue my doctoral research.

Summary

The complexity of biological systems has fascinated researchers for centuries, and grasping the details of such complexity in the context of basic biological processes as well as diseases continues to be a challenging endeavour. Large-scale functional high-throughput datasets enable the study of these processes on a systems-wide level. Deciphering the pathways and networks of biological functions, these methods can help to improve the understanding of diseases whose etiology cannot be linked to defects in just a single gene or protein. There is a wealth of biological and medical data on different scales available from the molecular to the physiological level. In recent years, we are witnessing breakthrough improvements of bio-molecular knowledge, and technologies have been developed to investigate various aspects of cellular processes, such as genomics, proteomics, imaging; kinetic (mobile, sensor) and clinical data. These high-throughput experiments are producing an explosion of different types of data and vast amounts of literature. However, these data resources are heterogenous in both content and format. The challenge is to transform these data into knowledge. The datasets need to be harmonised, integrated and analyzed in order to facilitate the data driven hypothesis generation and validation.

Bioinformatics approaches are key to create knowledge from biomedical data. In this thesis, the potential of data integration and knowledge management methods to improve the understanding of biological systems and to give input for translational medicine applications is demonstrated. In three different areas, from experimental model systems to viral diseases and clinical data sets, the benefit of different data driven methodologies is shown.

The first part of my thesis is focused on building a fine grained integrated database with intuitive web application, 'bioCompendium'. It facilitates the annotation of experimental results originating from different model organisms. Furthermore by using bioCompendium, one can analyze/compare/enrich one or more experimental datasets from different model organisms. It provides a wide spectrum of biological information including bio-annotations, transcription factor binding site profiles, diseases associations, interacting drugs and chemicals, patents information, different clustering and enrichment methods, and visualization tools. Another important web application, 'Human-gpDB', a database of G

protein - coupled receptors(GPCRs), G-proteins, effectors and their interactions was developed. This resource, consisting of drug targets, is of high interest for both the pharmaceutical industry and academia.

The second part is focused on text-mining of available Human Immunodeficiency Virus (HIV) full text articles to extract the mutations and map them on the HIV proteome. This knowledgebase is available as a web application - HIV Mutation Browser and it is a valuable addition to the currently available HIV resources that will allow researchers to quickly and intuitively access data on mutagenesis and phenotypic variation. The database will aid the process of experimental design and be a key resource for the HIV community.

The last part is focused on clinical and translational medicine data integration and visualization. A workflow for the analysis and interpretation of high-throughput translational medicine data was created, in which visualization is an important component at each step of data processing and exploration. In this workflow, three Web services - tranSMART (clinical and translational data integration and analysis platform), a Galaxy Server (workflow management system), and a MINERVA platform (disease maps visualization service) - are combined into one big data pipeline, TGM (tranSMART - Galaxy - MINERVA pipeline) to enable the interpretation of data and the derivation of new hypotheses.

Overall, my work demonstrates how data from experiments and the clinic together with prior knowledge can be utilised for hypothesis generation and validation to provide input into personalised medicine.

Zusammenfassung

Die im Hochdurchsatz generierten Datensätze in den Lebenswissenschaften können einen systemweiten Einblick in die zugrundeliegenden faszinierenden komplexen biologischen Prozesse auf Systemebene geben. Durch die Erforschung der molekularen Zusammenhänge auf der Ebene von Signalwegen und Netzwerken kann das Verständnis von Erkrankungen verbessert werden, bei denen die Ätiologie nicht auf einzelne Defekte in Genen oder Proteinen zurückgeführt werden kann.

Es gibt mannigfaltige biologische und medizinische Daten auf verschiedenen Ebenen vom molekularen bis zum physiologischen Bereich. In den letzten Jahren wurden enorme Fortschritte in biomolekularem Wissen erzielt und Technologien entwickelt, die es erlauben zelluläre Prozesse auf molekularer Ebene zu untersuchen wie zum Beispiel die Genom- und Proteinanalyse, aber auch die Bildgebung trägt wesentlich zum Fortschritt bei. Dies wird komplementiert durch klinische Daten und kinetische Daten, die mit Hilfe von mobilen Sensoren direkt von Patienten gewonnen werden. Diese Methoden führen zu einer Explosion von Daten verschiedenen Typs sowie zu einem enormen Zuwachs an Literatur. Diese Daten sind von unterschiedlichem Format und Inhalt. Eine der größten Herausforderungen in der Biomedizin ist es, aus diesen Daten Wissen zu generieren. Die Datensätze müssen harmonisiert, integriert und analysiert werden, um datengetriebene Hypothesengenerierung und Validierung zu ermöglichen.

Die Methoden der Bioinformatik sind essentiell um Wissen aus biomedizinischen Daten zu erlangen. In dieser Doktorarbeit wird gezeigt, wie Datenintegrations- und Wissensmanagementmethoden das Verständnis von biologischen Systemen verbessern können und dadurch Einblick geben in translationale medizinische Anwendungen. Der Nutzen von datengetriebenen Methoden wird in drei verschiedenen Feldern von experimentellen Modellsystemen zu viralen Erkrankungen und klinischen Datensätzen demonstriert.

Der erste Teil meiner Arbeit befasst sich mit "bioCompendium", der Erstellung einer detaillierten integrierten Datenbank mit einem intuitiven Internetportal. BioCompendium ermöglicht eine vergleichende Analyse experimenteller Ergebnisse von verschiedenen Tiermodellen. Die Nutzung erlaubt ferner, eine Anreicherung und Annotation der Datensätze. Es liefert ein großes Spektrum biolo-

gischer Informationen; hierzu gehören Bio-Annotationen, Profile der Bindungsstellen für Transkriptionsfaktoren, Assoziierung mit Erkrankungen, wechselwirkende Medikamente und chemische Substanzen, Patentinformation, verschiedene Clustering- und Anreicherungsmethoden und Visualiserungswerkzeuge. Eine weitere wichtige Anwendung ist die Datenbank Human-gpDB zu G-Protein gekoppelte Rezeptoren (GPCRs), G-Proteine, molekulare Effektoren und ihre Wechselwirkung. Durch den Bezug zu Drug Targets ist diese Ressource von großem Interesse sowohl für die Pharmazeutische Industrie als auch die öffentliche Forschung.

Der zweite Teil der Arbeit beschäftigt sich dem Text-Mining von im Volltext verfügbaren Artikeln zum humanen Immundefekt-Virus (HIV), um Mutationen auf dem HIV Proteom zu kartieren. Diese Wissensdatenbank ist als Internetanwendung verfügbar - der HIV Mutation Browser- und ist eine wertvolle Ergänzung zu den derzeit bestehenden HIV Informationsquellen, da es den Forschern einen schnellen und intuitiven Datenzugriff auf Mutagenese und phänotypische Unterschiede. Die Datenbank wird hilfreich sein für das experimentelle Design und zu einer wichtigen Ressource für die HIV Forschung werden.

Der letzte Teil der Arbeit konzentriert sich auf klinische und translationale medizinische Datenintegration und -visualisierung. Ein Ablauf für die Analyse und Interpretation von Hochdurchsatzdaten der translationalen Medizin wurde erstellt, in dem die Visualisierung ein wichtiges Element in jedem Schritt der Datenverarbeitung darstellt. In diesem Arbeitsablauf sind 3 verschiedene Webdienste - tranSMART (eine Datenintegrations- und Analyseplattform für klinische und translationale medizinische Daten), ein Galaxy Server (Managementsystem für IT Arbeitsabläufe), und die MINERVA Plattform (Visualisierung von molekularer Information zu Erkrankung als Karte) - zu einer großen Pipeline zusammengeführt, TGM (tranSMART - Galaxy - MINERVA Pipeline), um die Interpretation der Daten und das Ableiten neuer Hypothesen zu ermöglichen.

Insgesamt zeigt meine Doktorarbeit, wie Daten aus Experimenten und dem klinischen Umfeld mit bestehendem Wissen zur Hypothesenfindung und -validierung genutzt werden kann, um Input für eine personalisierte Medizin zu geben.

Contents

Chapter 1

Introduction

In this post-genomics era, advancements in high-throughput technologies made possible the paradigm shift from 'one gene - one postdoc' to 'genome-wide functional analysis'. High-throughput methodologies such as genomics, transcriptomics, proteomics, metabolomics, lipidomics, epigenomics, RNA interference (RNAi) facilitating investigations of cellular processes in both healthy and disease states at systems levels paved a path for systems biology and systems biomedicine.

These technologies are providing deeper insights into the systems biology of complex diseases for example cancer, diabetes, neurodegenerative diseases (e.g., Parkinson's disease, Alzheimer's disease) that are occurring as a result of dysregulation. Physiological functions such as cell division, apoptosis and pathophysiological functions occur as a results of interaction between various bio-entities (e.g., genes, proteins, chemicals, metabolites, regulatory elements, non-coding RNAs) in different biological processes and pathways at different levels: cell, tissue, organ and systems level. These are not local effects, but rather systemic, a network of systematically organised events taking place at the organism level. In order to study the systems biology of these events, both pathological and physiological, researchers are employing genome wide high throughput experimentation using techniques including microarrys, proteomics, metabolomics, next generation sequencing technologies (whole genome, exome, RNAseq). These high-throughput experiments are generating measurements for different sets of genes or gene products, metabolites, non-coding RNAs (for example miRNAs, siRNAs) that are playing an important role in the respective disease or healthy condition.

There is a pressing need to integrate these heterogenous experimental datatypes and prioritise the bio-entities for either early detection of the disease (as disease biomarkers) or efficient patient stratification and personalised therapies. On the other side there are large amounts of already published literature (>27millions scientific papers in PubMed until November 2017) and various publicly available biological databases that both are great sources of knowledge. This vast amount of public knowledge is helpful to annotate, validate and prioritise the bio-entries, markers, drug targets from above mentioned high throughput experimental techniques. The main problem with public data resources is, they are heterogenous in both content (e.g. genes, proteins, regulatory elements, chemicals, metabolites, diseases, ontologies, reactions, interactions, pathways, literature etc.,) and format (e.g. flat-file, XML, relational database). This public data need to be parsed and integrated in order to use this knowledge seamlessly.

The current field of systems biology increasingly depends on high-throughput experimental techniques. This trend challenges computational biologists to develop fast, accurate, and meaningful ways of analyzing this experimental data to make sense out of it. Some of these techniques are meant to study and understand physiological phenomena or a pathophysiological conditions (disease states). These experiments may result in list of genes or gene products (mRNAs, proteins, miRNAs and so on) that may play a role in that particular biological state. It is tedious and difficult for researchers to retrieve the information from several publicly available databases for each gene individually, process it and digest it in order to develop a hypothesis and/or prioritize them for further experimental validation.

Wider access of scientific data and knowledge catalyzes worldwide scientific progress. Enormous additional knowledge and insights could be extracted from existing projects if their datasets were publicly available. But often, these datasets are in silos and inaccessible for future research or data are stored in proprietary data formats that are not interoperable (Wruck et al., 2014). Therefore, governments, funding agencies such as European Commission (EC) H2020 mandate open access to all scientific publications. From 2017, research data is by default open and FAIR (Findable, Accessible, Interoperable and Re-usable), with possibilities to opt out. H2020 projects must have Data Management Plan

(DMP), it should provide information on what data the research will generate, how to ensure its curation, preservation and sustainability, what parts of that data will be open and how users can access the data(H2020, 2016). Similarly, Wellcome Trust (Welcome-Trust, 2017), NIH (NIH, 2016) embraces similar open access policy.

1.1 Data integration and knowledge management

1.1.1 Biological databases

Large number of biological databases, tools, resources, web-services are available as a result of data coming from life science experiments from different labs and research centers all over the world. This information was systematically collected and organised in a specific repository based on the nature of the data. Some examples of such databases are listed below:

- Genes/Proteins: Ensembl (Flicek et al., 2012), EMBL (Cochrane et al., 2009), GenBank (Benson et al., 2011), EntrezGene (Maglott et al., 2011), Unigene (Schuler, 1997), UniProt (Magrane and Consortium, 2011), IPI (Kersey et al., 2004), NCBI Protein (Sayers et al., 2012), RefSeq (Pruitt et al., 2012), HGNC (Seal et al., 2011), UCSC (Fujita et al., 2011), KEGG (Kanehisa et al., 2002), GeneCards (Safran et al., 2010)
- Diseases: OMIM (Hamosh et al., 2005)
- Protein structures: PDB (Berman et al., 2007), HSSP (Schneider and Sander, 1996), PSSH (Schafferhans et al., 2003), PredictProtein (Yachdav et al., 2014)
- Protein features: Pfam (Punta et al., 2012), SMART (Letunic et al., 2009), PRINTS (Attwood et al., 2012), InterPro (Hunter et al., 2012b)
- Chemicals: DrugBank (Wishart, 2008), HMDB (Wishart et al., 2009), PubChem (Bolton et al., 2008), chEBI (Degtyarenko et al., 2008), STITCH (Kuhn et al., 2008), MATADOR (Gunther et al., 2008), PDBLigand (Feng et al., 2004), chemicals from literature (AKS2)

- Pathways: KEGG (Kanehisa et al., 2012), PANTHER (Mi et al., 2005), Reactome (Croft et al., 2011)
- Ontologies: GeneOntology (GO) (Ashburner et al., 2000)
- Interactions (protein-protein, protein-chemical): STRING (Szklarczyk et al., 2011), STITCH (Kuhn et al., 2008)
- MicroRNA: miRBase (Kozomara and Griffiths-Jones, 2011)
- Literature: PubMed/MEDLINE (PubMed, 2015)

These resource are maintained and shared by organizations such as the European Bioinformatics Institute (EMBL-EBI `https://www.ebi.ac.uk`) and the National Center for Biotechnology Information (NCBI `https://www.ncbi.nlm.nih.gov`) to advance science and health by providing access to these biomedical and genomic information resources to researchers in academia and industry.

Each year the journal Nucleic Acids Research (NAR) publishes a database issue in the month of January and a web-services issue in July. The current 2017 NAR database issue, the annual collection of bioinformatic databases on various areas of molecular biology, consists of 1877 biological databases (Galperin et al., 2017) and their sub category distribution is shown in the Figure 1.1.

The Bioinformatics Links Directory from bioinformatics.ca (Brazas et al., 2010) features curated links to 176 molecular resources, 621 databases and 1548 tools. Similarly the Pathguide (`http://www.pathguide.org`, October 2017) provides the list of 688 biological pathway and molecular interaction related resources that are organised in different sub categories such as signaling pathways, metabolic pathways, pathway diagrams, transcription factors and gene regulatory networks, protein-protein interactions, protein-compound interactions, genetic interaction networks and other categories.

This vast amount of public knowledge is very useful. The main challenge with public data resources is exponential growth and these resources are heterogenous in both content and format. Most of these resources are available in the formats of flat-file (e.g., GenBank, EMBL, PDB, OMIM, UniProt, KEGG, GO), XML (e.g., NCBI EntrezGene, MEDLINE), relational database (e.g., Ensembl). Most of the systems biology data models are represented in XML based

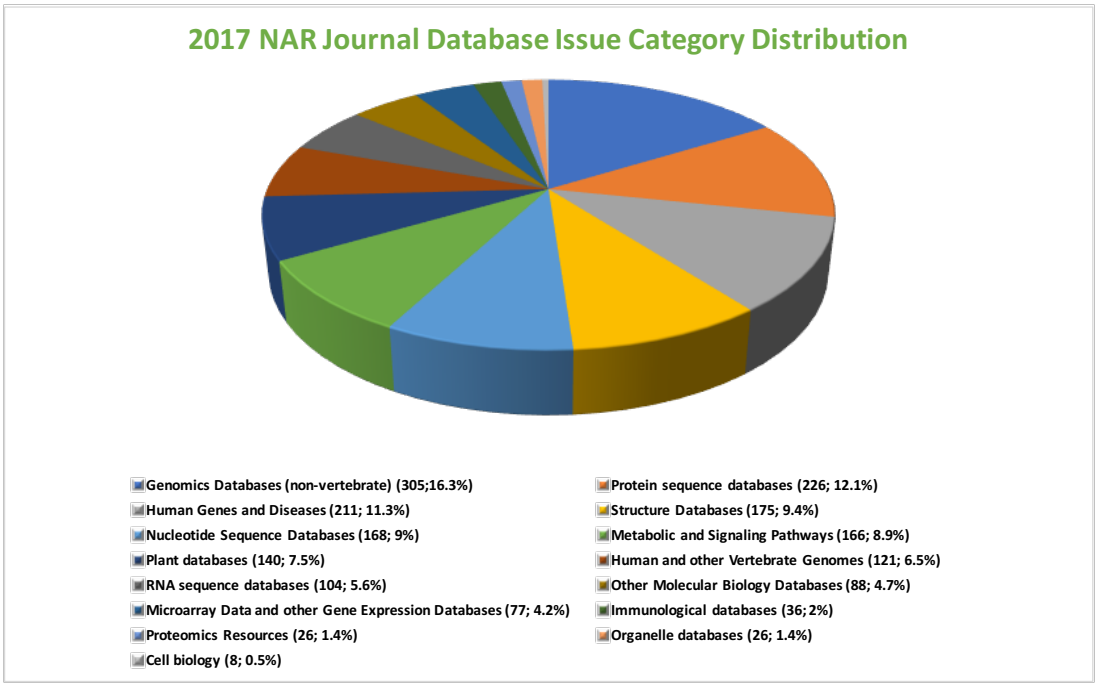


Figure 1.1: **The NAR database issue category distribution.** The figure shows the different classes of biological databases published in the 2017 Nucleic Acids Research (NAR) database issues with their absolute size and percentage of each category.

markup languages for example Systems Biology Markup Language (SBML), a standard for representing models of biochemical and gene-regulatory networks (Hucka et al., 2003; Le Novere, 2006); Cell System Markup Language (CSML) for creation of an exchange format for modeling, visualizing, and simulating biological pathways (Nagasaki et al., 2008); CellML, an XML-based language for describing and exchanging models of cellular and subcellular processes (Cuellar et al., 2015; Lloyd et al., 2004); System Biology Results Markup Language (SBRML), a markup language for associating systems biology data with models (Dada et al., 2010); Simulation Experiment Description Markup Language (SED-ML), a machine readable format for providing simulation and analysis experiments to apply to computational models. A SED-ML file includes details of which models to use, how to modify them prior to executing a simulation, which simulation and analysis procedures to apply, which results to extract and how to present them (Bergmann et al., 2015); Metabolic Flux Analysis Markup Lan-

guage (MFAML) for the representation and exchange of metabolic flux models (Yun et al., 2005); mass spectrometry standards from HUman Protein Organization Protein Standards Initiative (HUPO-PSI `http://www.psidev.info`) such as mzXML, mzData, mzML, TraML, mzIdentML, mzQuantML. This public data need to be parsed and integrated in order to use this knowledge seamlessly.

1.1.2 Data management and integration

Large systems biology, clinical and translational projects can have several work-packages and work-groups often from different parts of the world. The functionality of data management system that includes (i) data collection, (ii) storage, (iii) curation, (iv) harmonization, (v) exchange, (vi) integration, (vii) analysis and (viii) delivery of the results from the data in these research projects are very important to reach the objectives and goals of the the project. The greatest problems in data management in research are 5V's: Volume, Velocity, Variety, Value, Veracity (Fosso Wamba et al., 2015; White, 2012) as described below and the noisiness of the data generated by modern high-throughput methods, the lack of well-established data standards and globally unique identifiers, which is needed for mapping and data integration (Mayer, 2009a; Van Deun et al., 2009).

- Volume: Large volume of data that either consume huge storage or consist of large number of records (Russom, 2011)
- Variety: Data generated from greater variety of sources and formats, and contain multidimensional data fields (Russom, 2011)
- Velocity: Frequency of data generation and/or frequency of data delivery (Russom, 2011)
- Value: The extent to which big data generates economically worthy insights and or benefits through extraction and transformation (Fosso Wamba et al., 2015)
- Veracity: Inherent unpredictability of some data requires analysis of big data to gain reliable prediction (Beulke, 2011)

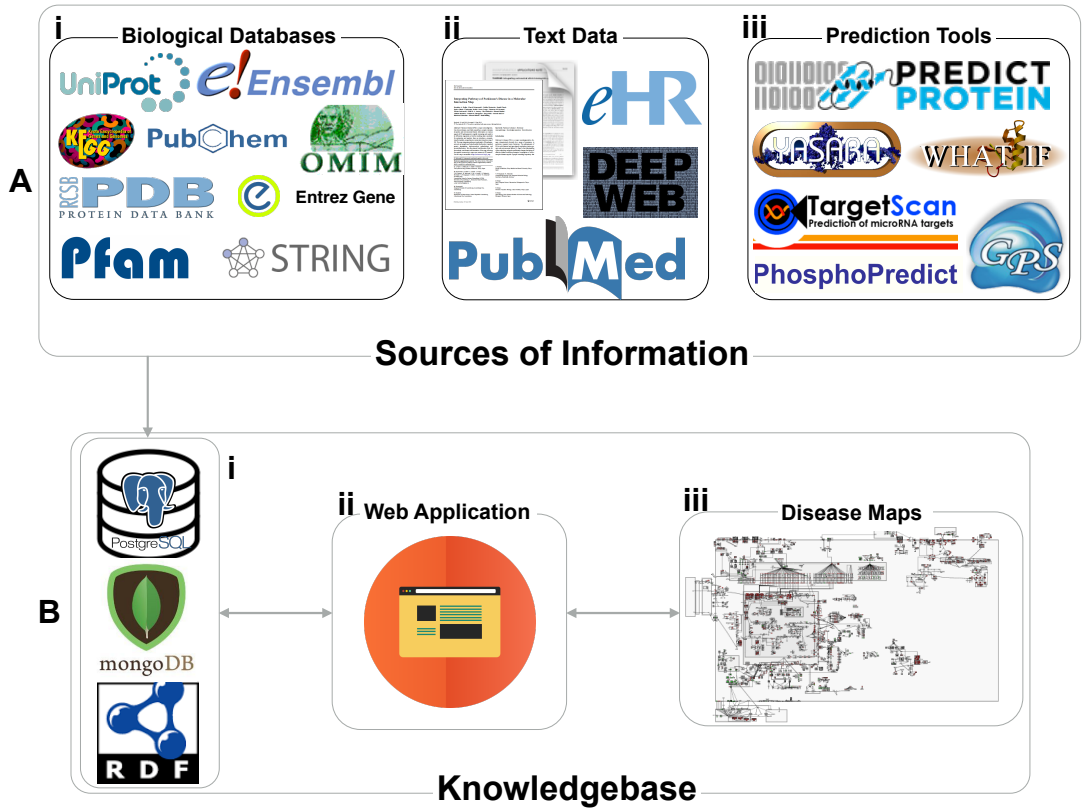


Figure 1.2: **Sources of biological information.** The figure shows (A) important sources of biological information, (A.i) several biological databases as shown in Figure 1.1 available as a result of deposition of experimental results coming from various labs from different parts of the world. Most of this data is structured (relational databases e.g., ENSEMBL, REACTOME, ChEMBL) or semi-structured (XML e.g., EntrezGene, GeneOntology; structured flat-files that needs parsing and indexing e.g., EMBL, UniProt, PDB, KEGG). (A.ii) Unstructured text data is other important type of data within the biomedical domain. For example, vast amounts of publications, conference proceedings, electronic health records, most of the current World Wide Web (WWW) content without semantics, even hidden behind HTML tags (deep web), contain large amounts of text which has been entered in a non-standardized formats. The next important source of information is (A.iii) bioinformatics prediction tools such as PredictProtein, TargetScan, PhospoPredict and so on. All these sources of biological information shown in section (A) are very vital and these are mined, analyzed, linked to fill the missing gaps and links that are important to unravel the bio molecular, cellular functions and to understand complex diseases. The extracted knowledge from these information sources are typically stored in a relational database systems such as (B.i) MySQL, PostgresSQL or NoSQL solutions like MongoDB or Resource Description Framework (RDF) triple stores (linked data). These structured knowledgebases (B) serve the fine grained knowledge to the rest of the world via web-services (B.ii) or APIs and to annotate, enrich one of the example applications such as (B.iii) disease maps (refer Chapters 6, 8 for more details).

Centralized versus federated data management: Centralized and federated data management are the two common approaches in the field of systems biology. Most of the data management solutions in systems biology are centralized as the majority of the analysis tools, algorithms that currently available are designed for pooled data. It is easier to merge the data from different resources and faster to develop new solutions. There are a lot of benefits in terms of ease to manage upgrades of systems, updates of data, its curation, application of standards and provide security as most of the data from systems biology projects are sensitive in nature. Though data update is a sisyphean task it is much easier if the data is available locally compared to remotely as in the case of federated data management. Many important resources provided by EMBL-EBI (`http://www.ebi.ac.uk`, 2017), NCBI (`https://www.ncbi.nlm.nih.gov`, 2017), Sequence Retrieval System (SRS) (Etzold and Verde, 1997), various BioMarts, clincial and translational medicine data management system - tranSMART, PredictProtein (Yachdav et al., 2014), solutions presented in this thesis for example bioCompendium, Human-gpDB, HIV Mutation Browser described in Chapters 2, 5, 7 respectively and all the databases mentioned in the section 1.1.1 of this chapter are examples of such centralized data management solutions.

Federated data management is rapidly becoming a mainstrem data management approach as we are dealing with large amounts of sensitive personalised data. Some of these data volumes are as high as from hundreds of terabytes to petabytes. For example Genomics England 100,000 Genomes Project expected to generate 15 to 20 petabytes (PB) of sequencing data and its analysis results (GenomicsEngland, 2017; Lonzer, 2014; University of Oxford, 2014). The Cancer Genome Atlas (TCGA) is providing 2.5 PB of data covering 33 different tumor types obtained from 11,000 patients (TCGA, 2017).

(Stephens et al., 2015) projected annual storage and computing needs in four domains of Big Data by 2025. In the domain of astronomy 25 zetta-bytes of data will be acquired annually, out of which 1 exabyte (EB)/year will be stored after real time processing. Where as in the case of Twitter (0.5 to 15 billions tweets/year), it demands 1 to 17 PB of storage. Metadata as well as topic and sentiment mining will be done. In the case of YouTube 500-900 millions hours data will be acquired annually, out of which 1 to 2 EB data will be stored. Then

coming to domain of our interest, genomics, will generate 1 zetta-bases per year which demands 2 to 40 EB/year. Heterogeneous data analysis and variant calling need roughly 2 trillions central processing unit (CPU) hours. In addition all-pairs genome alignments requires 10,000 trillion CPU hours. Apart from genomics, systems biology continues acquiring data from other related sub-domains - health, clinical/translational, imaging, other omics - proteomics, metabolomics, lipidomics, metagenomics, mobile sensors etc., at different scales and time points that demands far more acquisition, storage, analysis and distribution requirements than any other field in order to achieve personalised and precision medicine and treatment in the future.

Apart from these large volumes of data, the European General Data Protection Regulation (GDPR http://www.eugdpr.org) will be implemented by May 25, 2018 in each European member state, and this may influence data transfers. This might have some impact on movement of personal sensitive data from the member states, especially human sequencing data. This will depend on the subject informed consent and GDPR implementation in that particular European country. This forces us to implement federated data management solutions and take the bioinformatics analysis tools, algorithms, workflows to data and analyse the data in the federated environment rather than pooling (downloading) this massive data into one place and analysing locally which is a very expensive and time consuming effort.

The following approaches are the typical data management systems currently used in systems biology.

Spreadsheet based approaches: Most of the bench biologists follow this simplest approach to manage their experimental data using spreadsheet programs. This includes the use of Microsoft Excel and other template spread sheets like the TAB-based formats MAGE-TAB and ISA-TAB.

MAGE-TAB (Microarray Gene Expression - Tabular format) is a MIAME (Minimum Information About a Microarray Experiment) compliant, tab delimited format used to annotate microarray data and this data format is employed at the EBI repository Array-Express (https://www.ebi.ac.uk/arrayexpress/, 2017) for depositing microarray data. It uses simple spreadsheet-based for-

mat for representing primary data and associated metadata (http://fged.org/projects/mage-tab, 2017) (Rayner et al., 2006). The open-source ISA - Investigation represents the project context; Study represents a unit of research; and Assay represents an analytical measurement. The entire suite provides applications managing tab-delimited (TAB) format (ISA-TAB). It is a general purpose framework with which to collect and communicate description of the experimental metadata (for example sample details, technology and measurement types, sample to data linking) from 'omics-based' experiments so that the resulting data and discoveries are reproducible and reusable (http://isa-tools.org, 2017) (Rocca-Serra et al., 2010). The ISA software suite has been used in several projects (Sansone et al., 2012) and integrated into other data management systems, for example the Harvard stem cell discovery engine, an integrated database and analysis platform for cancer stem cell comparisons (Ho Sui et al., 2012). Another such TAB based format is Study Data Tabulation Model (SDTM) from the CDISC (Clinical Data Interchange Standards Consortium). It provides a standard for organizing and formatting data to streamline processes in collection, management, analysis and reporting. SDTM is one of the required standards for clinical study data submission to FDA (U.S.) and PMDA (Japan) (CDISC-Consortium, 2017). It's easy to use spreadsheet based approaches compared to XML based formats for example MAGE-ML. These TAB based formats are easily human-readable and can be processed by simple spreadsheet programs which are familiar to the biologists (Mayer, 2009a).

Web-based document sharing tools: In smaller projects, wiki or similar content management systems are often used for data management. There are several open source wikis (Giles, 2007; Singh et al., 2014; Zhang et al., 2014), some of them are combined with semantic web technologies (Burgstaller-Muehlbacher et al., 2016; Good et al., 2012) or groupware programs like Alfresco (Alfresco, 2017), BaseCamp (BaseCamp, 2017), EGroupware (EGroupware, 2017), BSCW - Be Smart Cooperate Worldwide (BSCW, 2017), Drupal (Drupal, 2017), or Joomla (Joomla, 2017). Solutions like Google drive (Google, 2017), Dropbox (Dropbox, 2017), ownCloud (ownCloud, 2017) are also used to exchange and access small volumes of data across the distributed project teams. Most of these so-

lutions are not based on Relational Database Management Systems (RDBMS), they can't guarantee the data consistency between the data submitted by different users as well as data security. Therefore, they are not suitable for the experimental data management (Arita, 2009; Mayer, 2009a).

Laboratory Information Management Systems (LIMS): LIMS or Electronic Lab Notebooks (ELN) are important class of laboratory data management tools, to record experimental details (meta-data), processes and results on daily basis instead of noting down in paper lab book and to link meta-data to associated high-throughout experimental results such as genomics or imaging data. Compared to paper-based lab books, these systems are searchable, and also enforce an elementary standardization of record keeping and allow the use of predesigned user-controlled templates (Mayer, 2009a). The ultimate vision of ELNs is paperless lab.

To choose a data management and integration system, there are 3 choices:

- Developing an in-house solution by using RDBMS as back end for processing and managing the data and web front-end to present the data from the database
- Use of open source software, list of open source data management systems are shown in table 1.1
- Use of commercial software, e.g. Genedata suite of Expressionist, Phylosopher, and Screener; GeneLogics, GeneBio, Accelrys Pipelien Pilot, Ingenuity Pathway Analysis Suite etc.

1.1.3 Semantic web and Linked data

Up to now I have discussed file based or relational data management and integration systems where the database schemas are specific to that application. That is why the traditional database concepts are called Closed World Assumption (CWA). But Semantic web and Linked data technologies are one step a

System	Description	Reference
AMEN	Annotation, Mapping, Expression and Network suite of tools for molecular systems biology	(Chalmel and Primig, 2008)
BASE	BioArray Software Environment to manage the transcriptomics/microarray data from Lund University, Sweden	(Vallon-Christersson et al., 2009)
BioMart	Data management system with predefined, query optimized relational schema that can be used to represent any kind of data	(Kasprzyk, 2011)
Bio-SPICE	A suite of current computational tools for biologists	(Garvey et al., 2003)
BRM	Bioinformatics Resource Manager provides the user with data management and data retrieval upon request	(Shah et al., 2007)
cBioPortal	The cBioPortal for Cancer Genomics provides visualization, analysis and download of large-scale cancer genomics data sets	(Cerami et al., 2012; Gao et al., 2013)
DIPSBC	Data Integration Platform for Systems Biology Cooperation and is a wiki based system with Solr, Lucene search functionality	(Dreher et al., 2012)
eTRIKS, tranSMART	Clinical and translational data management system	(Bierkens et al., 2015; eTRIKS consortium, 2017; Szalma et al., 2010)
i2b2	Informatics for Integrating Biology and the Bedside (i2b2)	(Murphy et al., 2009, 2010)
Gaggle	An open-source software system for integrating bioinformatics software and data sources	(Shannon et al., 2006)
ISA tools	ISA software suite consists of Java applications to manage data referring to the ISA-Tab format	(Rocca-Serra et al., 2010)
LabKey	An open source data management system focus on specimen management developed by LabKey Software	(Nelson et al., 2011)
MiMiR	An integrated platform for microarray data sharing, mining and analysis	(Tomlinson et al., 2008)
MIMAS	Data management system for multi-omics	(Gattiker et al., 2009)
OMERO	Microscopy imaging data management system	(Allan et al., 2012)
OpenClinica	Electronic clinical data capturing system. It provides tools to validate, annotate clinical data as well as to capture electronic patient reported outcomes. It also allows study audits, reporting, and data extraction	(OpenClinica, 2017; Shah et al., 2010)
openBIS	Distributed data management system for biologic information developed at the ETH Zürich.	(Bauch et al., 2011)
REDCap	Research Electronic Data Capture system (REDCap) is a secure web application for implementing and managing digital surveys and Case Report Forms (CRFs)	(Harris et al., 2009; REDCap, 2017)
SBEAMS	Systems Biology Experiment Analysis Management System form ISB, Seattle	(Marzolf et al., 2006)
SBW	Systems Biology Workbench (SBW) is a software framework for systems biology simulation and modelling	(Sauro et al., 2003)
SysMO-SEEK	Systems biology of microorganisms (SysMO) and corresponding transnational research	(Wolstencroft et al., 2011)
XperimentR	Standalone solution development at Imperial College in London by integrating BASE, Metabolomixed and OMERO for transcriptomics, metabolomics, proteomics and imaging data. It is not distributed as software	(Tomlinson et al., 2013)

Table 1.1: **Open source systems biology data management systems** - The table provides list of open source data management systems used to manage the data in the field of systems biology.

head in machine readable data and linking the data entities seamlessly. Semantic web is an extension of the web in which, not only humans, but also machines can read the information and data can be searched, found, interpreted, shared and reused among applications, organizations and communities (Sulè and Lapeyra, 2016). This is known as the Web of data. The semantic enrichment obtained with the help of ontologies to achieve true data integration, data exchange, efficient information and text mining approaches (Mayer, 2009b). There are several ontologies available in the biological domain. Currently OBO Foundry (`http://www.obofoundry.org`, 2017) lists 157 different ontologies, e.g., Gene Ontology (GO) (Lomax, 2005), Systems Biology Ontology(SBO) (Courtot et al., 2011) and Experimental Factor Ontology (EFO) (Malone et al., 2010).

Linked data is an important concept within the Semantic web and its aim is to semantically annotate the data with RDF (Resource Description Framework) model, RDFS (RDF Schema), OWL (Web Ontology Language) so that machines can browse the web. RDF is useful for describing static things or facts, because it uses a simple data model of "subject, predicate, object" triples. This RDF triple knowledge can be treated as a simple graph where subjects and objects are the nodes and edges are the predicates. New knowledge can be easily added to the existing graph, this is called OWA (Open World Assumption) (Mayer, 2009b). RDFS builds up on RDF and allows defining application-specific vocabularies by modelling of hierarchical class und subclass relationships. OWL is even more expressive than RDFS, allowing the definition of ontologies. OWL together with RDFS is suitable to model the concepts and relationships between data. One can query these triple stores with SPARQL (Simple Protocol And RDF Query Language). DBpedia, a project to convert Wikipedia articles into RDF and link to databases like GEO species, is an example of Linked data. Databases like UniProtKB, chEBI are available as RDF triple stores (Sulè and Lapeyra, 2016).

1.2 Literature mining

As shown in Figure 1.2, text is one of the important types of biomedical data. For example, vast amounts of publications, conference proceedings, medical health records. Most of the current World Wide Web (WWW) content without seman-

tics, contain large amounts of text which has been entered in non-standardized formats and different languages (Holzinger et al., 2014). However, this unstructured free text consists of very valuable information that needs mining and extraction of the knowledge. Literature mining or text mining is the use of automated methods to extract the valuable knowledge presented in the vast amounts unstructured literature. There are at least as many motivations for doing text mining work as there are types of bioscientists (Cohen and Hunter, 2008).

Text mining evolved from tagging of bio-entities by named-entity recognition methods in the text (Hirschman et al., 2002; Leaman and Gonzalez, 2008; Rebholz-Schuhmann et al., 2011; Settles, 2005), draws upon dictionaries of synonyms, homonyms, acronyms, orthographic variations, controlled terminologies, ontologies (Pafilis et al., 2013; Rzhetsky et al., 2009; Sasaki et al., 2008; Shah et al., 2009; Spasi et al., 2008) to application of co-occurrence based approaches (Hoffmann et al., 2005; Li et al., 2009a; Rosario and Hearst, 2004) and, finally, semantic and syntactic analysis that extracts complex events which seek to reveal cause-effect relation between various bio-entities involved in the biomedical processes (Björne et al., 2009; Kilicoglu and Bergler, 2009; Van Landeghem et al., 2012). Some approaches use textual data as the only source for the analysis (Gawronska et al., 2005; Peng et al., 2015) while some others combine it with experimental data available from dedicated databases (Liekens et al., 2011; Szklarczyk et al., 2011).

The literature contains rich information including bio-entities such as genes, proteins, diseases, chemicals (drugs, ligands, metabolites), cell types, tissues, organisms, biological processes, molecular functions, cellular components, ncRNAs, reactions, anatomical terms, side effects, genetic information (mutations, indels, structural variants) and so on. Genetic variations, natural selection and genetic drift are the main drivers of biological evolution. However, many mutations might be harmful (Rost, 1996; Rost et al., 2003; Sawyer et al., 2007). Most of the research results on mutations and their effects still remain published in papers. Databases such as OMIM, UniProtKB/Swiss-Prot rely on labor-intensive and time-consuming expert curation to extract knowledge from these precious experimental results available as unstructured text (Cejuela et al., 2017). There is a large information gap between literature and database annotations (Jimeno Yepes

and Verspoor, 2014). It is well described by Cejuel et al., in our publication (Cejuela et al., 2017) that PubMed search of query relevant to keywords *"mutation, mutagenesis, SNP, indel, polymorphism, sequence variant"* retrieved >1 Million articles, where similar query in UniProtKB/Swiss-Prot (Magrane and Consortium, 2011) retrieved $\sim$13,000 indexed publications. Similar trend was found with Human Immunodeficiency Virus (HIV) mutagenesis information that is available in UniProtKB/Swiss-Prot. Only 135 mutations are found in one of the highly-studied HIV group M subtype B HXB2 isolate that contains 9 proteins. But in reality there are much more mutation mentions in the literature. The clinical importance of HIV/AIDS (acquired immunodeficiency syndrome) leads to research across many diverse fields - basic research, clinical and therapeutic research. This research has produced large amounts of HIV literature, over 275,000 articles. Manual curation of this vast amount of literature is far from reality, but text mining of free literature is a solution that could scale and substantially narrow the gap (Krallinger et al., 2008).

Several resources are available with mutation data for researchers by manually curating mutagenesis and polymorphism data from HIV studies. These include the UniProt knowledgebase (UniProt-Consortium, 2014), which contains manually annotated articles describing mutagenesis of HIV proteins, the Stanford Drug Resistance database (Rhee et al., 2003), which contains curated mutations related to drug resistance and the Los Alamos HIV Database, which contains annotations from different HIV resources including epitope mutations and escape variants (`http://www.hiv.lanl.gov`). However, these resources are limited in scope because manual curation is difficult with vast amounts of available literature.

Resources such as Reflect (Pafilis et al., 2009) and MutationFinder (Caporaso et al., 2007) are available to quickly scan, tag, annotate the bio-entities computationally and organise large amounts of scientific literature systematically. These techniques should help HIV research. But in reality so few resources are available to access the literature in a structured and organised way. One such facility is PubMed Central (PMC `https://www.ncbi.nlm.nih.gov/pmc`), even though PubMed comprises more than 27 millions publications, only 4.4 millions (16% of PubMed) articles are archived in PMC and are available for scientific commu-

nity for manual reading. Many of these articles are subject to publishers access licenses and copyright restrictions and are not available for bulk downloading. Only a fraction (1.6 millions, 6%) of these publications are available for bulk retrieval and text mining. This was a major issue in mining full text articles.

Text-mining of available HIV literature and building of HIV mutation browser to serve the knowledge extracted from the literature are described in the Chapter 7.

1.3 Clinical and translational medicine

In this digital age, we are witnessing the beginning of revolutionary changes in health care and practice of medicine. Invention of internet, smart phones, world wide web(WWW), social media networking sites (Facebook, twitter, WhatsApp, WeChat, MySpace etc.,), search engines (Google, Bing, DuckDuckGo, Baidu etc.,) changed the way we live. DNA testing, sequencing and other high-throughput molecular analysis technologies paving a path for personalized medicine, which aims to be predictive, preventive, personalized and participatory (also known as P4 medicine) and establishing links between biomolecular characterizations, patient conditions, treatment effectiveness and adverse effects, and thus providing patients with the best individual treatment (Hood and Flores, 2012).

The European Society for Translational Medicine (EUSTM), defines Translational Medicine(TM) as "an interdisciplinary branch of the biomedical field supported by three main pillars: benchside, bedside and community. The goal of TM is to combine disciplines, resources, expertise, and techniques within these pillars to promote enhancements in prevention, diagnosis, and therapies" (Cohrs et al., 2015).

Translational medicine is a domain turning results of basic life science research and investigations in humans into new tools and methods in a clinical environment, for example, as new or improved diagnostics or therapies. It also includes, non-human or non-clinical studies conducted with the aim to advance therapies to the clinic or to develop basis for application of therapeutics to human diseases. This rapidly growing knowledge-based translational medicine discipline in biomedical research aims to expedite the discovery of new diagnostic tools and

treatments by using a multi-disciplinary and highly collaborative approaches. It involves the integration of multiple high dimensional datasets that capture the molecular profiles of patients, as well as detailed clinical information. To genuinely realise the promise of Big Data in healthcare, we must consistently and continuously collect the data, annotate it with consistent and useful ontologies, apply sophisticated statistical analysis and translate these findings to the clinical applications. It is not only a great opportunity but also a great challenge, as translational medicine big data is difficult to integrate and analyze, and requires the involvement of biomedical experts for the data processing. The rise of translational and personal medicine have been made possible due to the advancement in many high-throughput technologies to study the cellular processes and molecular functions of organisms at different levels as described earlier. These technologies such as genome, transcriptome, proteome, lipidome, metabolome, epigenome, and microbiome collectively called -omics of both single as well multi-cell samples, their integration and systems biology, have greatly advanced our understanding of human health and diseases (Canuel et al., 2015; Hawkins et al., 2010). However, the progress comes at a cost, as translational research data sets nowadays include genomic, imaging, and clinical data sources (Bender, 2015; Topol, 2015), making them large and heterogeneous. In effect, important steps of the data management: collection, integration, analysis, and interpretation are a challenge for biomedical research. Moreover, enabling biomedical experts to efficiently use big data processing pipelines is another challenge.

As translational medicine data become more and more rich and complex, their potential in informing both clinical and basic research grows (Regan and Payne, 2015). With constantly increasing presence of high-throughput molecular profiling, it becomes increasingly important to ensure that data interpretation capabilities follow the generation of large-scale biomedical data sets (Costa, 2014; Mardis, 2010). Visualization can support greatly the processing and understanding of complex data sets on each of the steps of the data life cycle. This opportunity is actively explored in various domains of biomedical research, including clinical big data (West et al., 2015) or multiscale biomedical ontologies (de Bono et al., 2012).

Modern translational medicine approaches aim to combine clinical and molec-

ular profiles of the patients to formulate informed hypothesis on the basis of stratified data (Tian et al., 2012). Integration of plethora of sources renders these data sets complex and difficult to process. Visualization of such integrated data sets can aid exploration and selection of key dimensions and subsets for downstream analysis. In turn, visually aided data analysis allows to comprehend even complicated workflows and aids interpretation of results.

Although many data standards (CDISC, 2017; HL7, 2017) and resources, such as Electronic Data Capture (EDC) systems, are available, the current state of clinical and translational research remains largely based on paper or spread sheet based case report forms (CRFs), especially in academia. But there is a growing trend towards the adoption of EDC systems. Even though some of the well structured and well organised projects such as PPMI (Marek et al., 2011), TCGA (Weinstein et al., 2013), are collecting clinical data via EDC systems, these projects are still using project specific terminologies. This lack of consistency prohibits groups around the world that are working on the same disease from effectively sharing and integrating their data. Therefore, there is a great and pressing need to use common ontologies and standards, such as those provided by the Clinical Data Interchange Standards Consortium, CDISC (CDISC, 2017). Indeed, major efforts are needed to access and harmonize the data due to the heterogeneous formats and terminologies. This challenge has been clearly identified by funding agencies (MRC, 2013), consequently, developing research policies or supporting the development of frameworks for standardized data integration, e.g., EMIF (EMIF, 2013), BioMedBridges (BioMedBridges, 2014).

1.3.1 Clinical and translational data

With the aim of stratification of the participants, early detection of biomarkers, understanding disease history, pathology and personalised medicine, clinical and translational studies collect a range of different datatypes. As shown in Figure 1.3, the typical data types are clinical observations (also call it clinical data or phenotypic data), imaging data, mass spec data, different types of omics - Genomics, transcriptomics, mobile sensor data and are described below:

Clinical data: Data collected by the characterization of a study participant by a medical professional, for example, demographics, medical history, adverse events, study-specific questionnaires, or examinations. It is collected with the help Case Report Forms (CRFs) and Electronic Data Capturing (EDC) systems as described below in great detail.

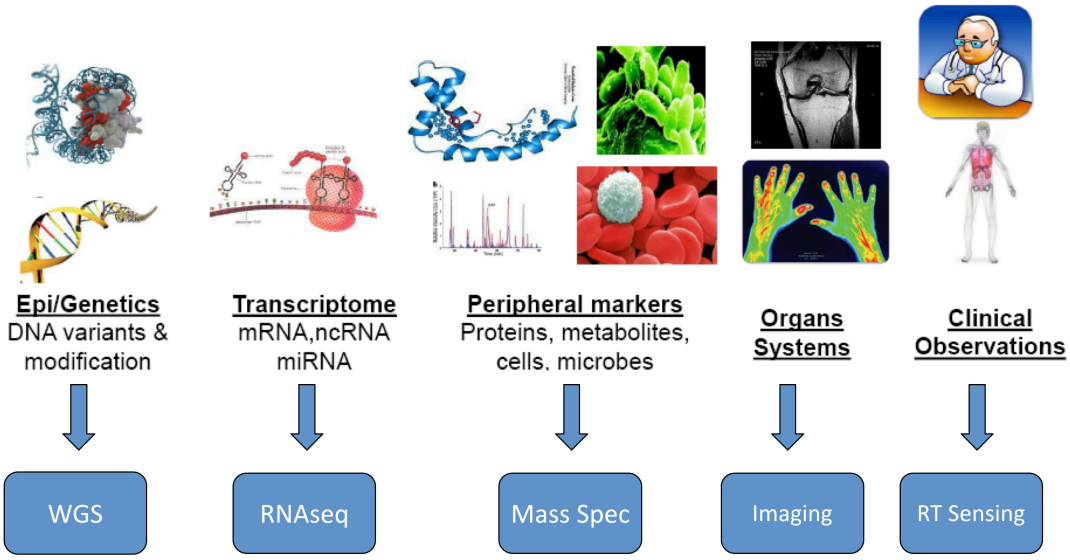


Figure 1.3: **Clinical and translational data types**. The figure shows the different data types typically collected in a clinical and translation study. These are clinical observation (also called clinical data or phenotypic data) including demographics, medical history, concomitant medication, adverse events etc., imaging of different organ systems, mass spec of different types of peripheral markers like proteins, metabolites, cells, microbes etc., transcriptome data (e.g. mRNA, ncRNA, miRNA), and genetics/epigenetics data.

Case Report Form (CRF): A CRF consists of either paper based or digital form of questionaries, domain (e.g. disease) specific scales, evaluation forms, fields to collect meta-data related to specimens collected from the participants. It is designed to collect the participant's (patient and healthy control) clinical data by a qualified clinician or a study nurse after signing the informed consent and fulfilling the inclusion-exclusion criteria. The International Conference on Harmonization Guidelines for Good Clinical Practice define the CRF as: "A printed, optical or electronic document designed to record all of the protocol

required information to be reported to the sponsor on each trial subject" (ICH-Guidelines, 1996).

The development of CRF represents a significant part of the clinical research and can affect study success (Nahm et al., 2011). It requires good planning in advance and one should pay attention to minute details in order to design a state-of-the-art CRF that should facilitate pooling of the data from other similar cohorts or cross-study comparisons and analysis. In addition, designing a CRF is crucial in a clinical trial as it will help in assessing the safety and efficacy of the drug or other medicinal product accurately. CRF should be designed for optimal collection of data in accordance with the study protocol compliance, regulatory requirements and shall enable the researcher test the hypothesis or answer the trial related questions (Bellary et al., 2014). A well-designed CRF should represent the essential contents of the study design and study protocol and ensure the data quality required by the protocol, ethics, safety needs and regulatory compliance. It should facilitate the compliance with standards like CDISC, controlled terminologies, vocabularies, standard ontologies and provide certain quality control of the data (Latimer, 2008; Lu and Su, 2010). In an ideal situation, CRF is designed once the study protocol is finalised. In some cases, it can be prepared in parallel, but it should fully comply with the study protocol (Bellary et al., 2014).

Poorly designed CRF results in frequent modification of backend database. Each change needs ethical and data protection approval that affects the study timelines, milestones and deliverables. CRF should collect the data that is required by the study protocol. Collection of large amounts of not-needed-data will results in waste of resources in collection, curation, integration and analysis. Questions in the CRF should be clear and unambiguous to avoid unnecessary confusions (CDISC, 2017).

In addition to CDISC Foundational Standards (PRM - Protocol Representation Model; SEND - Standard for Exchange of Nonclinical Data; SDTM - Study Data Tabulation Model; PGx - Pharmacogenomics / Genetics; CDASH - Clinical Data Acquisition Standards Harmonization; ADaM - Analysis Data Model), it also provides Therapeutic Area (TA) standards to represent data specific to particular disease area. These standards are developed in close col-

laboration with specific disease area experts and consortiums. They include disease-specific metadata, examples and guidance on implementing CDISC standards for a variety of uses, including global regulatory submissions. The list of disease areas for which TA standards are available can be found from https://www.cdisc.org/standards/therapeutic-areas.

Electronic Data Capturing (EDC) systems: Customarily, paper based CRFs or questionnaires have been used for gathering data as a part of clinical practice and research studies. Despite being simple and straightforward, manual data entry into spreadsheet or database and further quality control is tedious, time consuming and prone to errors (Velikova et al., 1999). With the advancement of information technology (IT), EDC systems have become an alternative option to paper based data collection. EDC systems are mostly web based applications with database backend, run in a web-browser and covers both single site or multi-site clinical studies. Some of the EDC systems are also standalone applications to run on a client desktop or server at a single site. These systems are designed to collect and manage clinical and laboratory data in digital format (Bart, 2003; Litchfield et al., 2005; Shah et al., 2010).

Although all the EDC systems are meant to capture the data digitally, there are different types of systems based on the technology and the intended use. In the case of studies using eCRF, the data of the participants will be directly enter into the EDC system during the clinical visit which is the ideal procedure. In other cases digitalize after collection on paper based CRF. Some EDC systems also provides interactive voice response systems (IVR) where a participant can report information through a phone, electronic diaries that capture participant reported outcomes and collection systems that use tablets to capture data. These sophisticated modules not only complement data collection in clinical and translational studies, but are also for building clinical registries and databases. Here are some other examples where EDC systems are generally used: phase 3 and Phase 4 clinical trials, pharmacovigilance studies, and safety surveillance activities (Koop and Mosges, 2002).

Usually an EDC system provides easy-to-setup, easy-to-configure CRFs and provides user friendly and intuitive interface. Other typical features range from

multi-level user access control with multi-factor-authentication, quality control of the data at the entry, multi-site and remote data entry. They also provide immediate feedback to site staff regarding inconsistencies, branching logic, multi-lingual support, digital signature, audit trial, data import/export in a variety of formats and the ability to reuse CRFs. They also comply with industry standards such as CDISC Operational Data Model (ODM), CDASH, E2B (Clinical Safety Data Management: Data Elements for Transmission of Individual Case Safety Reports) (Ahluwalia, 2016; FDA, 2014; ICH, 2017) and SDTM, Good Clinical Practice (GCP) and Electronic Records and Electronic Signature (ERES) compliant (CDISC, 2017; Medidata, 2013; Russell and McHale, 2010). Some EDC system variants also include the ability to analyze data and generate reports. Apart from these standards, EDC systems should comply with United States Food and Drug Administration (FDA) Federal Regulation - '21CFR part 11' as described below, in order to submit the clinical studies to FDA.

"Title 21 CFR part 11: Title 21 CFR Part 11 is the part of Title 21 of the Code of Federal Regulations that establishes the United States Food and Drug Administration (FDA) regulations on electronic records and electronic signatures (ERES). Part 11, as it is commonly called, defines the criteria under which electronic records and electronic signatures are considered trustworthy, reliable, and equivalent to paper records and handwritten signatures executed on paper (Title 21 CFR Part 11 Section 11.1 (a)). This regulation, which applies to all FDA program areas, was intended to permit the widest possible use of electronic technology, compatible with FDA's responsibility to protect the public health" (FDA, 2002, 2003, 2007, 2017; RISC, 2016; Wikipedia, 2017).

The number of published trials that use an EDC system has been rising (Paul et al., 2005), and there have been claims of a rapid uptake of this technology in clinical trials (Borfitz, 2007). However, the numbers are far from expected, a survey conducted in 2007 concluded that only 20% of trials are using EDC systems (El Emam et al., 2009) and by the end of the year 2017, it is expected to be 40% (Borfitz, 2017). There are several benefits associated with EDC systems. Online eCRF supports real time multisite data capture, and validation at data entry ensures accuracy and data quality characterized by low incidence of problematic or missing values (Velikova et al., 1999). Well defined eCRF with standards reduces

downstream data curation and standardisation efforts, which is tedious and manpower intensive. This eventually accelerates study processes and substantially reduces the study time thus reducing the overall study cost (Prokscha, 2011).

EDC systems can be broadly classified into two categories based on licensing model, as commercial and open source. An exhaustive list of EDC systems is available at URL: `http://edcmarket.appspot.com/edcsystems`, where one can compare the features of selected EDC systems. Commercial EDC applications are usually developed by industry group of developer for-profit. They charge the users for licenses and generally the source code is not published. Some examples include, Oracle Clinical, an integrated Clinical Data Management (CDM) and Remote Data Capture (RDC) solution that includes functionality in key areas such as integration, data collection, localization, and reporting (Oracle, 2017b); Oracle Health Sciences InForm, a cloud based global trial management system (Oracle, USA) (Oracle, 2017a); DATATRAK Electronic Data Capture and Medical Coding (DATATRAK, USA) (DATATRAK, 2017). On the other hand, free and open source applications are developed on voluntary basis by a single or group of developers. The source code and corresponding applications are published online and users can use them for free. For example REDCap (Vanderbilt University, USA) (Harris et al., 2009; REDCap, 2017); OpenClinica (Akaza Research, USA)(OpenClinica, 2017); DADOS P (Research on Research group, Duke University, USA) (Nguyen et al., 2006). These open-source EDC systems are described below:

- **REDCap:** Research Electronic Data Capture system- REDCap is a freely available secure web based application for building and managing online surveys, CRFs and databases. While REDCap can be used to collect virtually any type of data (including 21 CFR Part 11, FISMA (Federal Information Security Modernization Act (USA-Homeland-Security, 2014)), and HIPAA(The Health Insurance Portability and Accountability Act (USA-HHS, 1996))-compliant environments), it is specifically geared to support online or offline data capture for research studies and operations. The REDCap Consortium, a vast support network of collaborators, is composed of thousands of active institutional partners in over one hundred countries who utilize and support REDCap project (Harris et al., 2009; REDCap, 2017).

REDCap was initially developed and deployed in 2004 at Vanderbilt University Medical Center by Paul Harris to support a small group of clinical researchers who needed a secure data collection tool comply with HIPAA standards. Later in 2006, it was offered to other institutions and a consortium of domestic and international users was formed. Currently (November 2017), REDCap is used by 626,000 users from 2,597 institutions in 116 countries for around 475,000 projects and published around 4,217 articles so far (REDCap, 2017). REDCap is easy to setup and configure. Its features include: an intuitive interface to build the eCRF and secure data collection that supports data validation, quality control and data export in multiple formats (Excel, SPSS, SAS, Stata, R), supports multisite projects, double data entry, data management (lock forms). It also supports other advanced features such as branching logic, calculations, audit trails for tracking data manipulation, export procedures, file upload and provides reporting tools and an API (Shah et al., 2010). It is continuously developed by Vanderbilt University and consortium members voluntarily. In May 2017, a communication platform, REDCap Messenger, is built directly into REDCap, allowing users to communicate easily, efficiently, and securely. It is a chat application that supports one-to-one direct messages and group conversations, as well as project-linking, document and image sharing (REDCap, 2017).

- **OpenClinica:** OpenClinica is an open-source software for clinical research. First released in 2005, it has both EDC and data management capabilities. It is available to download and install on client environments and also available in the cloud. OpenClinica facilitates collection, validation, annotation of clinical data and also has features that allow study audits, reporting, data extraction along with tools to capture electronic patient reported outcomes (OpenClinica, 2017; Shah et al., 2010).

- **DADOS Prospective:** DADOS Prospective is a Web-based freely available tool for data collection on human subjects for clinical and translational trials. It is full compliance with HIPAA guidelines and 21 CFR Part 11 for collecting and storing patient data on secure database. Its features include,

eCRF setup, collect data in single and multiple studies, electronic signature, allows storage of source documents, medical images, audit trails, record locking after signature and extract data in an interoperable format (Nguyen et al., 2006; Shah et al., 2010).

Molecular data: Data collected by analyzing samples donated by a study participant using imaging (microscopy) or high-throughput molecular profiling ('-omics'). Examples are mammograms, brain scans, mass spec analysis of blood proteins, metabolites, transcription analysis of mRNAs, ncRNAs, miRNAs and sequencing analysis of genomes. Data sets generated from omics analysis are high dimensional, ranging from hundreds to hundreds of thousands of variables per sample.

1.3.2 Clinical and molecular (omics) data integration platforms

A combination of clinical and high-throughput molecular profiles (-omics) creates a very variable heterogeneous data set, where dimensionalities of different data types span several orders of magnitude (Wade, 2014). We can represent this data as a data-cube, where x-axis represent participants, y-axis represent clinical and molecular features, where as z-axis represents time points (followup visits). Moreover, ensuring veracity, quality of clinical data is a challenging and time-consuming task (Merelli et al., 2014; Stonebraker et al., 2013). This stems from a variety of collection methods, featuring manual data input, nondigital data capture, and nonstandard formats. It needs to be stressed that proper data curation is a mandatory step for accurate analysis of clinical data and proper interpretation of analytical results (Satagopam et al., 2016).

The emergence of big biomedical data sets, covering dozens of thousands of participants (Wade, 2014), raises questions on infrastructure necessary to host and analyze them. Especially genomic data, generated rapidly due to dropping sequencing costs, pose a problem in terms of storage and analytics. The perspective of cloud computing is postulated as a solution to this challenge, as summarized in recent and extensive reviews (Alyass et al., 2015; Costa, 2014; Luo et al.,

2016). Nevertheless, due to ethical and legal issues arising in cloud-based scenarios (Dove et al., 2015), incorporation of clinical data and processing of sensitive omics are still considered as an open question (Satagopam et al., 2016).

Translational medicine platforms integrating clinical and omics data need to ensure a protected environment for sensitive data processing. A number of solutions were developed to address this challenge, as summarized in an excellent review by Canuel et al. (Canuel et al., 2015). Platforms integrating clinical and omics data generally consists of four layers: (i) a data layer, to store the data, (ii) a semantic layer, to integrate and standardize the data by the use of ontologies, (iii) an application layer, to manage clinical databases, ontologies and data integration process, (iv) a presentation layer (web interface), to interaction with users (Miyoshi et al., 2013; Tradigo et al., 2014). These platforms can be divided into two groups: repositories with an existing infrastructure developed at a site and solutions requiring deployment. The first group is represented by technologies such as STRIDE (Lowe et al., 2009), iDASH (Ohno-Machado et al., 2012), Information Warehouse (IW) (Kamal et al., 2010), caGRID (Oster et al., 2007), and its follow-up, TRIAD (Payne et al., 2011) or BDDS Center (Toga et al., 2015). Certain platforms of this group focus on a specific disease, such as cBioPortal (Cerami et al., 2012), OncDRS (Orechia et al., 2015), G-DOC (Madhavan et al., 2011) for cancer or COPD Knowledge Base (Cano et al., 2014) for pulmonary dysfunction. The advantage of solutions based on existing computational infrastructure is direct use but at the cost of reduced flexibility in configuration and toolset management. The other group of solutions for translational medicine requires deployment on the user's infrastructure, often requiring substantial storage or high-performance computing (HPC) capabilities. Notable examples in this group are BRISK (Tan et al., 2011) and tranSMART (Szalma et al., 2010). Because of their highly configurable nature, such solutions are suitable in projects involving sensitive data, and where a repository is needed to support ongoing projects, such as the case of longitudinal cohort studies. Informative use cases of such platforms are SHRINE (Natter et al., 2013) and DARiS (Nguyen et al., 2015), where well-defined demands of clinical research projects drove the design and implementation of infrastructure supporting translational medicine (Satagopam et al., 2016).

Visually aided data exploration is an important component of clinical and omics integration platforms. A notable contributor in this field is the Informatics for Integrating Biology and the Bedside project (i2b2, www.i2b2.org), a scalable framework enabling the use of clinical data for discovery research (Murphy et al., 2009, 2010). The i2b2 Hive (Gainer et al., 2007) is a comprehensive collection of interoperable tools ranging from repository services to basic data conversions provided by i2b2 cells. Importantly, i2b2 Hive does not support directly the analysis of omics data, such as gene expression or whole-genome sequences by itself (Gainer et al., 2007), but enables key capabilities of clinical data exploration and processing to be used by other platforms (Satagopam et al., 2016).

Among these systems, tranSMART is a well established platform enabling translation of clinical, preclinical and translational research data into meaningful biological knowledge (Scheufele et al., 2014). It supports integration of low-dimensional clinical data and high-dimensional molecular data in a data warehouse architecture as shown in Figure 1.4. In addition it also stores the annotations from external databases and ontologies. Although tranSMART by default relies on a relational database technology, it extends toward storing the high-dimensional biological data using NoSQL technology HBase (Wang et al., 2014). The platform features data interoperability connectors, including clinical information collection (Blond and de Bruijn, 2015), imaging data (Vast, 2015), visual analytics (Herzinger, 2015; Herzinger et al., 2017), or bioinformatics workflow management (Bierkens et al., 2015). Finally, tranSMART features builtin data mining and analysis applications based on open-source systems, such as i2b2 and GenePattern (Szalma et al., 2010), and provides plugins to external tools, such as Dalliance Genome Browser (Down et al., 2011), or APIs for statistical packages, such as R.

1.3.3 Bioinformatics workflow management systems

Reusable and interoperable bioinformatics workflows become increasingly important in reproducible analysis of biomedical data and metadata, including clinical, omics, imaging, and sensor data (Affeldt et al., 2016; Afgan et al., 2015; Leipzig, 2017). A number of software frameworks were developed to support the scientific

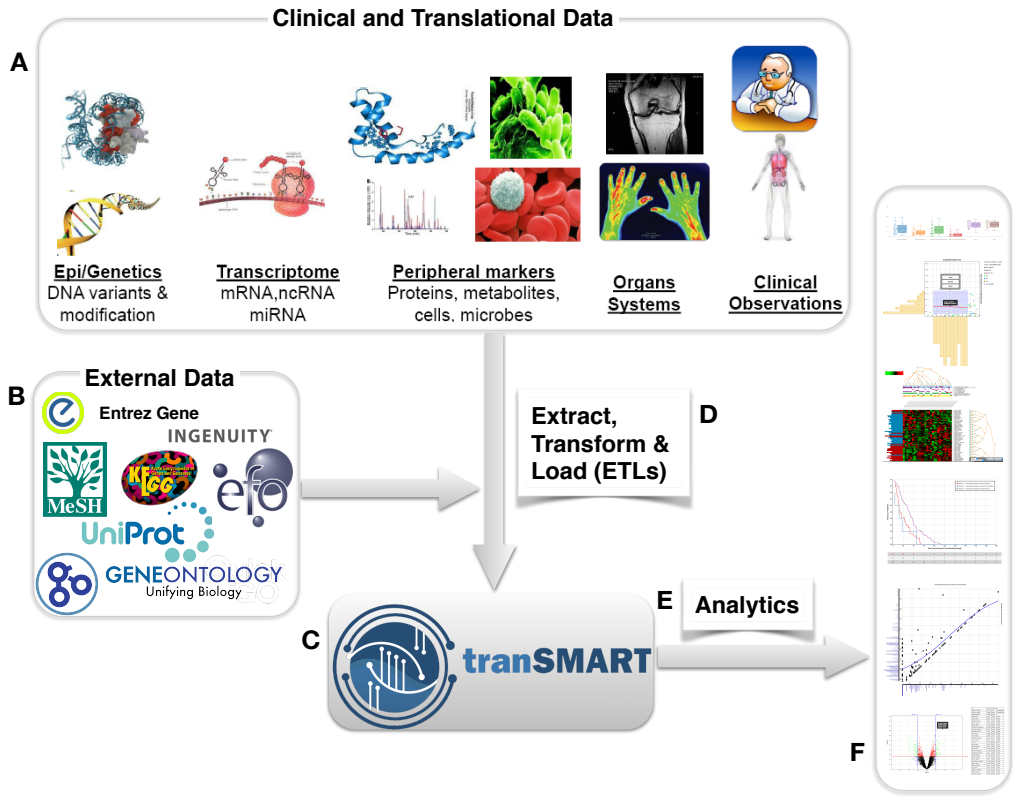


Figure 1.4: **tranSMART overview**. The figure shows overview of data integration and analysis in tranSMART, A) different data types typically collected in a clinical and translation study, B) publicly available annotation databases such as EntrezGene (NCBI Gene), KEGG, GeneOntology, Experimental Factor Ontology (EFO), literature can be integrated into tranSMART to annotate and enrich the clinical and translational data. The data from A) and B) loaded into C) tranSMART with the help of D) ETL (Extract, Transform and Load) scripts and tranSMART provides various E) analytics tools and advanced workflows to slice and dice the data, sub-cohorts selection and run the various dynamic visual analytics as depicted in F).

community in this goal. In a thorough review and classification of these workflow frameworks (Leipzig, 2017), Leipzig, J clustered existing technologies according to their interaction mode into command-line/application programming interface (API) and workbench approaches. The first group includes Snakemake (Koster and Rahmann, 2012), Yabi (Hunter et al., 2012a), Chipster (Kallio et al., 2011), or Mobyle (Neron et al., 2009) and relies on textual workflow construction in a script-like format. Certain tools in this group, such as Chipster, also enable Web-based collaborative development of workflows (Satagopam et al., 2016).

The second group of platforms provides the so-called 'workbench environment': a GUI enabling visually supported construction of workflows. Usually, workflows are represented as graphs, where nodes correspond to data processing steps, and edges to data flow. Workbench solutions include Galaxy (Goecks et al., 2010), Taverna (Wolstencroft et al., 2013), Pipeline Pilot (Warr, 2012), KNIME (Jagla et al., 2011), and gUSE (Kacsuk et al., 2012). Similar to data integration platforms, these tools need to be deployed on the user-provided infrastructure, and the extent of possible analysis is restrained by available storage and HPC capacities (Satagopam et al., 2016).

Ensuring computational resources may be a challenging task, and cloud computing becomes increasingly more important paradigm in development and execution of bioinformatics workflows. Cloud-oriented workflow management systems offer API support for construction of an analytical pipeline, including open-access solutions, such as Agave (Dooley et al., 2012) or Arvados (Arvados, 2013), or a number of commercial services (Costa, 2014). Workbench platforms are also available in the computational cloud environment. Interestingly, a number of open-access solutions use Galaxy as a workflow construction engine, including Galaxy Cloud (Afgan et al., 2011), Tavaxy (Abouelhoda et al., 2012), or Genomics Virtual Laboratory (Afgan et al., 2015). Commercial cloud workbenches, such as Seven Bridges (http://sbgenomics.com), are also available. In summary, cloud computing is an attractive scalable option on demand, especially for multisite collaborative research projects in terms of bringing the tools to the data. However, the speed of data transfer to the cloud, flexibility of the configuration of analytical pipelines, and the issues of privacy and security in data analytics remain challenges to address (Alyass et al., 2015; Leipzig, 2017).

1.3.4 Platforms for visualization of molecular interaction networks

Molecular interaction networks are a class of graphs, where nodes represent various biomolecules, and edges represent interactions between them. With the progress of systems biomedicine, molecular interaction networks became a popular form of representing knowledge about molecular mechanisms pertinent to

human health (Jacunski and Tatonetti, 2013). First, such networks provide a necessary format to encode multitude of interactions identified in biomedicine. Second, they provide a good support for visual exploration and analytics of complex knowledge (Gerasch et al., 2014). As such, they have a great potential in aiding the interpretation of analytical outcomes of translational medicine pipelines.

Molecular interaction networks can be constructed in various ways that determine their size and purpose. Experiment-derived networks are established from different types of molecular readouts, allowing, with a certain probability, to ascertain physical interaction between molecules, for example, protein-protein interaction (Pizzuti and Rombo, 2014) or chromatin immunoprecipitation assays (Kim and Park, 2011). Analysis-inferred networks are constructed by analyzing high-throughput omics data to identify molecules with similar properties or behavior, for example, using coexpression analysis (Guo and Wan, 2014). Finally, knowledge-based networks are established on the basis of existing body of knowledge, usually a set of published articles. Construction of knowledge-based networks is usually accomplished with text mining approaches (Neves and Leser, 2014) and/or expert curation (Fujita et al., 2014; Kutmon et al., 2016).

While experiment-derived and analysis-inferred networks offer a vast amount of unbiased information, they are usually large-scale graphs, requiring sophisticated network analysis to draw meaningful conclusions. Mapping translational medicine data sets on top of these networks may be considered an important step in the analysis (Glaab and Schneider, 2012) but not in the final interpretation of an analytical workflow. In turn, knowledge-based networks are usually established on the basis of low-throughput, in-depth experiments and allow for direct data interpretation. In particular, text mining generated networks are often used by the scientific community, where a number of commercial solutions, such as Ingenuity Pathway Analysis (Kramer et al., 2014), Pathway Studio (Pathway-Studio, 2016), or MetaCore (MetaCore, 2016), offer already established databases. These solutions, however, tend to contain the entire discovery pipeline inside their platforms, greatly reducing data interoperability (Satagopam et al., 2016).

Expert-curated networks focus on resources of high-quality confirmed knowledge and offer the highest degree of data set interpretation to translational medicine researchers. Important resources in the field of expert-curated net-

works are repositories called 'pathway databases', such as KEGG(Kanehisa and Goto, 2000), Reactome(Croft et al., 2014), or WikiPathways (Kutmon et al., 2016), which describe general biomolecular mechanisms. In contrast, the other type of networks focuses on representing mechanisms of human diseases as so-called "disease maps (Fujita et al., 2014; Kuperstein et al., 2015; Mizuno et al., 2012)". Detailed representation of domain knowledge and support by domain-related literature make disease maps a potentially interesting element of translational medicine analytical pipelines. Computational architectures supporting these maps provide dedicated APIs (Bonnet et al., 2015; Gawron et al., 2016), opening an interesting avenue in translational medicine data processing - from storage, through bioinformatics workflow analytics, to interpretation by visualization on the dedicated molecular interaction network (Satagopam et al., 2016).

In the Chapter 8, I have focused on translational medicine workflow providing the possibility of visually aided data exploration and informative hypothesis formulation. In this workflow, the data integration platform of choice was tranSMART as a server-based solution with i2b2 data exploration component, and Galaxy as a workflow management system, considering its flexibility and the availability of tools. Finally, to provide informative interpretation of analytical results, we bridged the Galaxy Server with MINERVA platform, allowing overlay of analysis results on disease-related mechanisms.

1.4 Biological application areas

The data integration, knowledge management and analysis methodologies described above are applied to different biological areas. Few of them are discussed briefly in the following sections:

Hutchinson - Gilford progeria syndrome (HGPS): HGPS is a rare premature aging disorder caused by mutations in LMNA gene encoding A-type nuclear lamins (Marji et al., 2010). The children with this disease age 10 times faster than normal humans. Due to this accelerated aging progeria children have the same physiological conditions (e.g., cardiovascular, respiratory, and arthritic conditions) as elderly people. On average, these children will die at the age of 13

due to progressive coronary atherosclerosis (Hennekam, 2006). There are very few children (∼100) living with progeria (http://progeriaresearch.org). Apart from progeria, mutations of lamin A/C (LMNA) cause a wide range of human disorders, including lipodystrophy, neuropathies and autosomal dominant Emery-Dreifuss Muscular Dystrophy (EDMD) (Bakay et al., 2006). Refer Chapter 3 for more details.

Huntington's Disease (HD): HD is an autosomal dominant, monogenic, fatal neurodegenerative disease (Landles and Bates, 2004). Currently, there are no effective ways to slow or prevent the neurodegeneration caused by the HD mutation (Leegwater-Kim and Cha, 2004). The disease is caused by expansion of a CAG repeat within the first exon of the gene encoding huntingtin locus (HTT), a large (>300kDa) protein of poorly characterized function. The mutated (HD) allele therefore includes an expanded polyglutamine stretch at the amino terminus of increased length (>35 Q residues). Clinically, the disease is characterized by progressive motor, cognitive and behavioural symptoms, resulting from degeneration of cortical and striatal neurons. Disease onset, progression and severity are variable and depend at least in part on the length of the CAG repeat. To date, nine such polyglutamine (polyQ) neurodegenerative disorders known to be caused by expansion of the CAG repeat in the coding region of the respective genes have been identified (Ashkenazi et al., 2017; Siwach and Ganesh, 2008). Among these human neurodegenerative pathologies, Huntington's Disease is a prototypic example where the disease is caused by a characterized mutation. Refer Chapter 4 for more details.

G-protein coupled receptors (GPCRs): Signal transduction refers to these cellular processes by which stimuli, either physical or chemical, induce specific cellular responses, through chosen molecular mechanisms. The specificity of a cellular response to a signal depends on the receptor expressed on the target cell (Satagopam et al., 2010). GPCRs are a very important superfamily of cell membrane receptors in eukaryotic cells. They may interact with both the environment outside and inside the cell and they play a crucial role in receiving stimuli signals from the environment. In response they induce certain cellular responses.

GPCRs have a characteristic structure comprised of seven transmembrane spanning α-helices, an extracellular N terminus, an intracellular C terminus and three interhelical loops on each side of the membrane (Oldham and Hamm, 2008). Ligands bind to GPCRs on the outside of the cell, activating the GPCRs by causing a conformational change, and allowing them to bind to G-proteins (McCudden et al., 2005). Through their interaction with G-proteins, several effector molecules are activated leading to many kinds of cellular and physiological responses. The great importance of GPCRs and their corresponding signal transduction pathways is indicated by the fact that they take part in many diverse disease processes and that a large part of efforts towards drug development today is focused on them (Satagopam et al., 2010). In this dissertation I present Human-gpDB, a database which currently holds information about human GPCRs and their interactions with G-proteins and effectors. Refer Chapter 5 for further details.

1.5 Aims of the project

Data integration and knowledge management efforts of my PhD project focused on integration of both structured and unstructured data from several publicly available sources and development of different knowledge rich resources and intuitive web applications to apply to large scale experiments and data collection in "production environments".

The purpose was to develop methods, build tools to analyse high-throughput experimental data like transcriptomics (gene expression analysis), proteomics, whole genome siRNA, miRNA knockdown experiments, next-generation sequencing (NGS) etc., obtained from different clinical, preclinical and translational projects. These projects are aiming to study complex diseases such as 'cancer', 'Hutchinson - Gilford progeria syndrome (HGPS)', neurodegenerative diseases: 'Huntington's disease', 'Parkinson's disease'. These tools also can be used to study physiological phenomenon such as 'Cell cycle regulation', in order to prioritise the biomarkers and targets for further functional validation. Thus eventually leads to the discovery of novel biological processes, pathways, diagnostics makers, drug targets and drugs.

Major part of the work focused on the collection of several publicly available

biological databases, full text articles as well as experimental data from projects such as HGPS (Marji et al., 2010) and TAMAHUD (Marji et al., 2010). I have used advanced data integration and management techniques, that are described in great detail in the Chapter 2, bioCompendium, a publicly accessible high-throughput experimental data analysis platform. The system is aimed to work with large lists of genes or proteins and to collect a wide spectrum of biological information. It is also aimed to facilitate the analysis, comparison and enrichment of experimental results; either proprietary or publicly available data sets. Typical use cases are the prioritization of potential targets from gene expression analysis studies or from RNAi studies. The current version is to work best for human, mouse and yeast model organisms. Main features of the system aimed at:

- Conversion of a wide range of input ID's like UniProt, GO, Affymetrix and RefSeq,
- Extraction of bio-entities from different file formats (MS-Office, PDF and flat text),
- Comprehensive knowledge collection from different biological database for a given list(s) of genes,
- Search interface to the knowledge collection to find information like gene annotations, disease associations, sequences domain architectures, interacting chemicals and involved pathways,
- Enrichment analysis for GeneOntology terms, diseases, pathways and other biological concepts,
- Extraction of the protein-protein, protein-chemistry interactions networks,
- Compilation of clusters based on sequence homology and sequence domain architectures in a given list(s) of genes,
- Analysis and clustering of transcription factor binding site (TFBS) profiles,
- Access to orthology information, clinical trial and patent information,

- Comparison of results derived from different experimental conditions, time series or treatments.

Chapter 3 focuses on HGPS, where as Chapter 4 on Huntington's disease, both are use cases for bioCompendium. In the former Chapter, I have analyzed gene expression data obtained from HGPS patients and healthy controls. The Differentially Expressed Genes (DEGs) were analyzed using bioCompendium, annotated and enriched with knowledge obtained from several databases. A permanent session has been created in bioCompendium with analysis results and are compared with previously published related datasets.

Chapter 4, TArgets and MArkers in HUntington's Disease (TAMAHUD) project aims at identification of early disease markers, novel pharmacologically tractable targets and small molecule phenotypic modulators in Huntington's Disease (HD). In this collaborative project, I have analyzed data obtained from RNAi screen in a human embryonic kidney (HEK293) T-REx cell line overexpressing a full length mutated Huntingtin construct. TAMAHUD knowledgebase with both TAMAHUD experimental results as well as public high-dimentional Huntington's disease datasets has been developed and is bridged with bioCompendium platform to analyze and/or compare different datasets from TAMAHUD knowledgebase.

Chapter 5 Human-gpDB is another application of data integration and knowledge management. In this project, I have build a publicly accessible knowledgebase consists of human GPCRs, G-proteins, effectors and their interactions. It is integration with several external data sources and it is a simple, yet a powerful tool for researchers in the life sciences field as it integrates an up-to-date, carefully curated collection of information. The database may be a reference guide for medical and pharmaceutical research, especially in the areas of understanding human diseases and drug discovery.

In Chapter 6, I have presented the development of two services: synonym database service, a comprehensive knowledgebase obtained from mining of various publicly available biological databases. It consists of several bio-entities (e.g., genes, proteins, drugs, chemicals, metabolites), their synonyms, annotations and cross-references to other resources. Another service, SBML maps annotation service to enrich different molecular interaction networks and disease maps e.g., Parkinson's disease map by providing rich annotations to various species elements

of the map. I have developed RESTful APIs for both these services. Systems biology tools including - CellDesigner, GARUDA, MINERVA platform - took the advantage of these biological data driven services to develop plugins (gadgets) using provided RESTful APIs to standardise and harmonise bio-entities and/or annotate them.

Then in Chapter 7 HIV Mutation Browser, I have built a publicly available residue-centric resource of HIV mutagenesis and polymorphism literature designed for use by those carrying out basic and applied HIV research. The HIV Mutation Browser is one of the first resources to computationally text-mine mutagenesis and polymorphism data, and the first to apply such methods to the extensive corpus of HIV literature. In this project, I have negotiated with publishers for bulk download and text-mine the HIV related full text articles. This literature was text-mined for mutations and mapped them to the corresponding HIV proteins.

In Chapter 8, I have demonstrated a workflow for analysis and interpretation of high-throughput translational medicine data, in which visualization is an important component at each step of data processing and exploration. In this workflow, three Web services - a tranSMART server, a Galaxy server, and a MINERVA server - are combined into one big data pipeline. It is called TGM (tranSMART - Galaxy - MINERVA) pipeline.

Finally, in Chapter 9 Conclusions, I have concluded my whole dissertation and provided my future directions.

Chapter 2

bioCompendium: High-throughput experimental data analysis platform

2.1 Description

The current field of systems biology increasingly depends on high-throughput experimental techniques like transcriptomics (gene expression analysis using microarrays or RNA sequencing), proteomics (using mass spectrometry), next generation sequence analysis, whole-genome knock-down experiment using for example miRNAs or siRNAs. This trend challenges computational biologists to develop fast, accurate, and meaningful ways of analyzing this experimental data and make sense out of it. Some of these techniques meant to study and understand physiological phenomena or a pathological conditions (disease states). These experiments may result a list of genes or gene products (mRNAs, proteins, miRNAs and so on) that may play a role in that particular biological state. The size of these lists may range from few genes to few thousands of genes. It is tedious and difficult for researchers to retrieve the information from several publicly available databases for each gene individually, process it and digest in order to develop a hypothesis and/or prioritize them for further experimental validation.

To assist the researchers in the analysis of data obtained from above mentioned

high-throughput experimental techniques, I have developed a information rich data driven approach by taking the advantage of enormous amounts of data deposited in the public databases and the literature. This approach is available as a simple and easy to use web-application as shown in Figure 2.1 and also provide an API for programmers and bioinformaticians. The bioCompendium functionality is described in the following sections.

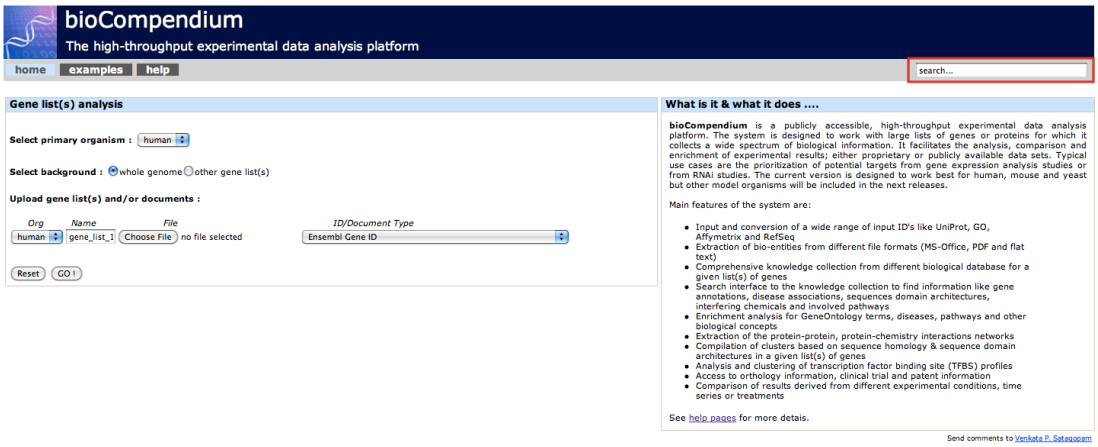


Figure 2.1: **The bioCompendium web interface**. The figure shows the front page of the bioCompendium web application. It is equipped with simple search, highlighted with red rectangle, gene list(s) analysis input block consisting of selection of primary organism, background selection, upload of gene list(s) and gene list parameters like name of organism, gene list, place holder to upload gene list, type of identifier or document. It is also providing short description of the resource.

2.2 bioCompendium implementation

bioCompendium is a three-tier web application as shown in the Figure 2.2. The first tier, the presentation layer, is developed using Perl-CGI, JavaScript, HTML, DHTML, CSS. The second tier, the application (logic) layer, is implemented in Perl, Perl-CGI, R and BioConductor, Apache HTTP web server. In the third tier, the database layer, the data is stored in a MySQL database.

Currently bioCompendium is developed for three model organisms (human, mouse and yeast) and their genome assembly versions were shown in the Table 2.1. The other model organisms will be added in future releases. The bioCompendium

Figure 2.2: **Web application architecture**. The figure shows three layered architecture of the bioCompendium.

overview is shown in the Figure 2.3 and methods used in the construction of this resource are detailed in the below sections.

Species	Common name	Genome build
Homo sapiens	human	GRCh37
Mus musculus	mouse	NCBIM37
Saccharomyces cerevisiae	yeast	EF4

Table 2.1: **The bioCompendium species table** - The list of three organisms currently integrated in bioCompudium along with their names and version of the genome build.

2.2.1 Data integration

In bioCompendium data warehousing is achieved by the concept of ETL (Extract, Transform and Load) procedure (Mayer, 2009a), where the source databases were regularly scanned and loaded into the central repository.

Ensembl gene identifiers (IDs) were used as starting points to integrate the data with various external sources. The systems that were used to help with this integration were Ensembl (Flicek et al., 2012), BioMart (Kinsella et al., 2011) (Guberman et al., 2011), SRS (Etzold and Verde, 1997) (with more than 80 different important biological databases) and a text mining resource, 'Alma Knowledge Server2' (AKS2). These resources were installed locally, that facilitates faster access to the data compared to accessing them from remote services through the web. In our studies, these data resources were scanned for each gene from the above selected three species and the relevant information as for example, gene names, description, synonyms, chromosomal location, gene function, cellular

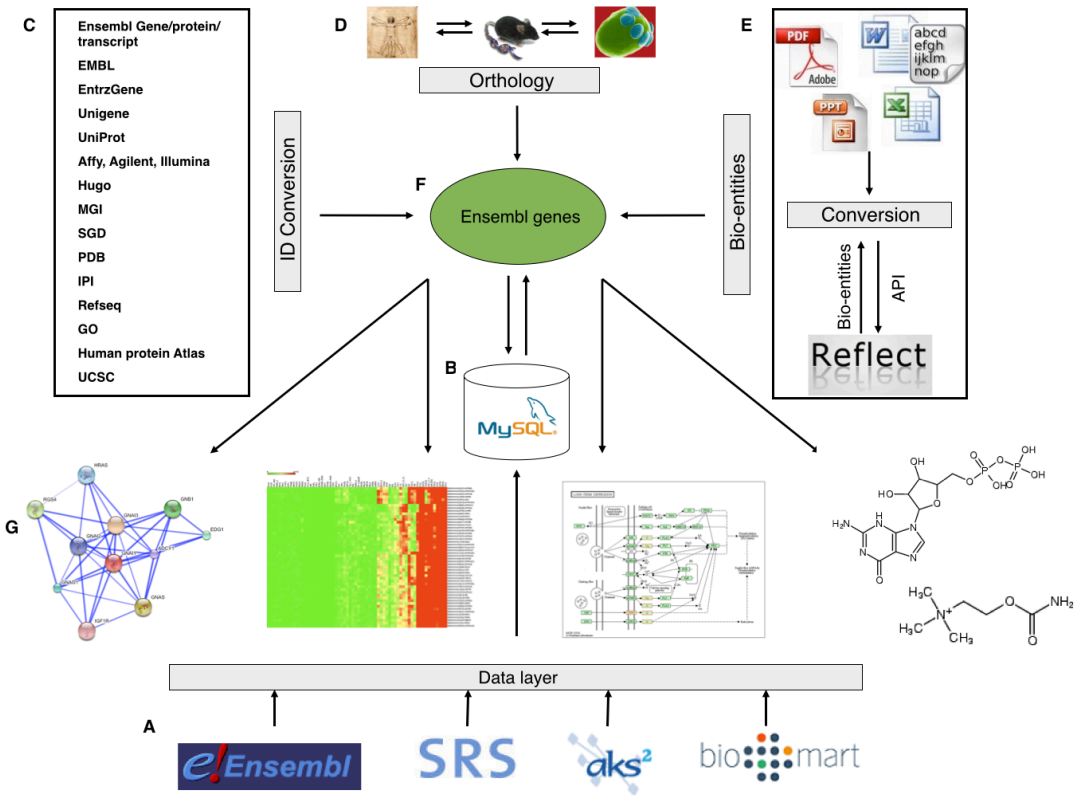


Figure 2.3: **The bioCompendium overview**. The above figure shows a schematic overview of bioCompendium consisting of (A) Data layer, in which data from more than 80 important biological databases in different formats (flat-file, XML,RDBMS) is collected locally and integrated into the SRS system. The local Ensembl, BioMart, a text mining resource (AKS2) are also used. All these databases are scanned for each gene from three currently integrated organisms depicted in (D) and the relevant information is stored in a local MySQL database (bioCompendium knowledge base)(B). As scientists use different database identifiers (IDs) to represent their experiment results, in order to compare the results from one format of IDs to another, I have implemented an ID Conversion service (C). Researchers also do experiments in different model organisms. Therefore to achieve the cross-species comparison and enrichments, I am converting information from one genome to another by using orthology relationships (D). One more unique feature of bioCompendium is handling of documents (PDF, MS-Word, Excel, PowerPoint, plain text) to extract gene lists. I have implemented an API to access the Reflect (O'Donoghue et al., 2010), a bio-entity tagging service. After converting any of the above mentioned documents to ascii text, that text is sent to the Reflect via an API and the tagged version of the text is received back. The Bio-Entities (genes/proteins) are extracted and processed like a gene list. The information in bioCompendium is Ensembl gene centric, all the information from methods depicted in (C), (D) and (E) boils down to Ensembl genes (F) and it provides several bioinformatics analysis results. Few of these are shown in Figure (G) - protein-protein, protein-chemical interaction network, transcription factor binding site profiling, KEGG pathway enrichment and related drugs and chemicals.

location, involved biological processes, pathways, reactions, regulatory elements, encoding proteins, protein features, 3D structural information, disease association, sequence, cross references, interacting proteins, drugs etc., were then stored in a local MySQL database (bioCompendium knowledge base) as shown in Figure 2.3(B). All the data sources under the 'Data layer' depicted in Figure 2.3(A) are detailed in the following sections.

Ensembl: This database provides up-to-date genome annotation information for several model organisms e.g., human, mouse, rat and zebrafish. The version 76 consists of 68 species comprising of vertebrates and other eukaryotic organisms. It provides gene, regulatory annotations and comparative genomics resources including homology, paralogy and orthology relationships. It also provides variation (Single Nucleotide Polymorphisms (SNPs)) information and incorporates data from ENCODE project (ENCODE-Consortium, 2011). The knowledge from this locally installed database is accessed via Ensembl Application Programing Interface (API).

BioMart: It is an open source data management system that enables scientists to perform advanced querying of a biological data source through a single web interface. It also provides an API to access the data as well. Further, this resource is used to get the identifier mappings to Ensembl database.

SRS: Sequence Retrieval System (SRS) is a commercial platform for integration of biological databases that are heterogenous in content and format. It provides rapid and easy access to the large amounts of diverse data stored in several public domain databases through a single user interface. One of the important feature of SRS is the linking functionality, as these databases have cross-references to other databases and is also able to explore the relationships between the different sources of biological data. More than 80 important biological databases are integrated in local SRS and it is an important source of data for bioCompendium. All these database are updated on nightly basis.

AKS2: AKS2 (Alma Knowledge Server 2) is a commercial (now retired) biomedical knowledge management system. It is based on the advanced text mining technology to extract information from the Medline abstracts, and make it available to researchers through an easy to use interface (Bioalma, 2010). This resource is used to extract the drugs and chemicals interacting with the proteins.

2.2.1.1 Biological database resources

The databases used in the development of bioCompendium, Human-GpDB, SBML map annotation service can be broadly categorized into the following different knowledge domains -

- Genes/Proteins: Ensembl (Flicek et al., 2012), EMBL (Cochrane et al., 2009), GenBank (Benson et al., 2011), EntrezGene (Maglott et al., 2011), Unigene (Schuler, 1997), UniProt (Magrane and Consortium, 2011), IPI (Kersey et al., 2004), NCBI Protein (Sayers et al., 2012), RefSeq (Pruitt et al., 2012), HGNC (Seal et al., 2011), UCSC (Fujita et al., 2011), KEGG (Kanehisa et al., 2002), GeneCards (Safran et al., 2010)
- Diseases: OMIM (Hamosh et al., 2005), KEGG diseases (Kanehisa et al., 2010)
- Protein structures: PDB (Berman et al., 2007), HSSP (Schneider and Sander, 1996), PSSH (Schafferhans et al., 2003)
- Protein features: Pfam (Punta et al., 2012), SMART (Letunic et al., 2009), PRINTS (Attwood et al., 2012), InterPro (Hunter et al., 2012b)
- Chemicals: DrugBank (Wishart, 2008), HMDB (Wishart et al., 2009), PubChem (Bolton et al., 2008), chEBI (Degtyarenko et al., 2008), STITCH (Kuhn et al., 2008), MATADOR (Gunther et al., 2008), PDBLigand (Feng et al., 2004), chemicals from literature (AKS2)
- Pathways: KEGG (Kanehisa et al., 2012), PANTHER (Mi et al., 2005), Reactome (Croft et al., 2011)
- Ontologies: GeneOntology (GO) (Ashburner et al., 2000)

- Interactions (protein-protein, protein-chemical): STRING (Szklarczyk et al., 2011), STITCH (Kuhn et al., 2008)
- Transcription factor binding site position weight matrices: JASPAR (Bryne et al., 2008)
- Patents: EPO Proteins (www.epo.org), JPO Proteins (www.jpo.go.jp), USPTO Proteins (www.uspto.gov), KIPO Proteins (www.kipo.go.kr/en/)
- Clinical trials: http://clinicaltrials.gov database (Zarin et al., 2011)
- Orthology: Ensembl compara (Vilella et al., 2009)
- Gene expression arrays: Affy (Kinsella et al., 2011), Agilent (Kinsella et al., 2011), Illumina (Kinsella et al., 2011)
- MicroRNA: miRBase (Kozomara and Griffiths-Jones, 2011)
- Literature: PubMed/MEDLINE (PubMed, 2015)

2.2.1.2 Database identifiers

Currently the bioinformatics field has to deal with a wide range of biological database identifiers. For example, some scientists tend to use Ensembl based identifiers to represent their experimental results, whereas other scientists tend to use more of NCBI based identifiers; others use either of these resources. To compare the experimental results coming from different parts of the world, one should be able to convert one type of identifiers to other types easily. In order to address this issue, I have developed an ID conversion service.

ID conversion service: This service converts experimental results (gene/protein list(s)) represented in identifiers from any of the above mentioned knowledge domains to Ensembl gene identifiers (IDs) with the help of cross references mentioned in the databases. Once any type of IDs are converted into one specific type (i.e. Ensembl gene identifiers in our case), then it is easy to compare, analyze and further enrich the information.

2.2.1.3 Document handling

bioCompendium also provides a unique facility to compare the experimental results with literature directly, i.e., one can compare list(s) of genes with a pdf, word, excel, ppt or plain text document directly. In this section, the bioCompendium extracts the bio-entities present in the document and does the further comparison and analysis. To achieve this functionality programmatically, an Application Programing Interface (API) has been implemented for the Reflect resource (Pafilis et al., 2009).

Reflect API: I have developed a REST (REpresentational State Transfer) based API that allows more precise control of how a document is tagged. The API can be accessed via REST (http://reflect.ws/REST_API.html) using HTTP 'post' (O'Donoghue et al., 2010). Here is a Perl implementation that uses HTTP 'POST' method to tag small molecule and protein names in a sample HTML document:

```
#!/usr/bin/perl
use LWP::Simple::Post qw(post post_xml);
my $input = "<html><head></head><body>Omeprazole is ATP4A inhibitor\
</body></html>";
my $response = post(
"http://reflect.ws/REST/GetHTML",
"document=$input"
);
```

The document handling module in bioCompendium takes one or more documents of type PDF, MS-Word, Excel, Powerpoint, text and converts them to plain text before sending the request to Reflect through API. As a response bioCompendium receives the tagged version of the document where the bio-entities (genes/proteins, chemicals, wikipedia terms) are highlighted. bioCompendium extracts the genes/proteins present in the tagged document(s) and maps them to the Ensembl gene IDs. The important information, genes/proteins in a document or supplementary material is returned as a list of Ensembl IDs, and that can be easily compared with other experimental results and can be further analyzed.

2.2.2 Interspecies comparison

In bioCompendium, experimental results coming from different species are compared with the help of orthology relationships between the proteins. This orthology information obtained from Ensembl Compara database is based on Ensembl GeneTrees (Vilella et al., 2009).

2.2.3 Simple search

bioCompendium is equipped with an auto complete enabled search functionality. I have used Ajax (Garrett, 2005) and scriptaculous (Fuchs, 2010) JavaScript libraries to achieve the auto complete. User can search the whole database by using this simple search. The Figure 2.4 shows the tabular view of hits from bioCompendium database for the search term 'Alpha-synuclein'. It shows the number of records found in each organism for the given search term and followed by the results from each organism segregated in different sub tables. Each tables consists of gene name, description and corresponding Ensembl Gene Identifier with gene name hyper linked to bioCompendium summary sheet (refer section 2.3.3), where as Ensembl Gene Identifier hyperlinked to Ensembl database.

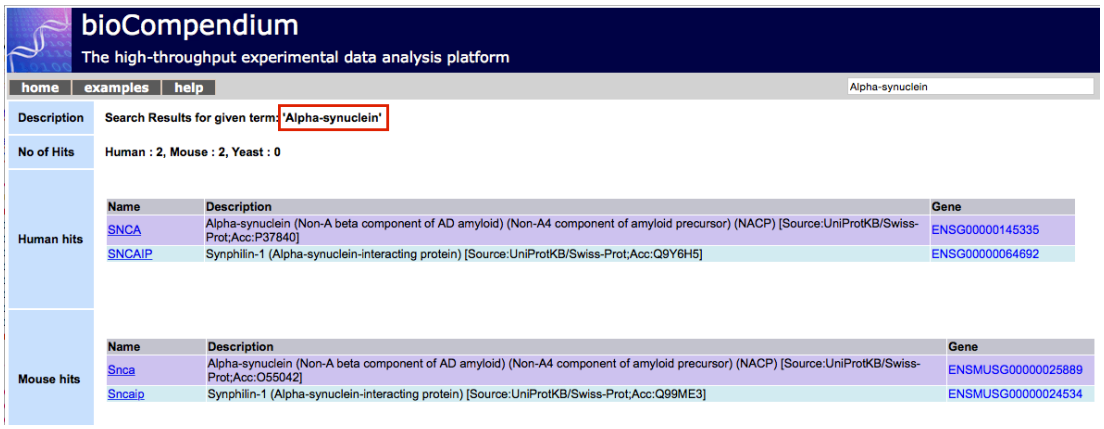


Figure 2.4: **Simple Search in bioCompendium**. The figure depicts the search results for a given search term e.g. 'Alpha-synuclein'. Initially it provides the number of records (hits) mentioning the given search term in each organism, then provides the name, description and corresponding Ensembl Gene ID for each hit in table view organised in subsections for each organism.

2.2.4 Experiment set up

In a typical experimental set up, where a bench scientist is working on a disease, he/she will compare the disease condition with the control as outlined in Figure 2.5(A). This may lead to a single gene/protein/probe-set list that may playing an important role in the disease state when compared to the control or healthy state. In addition to this experimental set up, researcher may want to find out the effect of a drug or chemical perturbation on this disease, this may lead to many gene lists as shown in Figure 2.5(B), each line represents a gene list. These list(s) need to be further compared and analyzed in order to prioritize the genes/proteins for further experimental validation. bioCompendium will help to achieve this goal. Other important scenario is before publishing the results scientists need to compare their experimental results with already published data even if the experiment done in different model organism, bioCompendium is useful here as well.

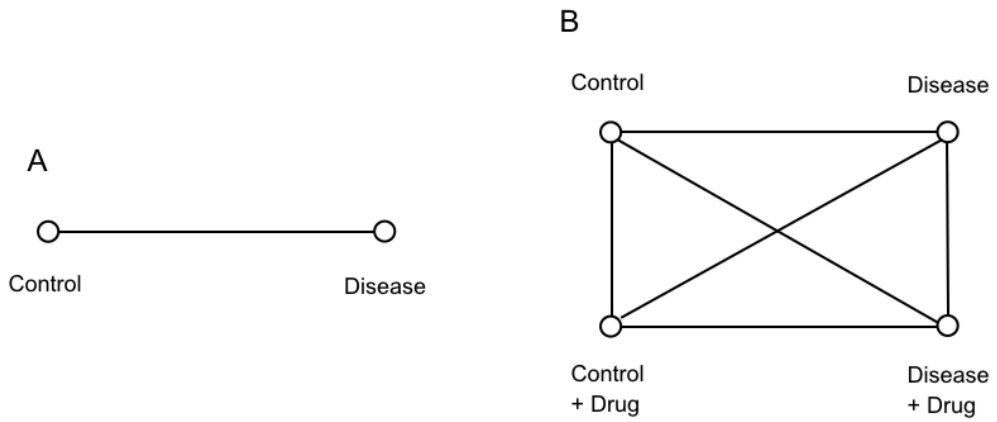


Figure 2.5: **A typical experiment setup**. The figure outlines different experimental setups, A) represents the study of a biological phenomena, for example a disease when compared to healthy state. Where as B) represents a similar experimental set up with more experimental variables such as perturbation of a drug or a chemical.

2.2.5 Data input and comparison

bioCompendium provides dynamic HTML elements with advanced JavaScript methods to input the data through a web form as shown in Figure 2.6. It also provides an API which is described in the 'bioCompendium API' subsection in detail.

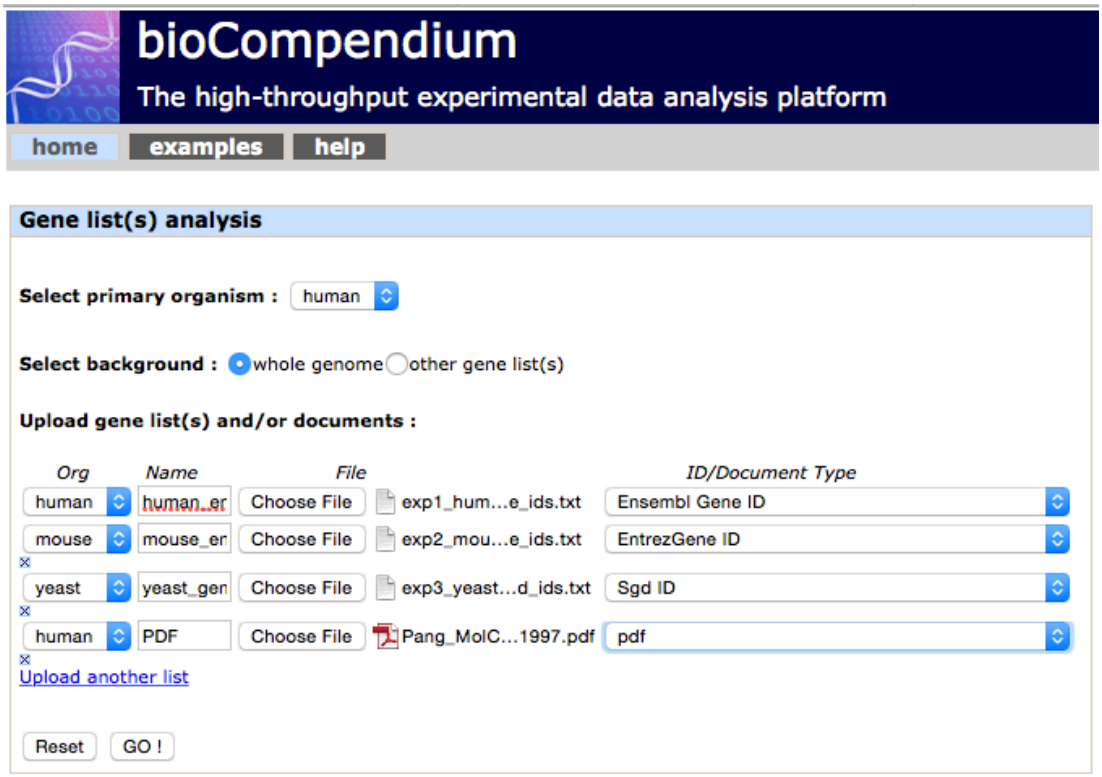


Figure 2.6: **The bioCompendium gene list(s) analysis data input form**. The figure shows different elements of data input form starting from selection of primary organism, background selection, upload of gene list(s) and each gene list parameters like name of organism, gene list, place holder to upload gene list, type of identifier or document.

Primary organism: A user can select one of the three organisms (human, mouse and yeast) currently integrated in the bioCompendium as a primary organism. All the uploaded gene list(s) will be mapped to this selected primary organism via the orthologous relationships.

Statistical background selection: In bioCompendium the user has the choice to select either whole genome or other gene lists as background and can analyze a single gene list or a few gene lists. In case of a single gene list, the background will be the whole genome, because in this case there is no other gene list. More precisely it is the whole genome minus the given gene list as background. But in the case of several gene lists, user can select either one of the two background choices. Here also the query gene list is subtracted from the whole genome or the other gene list(s).

This background data set is used to calculate how significant the selected features are compared to the random selection of elements from the background as mentioned in the pathway and GO enrichment subsections (see subsections 2.3.7 2.3.9 respectively).

Upload gene lists/documents and session management: With the help of the available dynamic file upload functionality, the user can add or remove files easily. JavaScript dynamically generates the list items for the top down menu offered by 'ID/Document type' filed upon the selection of the respective organism. Generally there is no limit except browser timeout, on the number of experimental results e.g., gene list(s) that can be analyzed at a time. For the first 10 gene lists, I am pre-calculating the comparisons and showing the results in bar diagrams. The later visualize how many genes are in each gene list and how many genes are common across different gene lists in different combinations. For the gene lists after 10, the user has to specify the query and the bioCompendium will generate the bar diagram specific to that query, this query functionality need to be implemented. One can further explore the biological information behind each subset represented by colored bars as in Figure 2.9(i).

Once the data is uploaded, bioCompendium creates a temporary session with a unique session ID and processes each gene list. Initially each list is mapped to the corresponding organism specific Ensembl gene IDs with the help of 'ID conversion service' as mentioned above. Later these organism specific Ensembl gene IDs are mapped to the selected primary organism specific Ensembl gene IDs as described in the 'Primary organism' section. If the user gives a document as input, the bio-entities (genes/proteins) mentioned in that document is obtained with the help

of Reflect API. These genes are also mapped to the primary organism specific Ensembl gene IDs. So at this stage all the input gene lists or documents are boiled down to one common format i.e., list of Ensembl gene IDs.

bioCompendium core knowledge obtained as a result of scanning several biological databases is stored in a MySQL database as shown in the Figure 2.3. If each gene from each list (typical size ranging from few to few thousands) is queried separately for the core information, the number of queries will be enormous and it will be very inefficient. In order to address this issue, I'm creating a temporary database in MySQL for each session with session ID as a database name. Each gene list is a temporary table in that database and now bioCompendium core knowledge is obtained for whole gene list in a single SQL join query.

The initial user input (gene lists and documents), the files generated after conversion to common Ensembl format, files that are necessary to create temporary MySQL database and tables, all are stored in the temporary session. Further bioinformatics analysis results obtained for the given input are also stored in the same session directory. These temporary session directories and databases are removed at regular intervals with a cron job.

Lets take a synthetic example as shown in Figure 2.7, a biologist studying a disease in a human model and got the list of significantly relevant genes to that particular disease and represented them as "human Ensembl genes" and he/she want to compare the already published data, but in different model organisms. The primary organism in this example is human and so all other datasets will be converted to human. The second dataset is from mouse and is represented as 'mouse Entrez Gene identifers', third dataset is from yeast and is represented as 'Yeast SGD identifiers' and the fourth dataset is a PDF file related to the same disease model in human. At this point all the datasets are heterogeneous and are coming from different model organisms and represented with different database identifiers. We can not compare them at this stage and the ID (identifier) conversion service (more details in section 2.2.1.2) convert different ID types into organism specific Ensembl genes and the documents are converted to plain text initially and extract the genes/proteins mentioned in the text using Reflect API (more details in section 2.2.1.3) and mapped them to Ensembl gene identifiers. At this state, all are converted to organism specific Ensembl gene identifiers, but

still we can not compare them and genes which are from other than human in this example are converted to human genes by following orthology information. At this stage all the data sets are converted to human Ensembl gene identifiers, so that we can compare and further analyze them.

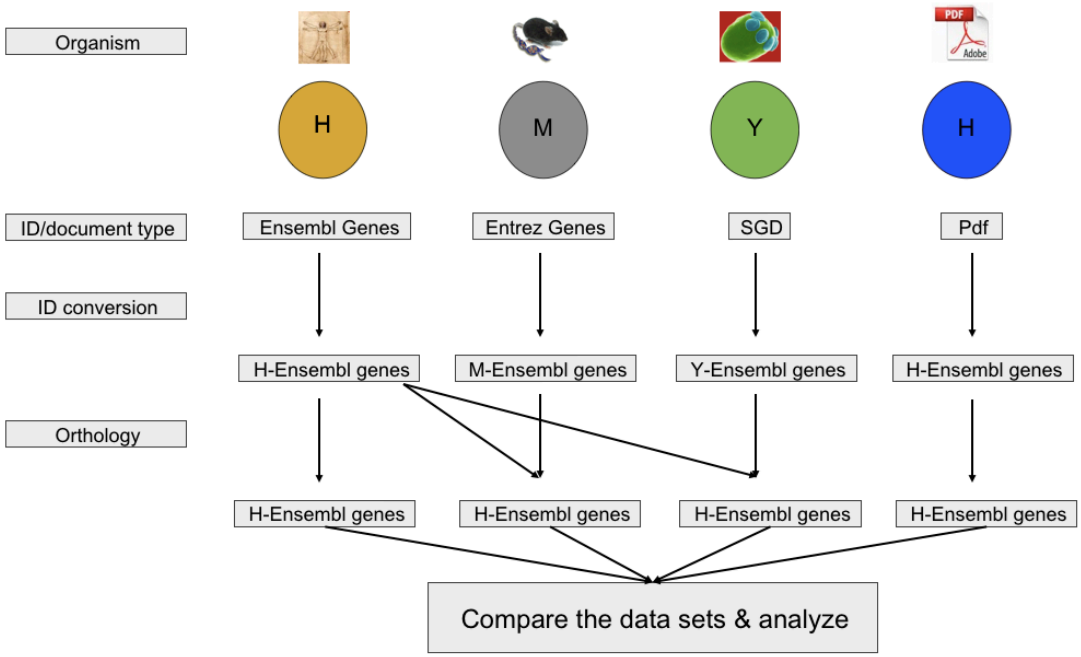


Figure 2.7: **The data input example**. The figure outlines the schematic representation of an example data input consisting of four heterogeneous datasets from human (H), mouse (M), yeast (Y) and a PDF from human. These datasets are represented with different ID/document types, 'Ensembl gene IDs, Entrez Gene IDs, SGD IDs and PDF' respectively. The ID conversion service converts these heterogenous IDs into organism specific Ensembl gene IDs and Reflect API extract the gene/proteins from PDF file after converting it to plain text and mapped them to organism specific Ensembl gene IDs. With the help of orthology information all the different organism specific Ensembl gene IDs mapped to the primary organism specific Ensembl gene IDs. All the datasets are boil down to Ensembl gene IDs and can be compared and further analyzed.

Up to four or five datasets one can easily represent them as venn diagrams, beyond that it will be bit difficult to show as venn diagrams. That is why I'm showing the comparisons as bar diagrams as shown in Figure 2.8.

A

#	Name	Size	Organism
1	exp1_human_ensembl_gene_ids	50	human
2	exp2_mouse_entrezgene_ids	41	mouse->human
3	exp3_yeast_sgd_ids	13	yeast->human
4	Pang_MolCellBiol1997	37	human

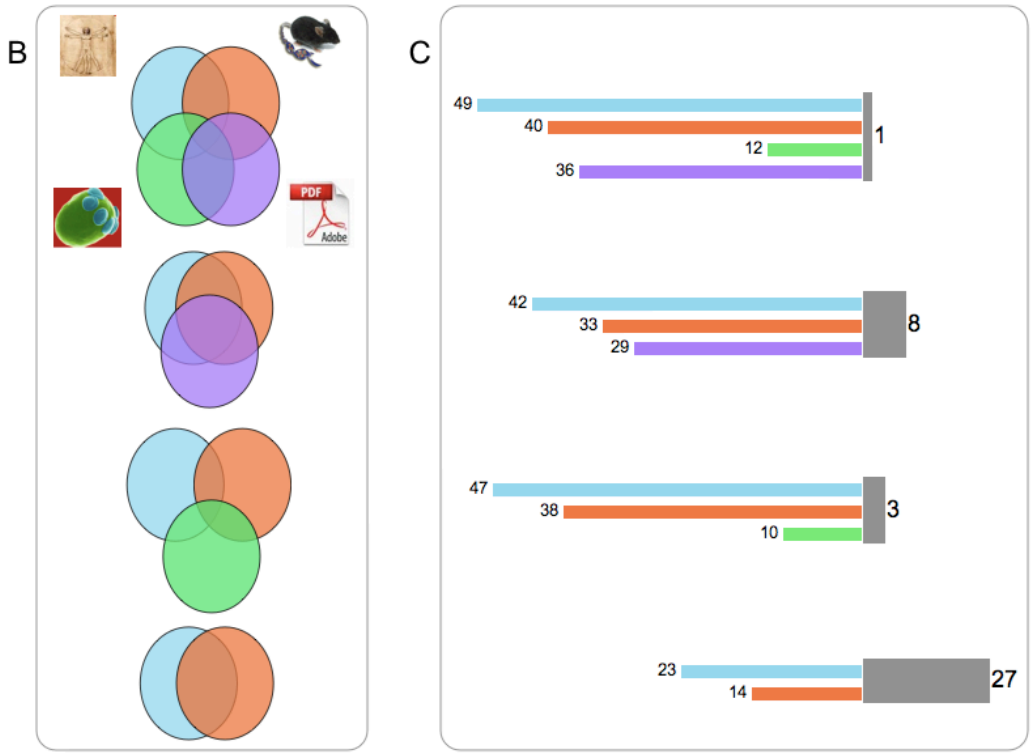


Figure 2.8: **Venn diagrams versus bar diagrams**. The figure A) depicts a table with the information after the conversion of the input datasets to human specific Ensembl gene identifiers. It shows names of the datasets, their size after conversion, a colored bar for each dataset and these colored bars helps in the visualization of the combinations in B) and C), to easily navigate each dataset. It also provides the organism names, the second dataset organism name is 'mouse ->human', this means the original dataset is obtained from mouse and is mapped to human genes by orthology, similarly the third dataset as well i.e 'yeast ->human'. Some of the combinations were shown as venn diagrams in B), each circle represent a dataset and color coding similar to the bars in the table and in bar diagrams in C). As in bioCompendium user can enter any number of datasets and generate the combinations between first ten datasets because of total number of combinations (more details in section 3.1.5), it will be difficult show them as venn diagrams that is why in the combinations were depicted as bar diagrams as shown in C). Here 'intersection' set was shown in grey colored bar, where as 'unique only' dataset shown with respected colored bar and their size.

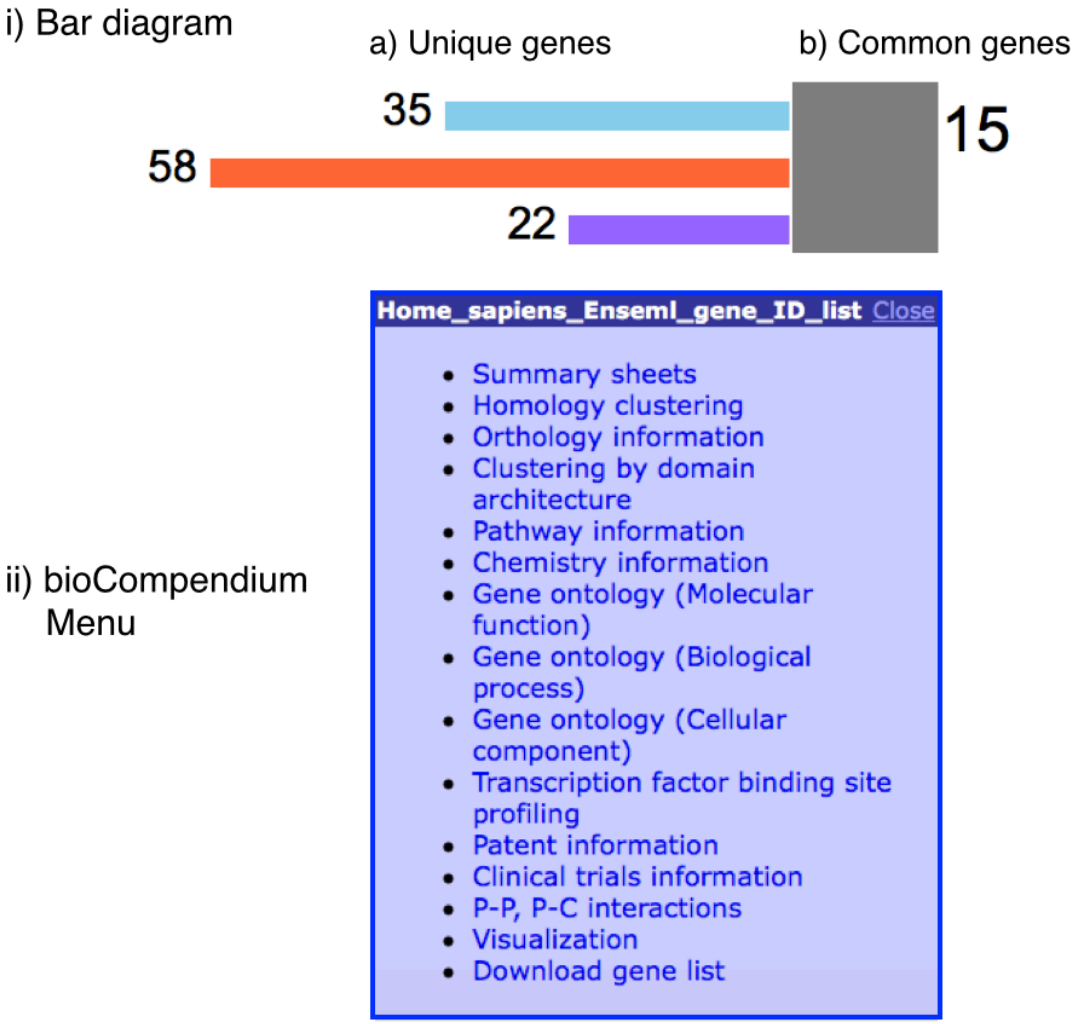


Figure 2.9: **The bar diagram and menu in bioCompendium**. The figure shows an example bar diagram (i) that results from processing of 3 gene lists. The unique gene are represented by a colored bar. The colors of these bars are randomly selected and the length corresponds to the gene set size (i)(a). The intersection part (the common gene set) is represented as grey bar (i)(b). By clicking on any of these bars or size numbers, a menu opens as shown in (ii). By following each hyperlinked menu item, one can view bioinformatics results obtained from that particular method or section.

2.2.6 Bar diagrams and menu

In bioCompendium the user can submit any number of gene lists as input. The overlapping and unique gene sets are represented as bar diagrams as shown in Figure 2.9(i), instead of venn diagrams. Up to 5 or 6 gene lists can be represented as venn diagrams, but beyond that it will be difficult to represent as venn diagram. bioCompendium calculates all possible combinations for up to 10 uploaded gene lists. For example, if the user gives 4 gene lists as input, then it generates all 11 possible combinations up to pairs as represented in the equation 2.1 and calculates overlapping genes and represents the results as bar diagrams. An example of such a bar digram is shown in Figure 2.9(i).

$$\sum_{r=2}^{n} ({}^{n}_{r}\mathrm{C}) = \sum_{r=2}^{n} \left(\frac{n!}{(n-r)!\,(r!)} \right) \tag{2.1}$$

Each gene set from the Figure 2.9(i) can be further explored for different bioinformatics analysis results by clicking the colored bar. This will generates small pop up with the menu shown in Figure 2.9(ii). The materials and methods used for each of these menu items are detailed in the following subsections and the corresponding results are also discussed in the subsequent sections.

2.3 bioCompendium functionality

2.3.1 Single gene list analysis

The functionality of bioCompendium is explained with the following examples. Lets take a microarray expression dataset for diabetes mellitus (GEO Accession number : GSE25724) (Dominguez et al., 2011). The raw data has been downloaded from the Gene Expression Omnibus (GEO) (Barrett et al., 2012) and analyzed with LIMMA R package (Smyth, 2004). The differentially expressed genes between type 2 diabetic human islets and non-diabetic islets samples with p-value ≤ 0.01 and log2 fold change ≥ 1 were selected for bioCompendium analysis.

The first step is to go to the bioCompendium web interface as shown in Figure

2.1, and select the 'human' as a primary organism. At this point user can select one of the other organisms as a primary organism. Depending on the organism selected the gene list(s) will be converted to the selected primary organism by following the orthology relations. Next, select the background, here user has two possibilities: 1) whole genome and 2) other gene list(s). In this current example as we have only one gene list, the background will be by default 'whole genome', because we don't have any other gene list. Even if the user selects by mistake the second option i.e 'other gene list(s)', it will set back to 'whole genome'.

The second step upload the gene list. The list should contain only one type of identifiers. First enter the organism name, which in our current example is 'human'. Then enter a name. The user can enter any name but it would be useful if it represents the gene list or the experiment. After loading the corresponding gene list file, select the appropriate identifier type or document type in case user uploaded a document (PDF, MS-Word, Excel, PowerPoint, plain text). The differentially expressed genes from GSE25724 were Ensembl gene identifiers, so 'ID/Document Type' is 'Ensembl Gene ID' and then submit the job. This will open an initial results interface as shown in Figure 2.10 and it shows the background information selected during data input step and a table that provides gene list name, its size after mapping to the Ensembl genes, a colored bar and the organism name. The user can open a menu to view the bioCompendium analysis results by clicking on the colored bar. This menu will be explained in greater detail in the next sections.

At this point in the backend server side, a session directory is created and stores all the input details and gene list(s) in that directory. The input details go into a metafile, which will facilitate to reload the session later. Generally a temporary session is created and is deleted after two days due to limited resources. I also provide permanent sessions on demand. It also creates a temporary database with tables in MySQL with the same session identifier and loads the data into this database by using ETL (Extraction, Transform and Loading) scripts.

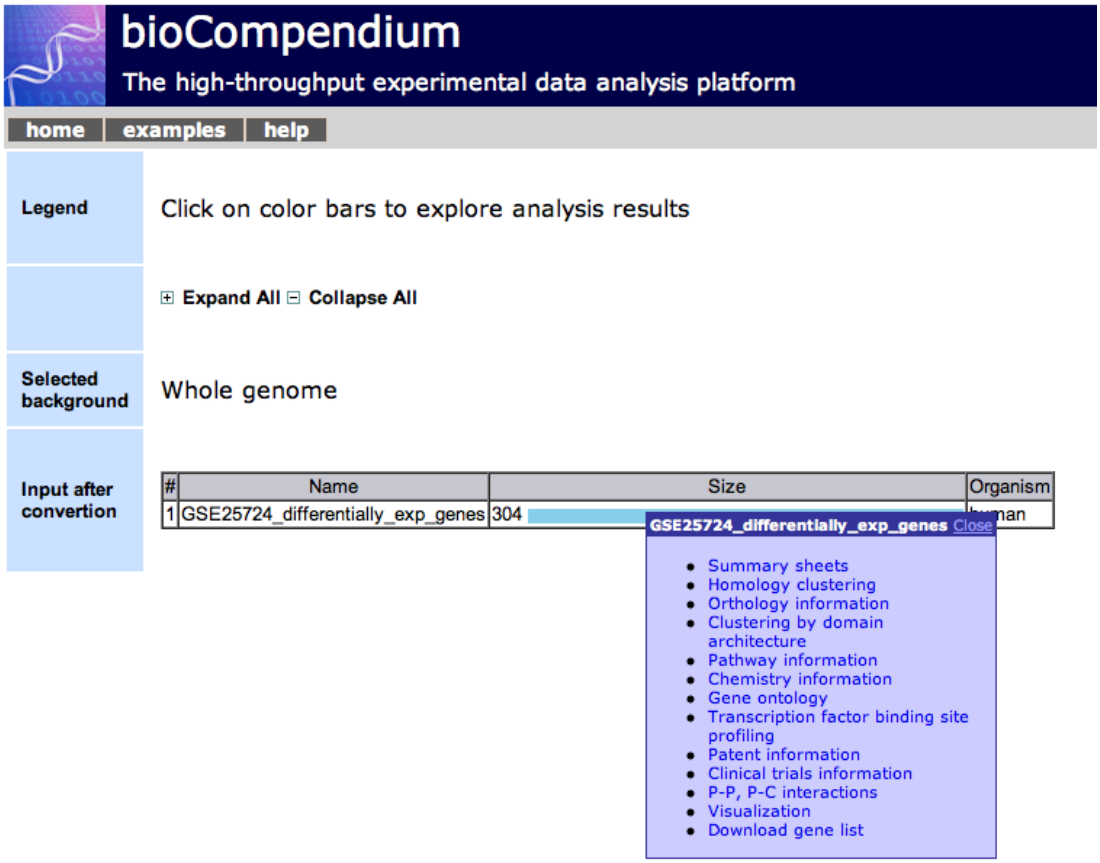


Figure 2.10: **The bioCompendium single gene list result interface**. The figure shows the initial results interface for a single gene list. It shows selected background and a table with gene list name, its size, a colored bar after mapping the selected identifiers to 'Ensembl Genes' and the organism name from which the gene list is obtained. By clicking on the colored bar a small floating window appears that provides a menu to view the bioCompendium analysis results.

2.3.2 Multiple gene list analysis

In the above example, we have seen how to analyze a single gene list from a transcriptomics dataset for diabetes mellitus (GEO Accession : GSE25724) (Dominguez et al., 2011) and now we will focus on analysis of multiple gene lists. Let's compare the differentially expressed genes from the experiment GSE25724 with a proteomics experiment and a PDF paper related to type 2 diabetes mellitus. The proteomics study is about discovery of the novel glucose-regulated proteins from

human pancreatic islets using LC-MS/MS based proteomics (Schrimpe-Rutledge et al., 2012) and the PDF paper is related to the same disease but the experiment is done in a mouse model (de Wilde et al., 2008), whereas transcriptomics and proteomics studies are done in human models. The transcriptomics data is presented as 'Ensembl genes', whereas proteomics data as 'IPI identifiers' and PDF will be processed to extract genes/proteins mentioned in the free text by using Reflect API. In this example the IPI IDs are mapped to human Ensembl gene IDs by using ID conversion service and the genes/proteins obtained from PDF paper are related to mouse species and are mapped to corresponding human genes by following orthology relationships.

The bioCompendium multiple gene list initial result interface is shown in Figure 2.11. By clicking on any colored bar it will open a menu with hyperlinks that will provide analysis results for that selected dataset. We will explore each menu item one by one in the next sections.

2.3.3 Summary sheets

Biological knowledge acquired from various databases (refer to section 2.2.1.1) for the selected gene list is displayed in a sortable table view. This section provides gene name, description and links to the summary sheet and Ensembl gene entry page. The summary sheet displays this knowledge in different sections labeled as genes, proteins, domains, structural features, diseases, gene ontology (GO), pathways, chemicals, interactions and orthology. The information of each section can be further explored, thus giving researchers the opportunity to go deeper into the volume of knowledge in a more efficient way. All the information is presented under the same web page, which not only makes the exploration easier but it also reduces dramatically the loss of time that occurs if someone needs to find the information by browsing and querying several available databases individually. For example, if one checks the information in the section 'Proteins', (s)he can see that it comes from four resources Ensembl Proteins, Entrez Proteins, RefSeq Proteins and UniPort. The information in these four resources related to the protein of interest is either the same or complementary. Furthermore, a Dasty2 (Jimenez et al., 2008) DAS client provides protein features in a nice

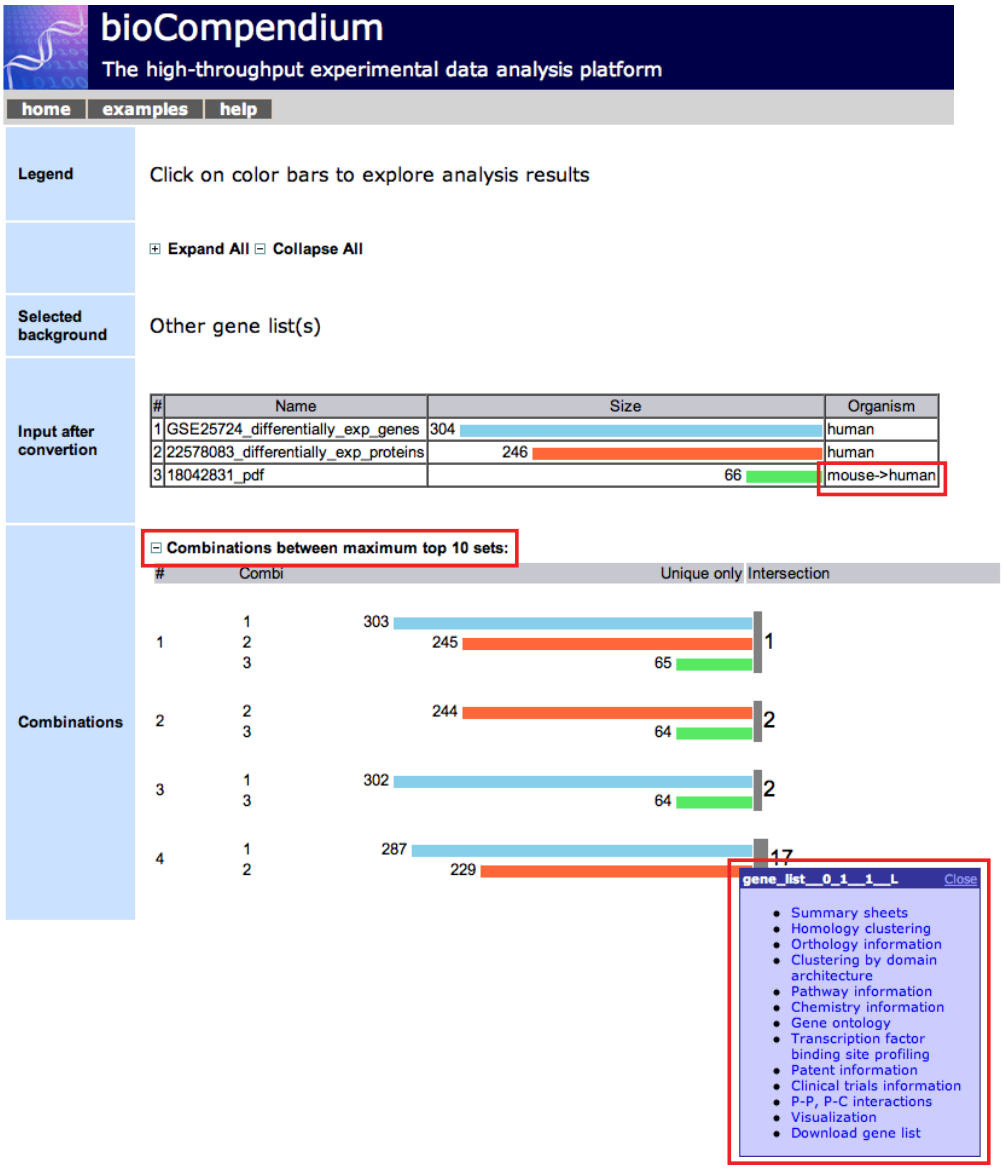


Figure 2.11: **The bioCompendium multiple gene list result interface**. The figure shows the initial results interface when a user uploads more than one gene list. It shows the selected background and a table with gene list names, their size, colored bars after mapping the selected identifiers to 'Ensembl Genes' and the organism name from which the gene lists are obtained. In this example the third gene list obtained from a PDF file and its original organism is 'mouse', but the primary organism is 'human' that is why the organism is mentioned 'mouse -> human' which means the mouse genes are mapped to human genes by following orthology relationships and are highlighted with a red rectangle. Next section show the combinations between the input gene lists, which can be visualized as bar diagrams by clicking on the '+ Combinations between maximum top 10 sets' button highlighted with a red rectangle. By clicking on any colored bar a small floating window appears that provides a menu to view the bioCompendium analysis results.

graphical representation and also provides the sequences as well.

Summary sheets give users access to a variety of correlated information about an entity (gene or protein) in a simple interface. Instead of spending time looking in multiple places for information, users can open a summary sheet and find all the relevant information organized in one location.

Description: Information from various biological sources in a short tabular view

No of Genes 304

Legend : *The following table is sortable*

Name	Description	Gene
AASDHPPT	L-aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase (EC 2.7.8.-) (4'-phosphopantetheinyl transferase) (Alpha-aminoadipic semialdehyde dehydrogenase-phosphopantetheinyl transferase) [Source:UniProtKB/Swiss-Prot;Acc:Q9NRN7]	ENSG00000149313
AC005000.2	SMT3 suppressor of mif two 3 homolog 2 (S. cerevisiae) (SUMO2), transcript variant 2, mRNA [Source:RefSeq_dna;Acc:NM_001005849]	ENSG00000188612
AC016940.7	Protein CXorf40A (Endothelial-overexpressed lipopolysaccharide-associated factor 1) [Source:UniProtKB/Swiss-Prot;Acc:Q8TE69]	ENSG00000197620
AC025917.8	cAMP-regulated phosphoprotein 19 (ARPP-19) [Source:UniProtKB/Swiss-Prot;Acc:P56211]	ENSG00000128989
AC079061.8	Syntabulin (Syntaxin-1-binding protein) (Golgi-localized syntaphilin-related protein) [Source:UniProtKB/Swiss-Prot;Acc:Q9NX95]	ENSG00000147642
AC100778.5		ENSG00000215472
ACP1	Low molecular weight phosphotyrosine protein phosphatase (EC 3.1.3.48) (EC 3.1.3.2) (LMW-PTPase) (LMW-PTP) (Low molecular weight cytosolic acid phosphatase) (Red cell acid phosphatase 1) [Source:UniProtKB/Swiss-Prot;Acc:P24666]	ENSG00000143727
ACSL1	Long-chain-fatty-acid--CoA ligase 1 (EC 6.2.1.3) (Long-chain acyl-CoA synthetase 1) (LACS 1) (Long-chain acyl-CoA synthetase 2) (LACS 2) (Palmitoyl-CoA ligase 1) [Source:UniProtKB/Swiss-Prot;Acc:P33121]	ENSG00000151726
ADSS	Adenylosuccinate synthetase isozyme 2 (EC 6.3.4.4) (AdSS 2) (Adenylosuccinate synthetase, acidic isozyme) (IMP--aspartate ligase 2) (AMPSase 2) [Source:UniProtKB/Swiss-Prot;Acc:P30520]	ENSG00000035687
AHCY	Adenosylhomocysteinase (EC 3.3.1.1) (AdoHcyase) (S-adenosyl-L-homocysteine hydrolase) [Source:UniProtKB/Swiss-Prot;Acc:P23526]	ENSG00000101444
AKAP11	A-kinase anchor protein 11 (Protein kinase A-anchoring protein 11) (PRKA11) (A kinase anchor protein 220 kDa) (AKAP 220) (hAKAP220) [Source:UniProtKB/Swiss-Prot;Acc:Q9UKA4]	ENSG00000023516

Figure 2.12: **The Summary sheet table view**. The figure shows a sortable table view with gene name, description and its corresponding Ensembl gene identifier. By clicking on the gene names it will open a detailed summary sheet and by clicking on Ensembl gene ID, it will take the user to the Ensembl resource.

Summary sheets can be launched from the menu by clicking the bar diagrams shown in Figures 2.8, 2.9, 2.10, 2.11 or from the simple search shown in Figure 2.4. In both cases, bioCompendium creates a sortable table view shown in Figure 2.12 with gene name, description and Ensembl gene. A summary sheet will be opened by following the gene name hyper link. Once opened, the content is organized in different subsections - name, description, genes, proteins, domains, structural information, diseases involved, related Gene Ontology terms, pathways involved, drugs, ligands, metabolities, other chemicals, protein-protein, protein-chemical interactions, and orthology relationships to all Ensembl model organisms. One can easily navigate the information in the summary sheet with the help of '+ **Expand All**' and '**- Collapse All**' buttons (highlighted with red rectangles in the Figure 2.13).

An example of such a summary sheet for a gene "SNCA" involved in Parkinson's disease is shown in the Figure 2.13.

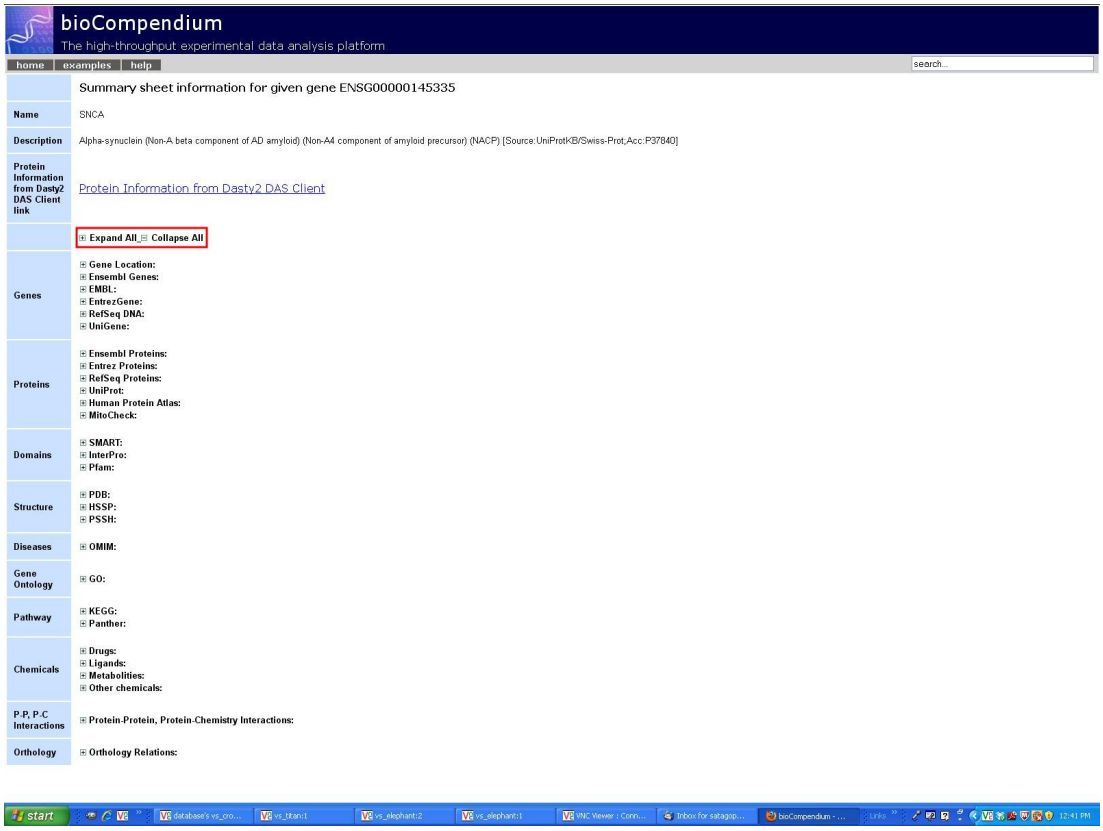


Figure 2.13: **The summary sheet example**. The figure shows the collapsed view of summary sheet for a gene "SNCA". The information is grouped in several subsections and can be easily explored by toggle between '+', '-' buttons and all the subsections can be expanded and collapsed by using '**+ Expand All**' and '**- Collapse All**' buttons, highlighted with red rectangle.

2.3.4 Homology clustering

In this section, the selected list of proteins were clustered by sequence similarity. Initially protein sequences were collected for all the proteins and 'all against all blast' (Gish, 1996), (Lopez et al., 2003) search was carried out to obtain the sequence similarity. Generally blast jobs take longer to calculate the sequence similarity. Therefore, in bioCompendium 'all against all blast' for whole proteome was calculated in advance and the resulting pairwise similarity matrix is stored into the main memory, which accelerates the homology clustering analysis. A Remote Procedure Call (RPC) based method is implemented to access the pair-

wise similarities for the given input of genes/proteins from the whole proteome similarity matrix.

The Markov Clustering Algorithm (MCL) is used to obtain the sequence similarity clusters from the above mentioned pairwise sequence similarity matrix. MCL is an unsupervised clustering algorithm for networks based on the stochastic flow in graphs (Enright et al., 2002). I have used default inflation parameter (-I) value 2, this parameter controls the granularity of the output clustering. An example of such clusters for the GSE25724 type 2 diabetes dataset is shown in Figure 2.14. The clusters were sorted by their size, largest at the top. Elements of each cluster can be explored with the help of '+' and '-' buttons (highlighted with red rectangles in the Figure 2.14).

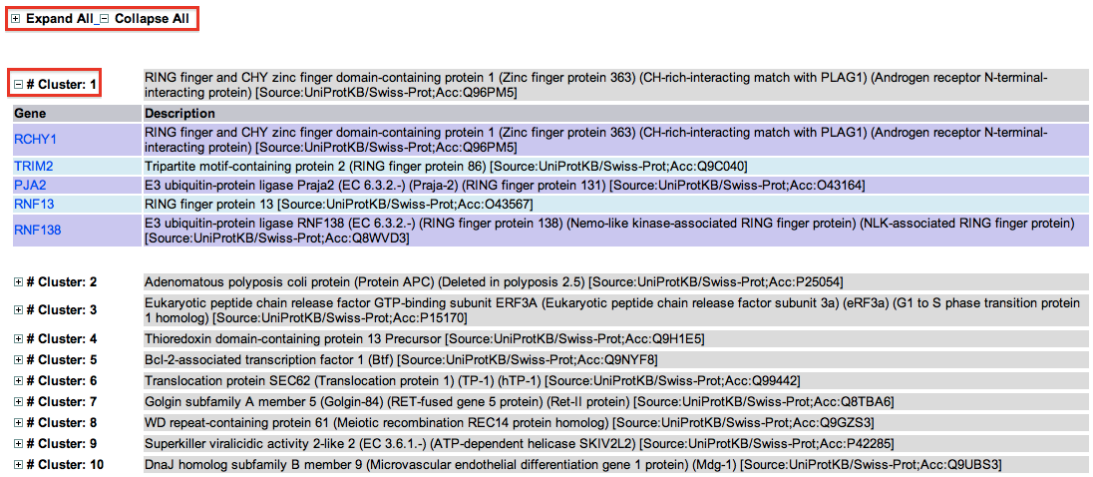


Figure 2.14: **Homology clustering example**. The figure shows the overview of the first 10 clusters based on sequence similarity and the first cluster is showing its elements - all are 'RING finger' proteins.

2.3.5 Orthology information

This section provides orthologs to all other Ensembl genomes for the given dataset in a table view and it also details the type of orthology relationships. I am using same orthology information for the cross species comparisons. The Figure 2.15 shows the overview of orthology results for given input gene list and upon clicking

on '+' and '-' buttons, othologs for each gene can be explored, an example of such information is shown in Figure 2.16.

2.3.6 Clustering by domain architecture

Protein domain information for each protein is obtained from Pfam, SMART, InterPro and PRINTS. By having special access to the SMART backend database and domain graphical representations, SMART domains were chosen for clustering based on domain architecture. The pairwise domain architecture similarity between proteins was calculated by using tf-idf (term frequency - inverse document frequency) method (Salton and Buckley, 1988), which is widely used to find out document similarity in text-mining, as outlined below.

Term frequency (tf): The $tf_{d,p}$ of a domain d in protein p is defined as the number of occurrences of domain d in protein p. This value is normalized by dividing by the total number of domains in that protein in order to prevent a bias towards longer proteins with many domains. The long proteins have a higher domain count regardless of the actual importance of that particular domain in the protein. Normalization gives a measure of importance of the domain d within the particular protein p.

Inverse document frequency (idf): The df_p, the document frequency of domain d is the numbers of proteins containing the domain d. The inverse document frequency, $idf_{d,P}$ is obtained by dividing the total number of proteins P in the proteome by the document frequency df_p, and then taking the logarithm of the value.

$$idf_{d,P} = log\left(\frac{P}{df_p}\right) \tag{2.2}$$

This inverse document frequency is a measure of whether the domain is common or rare across the proteome.

tf-idf weight of a domain is the product of $tf_{d,p}$ value and $idf_{d,P}$ value.

$$(tf - idf)_{d,p} = (tf_{d,p} * idf_{d,P}) \tag{2.3}$$

Each protein is now represented by a real-valued vector of tf-idf weights $\in$ $R^{|V|}$. This results in a $|V|$ dimensional vector space, where domains are axes of

Figure 2.15: **bioCompendium orthology overview**. The figure shows the overview of orthology information for a given list of genes. Initially it shows only the gene name and description, upon clicking '+' button, orthology information can be explored.

Orthology Relations:

−MLH3	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]	
Organism	**Orthology type***	**Orthologs**	**Orthologs Description**
Pan_troglodytes	one2one	MLH3	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Pongo_pygmaeus	one2one	ENSPPYG00000005997	MHL3 (Fragment). [Source:UniProtKB/TrEMBL;Acc:A0EJJ1]
Macaca_mulatta	one2one	ENSMMUG00000002911	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Microcebus_murinus	one2one	ENSMICG00000006104	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Tarsius_syrichta	one2one	ENSTSYG00000000754	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Equus_caballus	one2one	LOC100057292	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Pteropus_vampyrus	one2one	ENSPVAG00000000664	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Otolemur_garnettii	one2one	ENSOGAG00000008240	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Tursiops_truncatus	one2one	ENSTTRG00000001923	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Spermophilus_tridecemlineatus	one2one	ENSSTOG00000007304	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Myotis_lucifugus	one2one	ENSMLUG00000005949	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Felis_catus	one2one	ENSFCAG00000003082	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Vicugna_pacos	one2one	ENSVPAG00000003084	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Canis_familiaris	one2one	MLH3	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Bos_taurus	one2one	MLH3	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Dasypus_novemcinctus	one2one	ENSDNOG00000001959	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Tupaia_belangeri	one2one	ENSTBEG00000000655	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Oryctolagus_cuniculus	one2one	ENSOCUG00000010182	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Cavia_porcellus	one2one	ENSCPOG00000024679	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Dipodomys_ordii	one2one	ENSDORG00000004893	mutL homolog 3 [Source:RefSeq_peptide;Acc:NP_780546]
Procavia_capensis	one2one	ENSPCAG00000008569	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Echinops_telfairi	one2one	ENSETEG00000010598	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Ochotona_princeps	one2one	ENSOPRG00000011908	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Mus_musculus	one2one	Mlh3	mutL homolog 3 [Source:RefSeq_peptide;Acc:NP_780546]
Rattus_norvegicus	one2one	ENSRNOG00000006699	mutL homolog 3 [Source:RefSeq_peptide;Acc:NP_780546]
Sorex_araneus	one2one	ENSSARG00000010893	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Erinaceus_europaeus	one2one	ENSEEUG00000014034	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Monodelphis_domestica	one2one	ENSMODG00000006037	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
Gallus_gallus	one2one	ENSGALG00000010304	DNA mismatch repair protein Mlh3 (MutL protein homolog 3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UHC1]
+MSX2			Homeobox protein MSX-2 (Hox-8) [Source:UniProtKB/Swiss-Prot;Acc:P35548]
+PMS2			Mismatch repair endonuclease PMS2 (EC 3.1.-.-) (PMS1 protein homolog 2) (DNA mismatch repair protein PMS2) [Source:UniProtKB/Swiss-Prot;Acc:P54278]
+ATP7B			Copper-transporting ATPase 2 (EC 3.6.3.4) (Copper pump 2) (Wilson disease-associated protein)(WND/140 kDa) [Source:UniProtKB/Swiss-Prot;Acc:P35670]
+SYMPK			Symplekin [Source:UniProtKB/Swiss-Prot;Acc:Q92797]
+ATP6V1F			Vacuolar proton pump subunit F (EC 3.6.3.14) (V-ATPase subunit F) (V-ATPase 14 kDa subunit) [Source:UniProtKB/Swiss-Prot;Acc:Q16864]
+ATP1B2			Sodium/potassium-transporting ATPase subunit beta-2 (Sodium/potassium-dependent ATPase subunit beta-2) [Source:UniProtKB/Swiss-Prot;Acc:P14415]
+ATP6V1E1			Vacuolar proton pump subunit E 1 (EC 3.6.3.14) (V-ATPase subunit E 1) (V-ATPase 31 kDa subunit) (P31) [Source:UniProtKB/Swiss-Prot;Acc:P36543]
+RAF1			RAF proto-oncogene serine/threonine-protein kinase (EC 2.7.11.1) (C-RAF) (cRaf) (Raf-1) [Source:UniProtKB/Swiss-Prot;Acc:P04049]

Figure 2.16: **bioCompendium orthology results example**. The figure shows the orthology information for a gene 'MLH3' in a tabular format consisting of organism name, type of orthology based on the protein trees (Vilella et al., 2009), (Ensembl, 2009) ortholog name and corresponding ortholog description.

the space and proteins are points or vectors in this space. Domain architecture similarity between pairs of proteins (p,q) is obtained by normalizing each vector by its length that results in unit vectors and is followed by calculating the inner product of the two unit vectors.

$$cos(p,q) = \left(\frac{p}{|p|} \bullet \frac{q}{|q|}\right) = \left(\frac{\sum_{i=1}^{|V|} p_i q_i}{\sqrt{\sum_{i=1}^{|V|} p_i^2}\sqrt{\sum_{i=1}^{|V|} q_i^2}}\right) \tag{2.4}$$

In the Equation 2.4, p_i is the td-idf weight of domain i in protein p and q_i is the td-idf weight of domain i in protein q. *cos(p,q)* is the cosine similarity of vectors $\vec{p}$ and $\vec{q}$ i.e cosine of the angle between $\vec{p}$ and $\vec{q}$, which results in the domain architecture similarity between the proteins p and q. The resulting similarity ranges from 0 (no similarity) to 1 (same domain architecture).

By using the above methodology, all-against-all domain architecture similarities were calculated. This task is also carried out in advance, similar to the 'Homology clustering' and the resulted pairwise similarity matrix for whole proteome is stored in the main memory. A Remote Procedure Call (RPC) based method is implemented to access the pairwise similarities for the given input of genes/proteins.

The Markov Clustering Algorithm (MCL) (Enright et al., 2002) mentioned above was used to obtain the protein clusters from the above mentioned pairwise domain architecture similarity matrix.

A 2D graphical view is implemented for each cluster obtained above using SMART domain representations, DHTML, JavaScript and CSS. Figure 2.17 shows an example of such clustering results. It lists the proteins of each cluster in a tabular view and also provides domains and their positions on the protein. The user can get more details of each of the domains by having the mouse over it. An interactive graphical view is implemented for each cluster to highlight how the domains are conserved in these clusters, an example is shown in Figure 2.18.

2.3.7 Pathway information

Pathway enrichments have been calculated by using KEGG (Kanehisa et al., 2012), PANTHER (Mi et al., 2005), Reactome (Croft et al., 2011) pathway

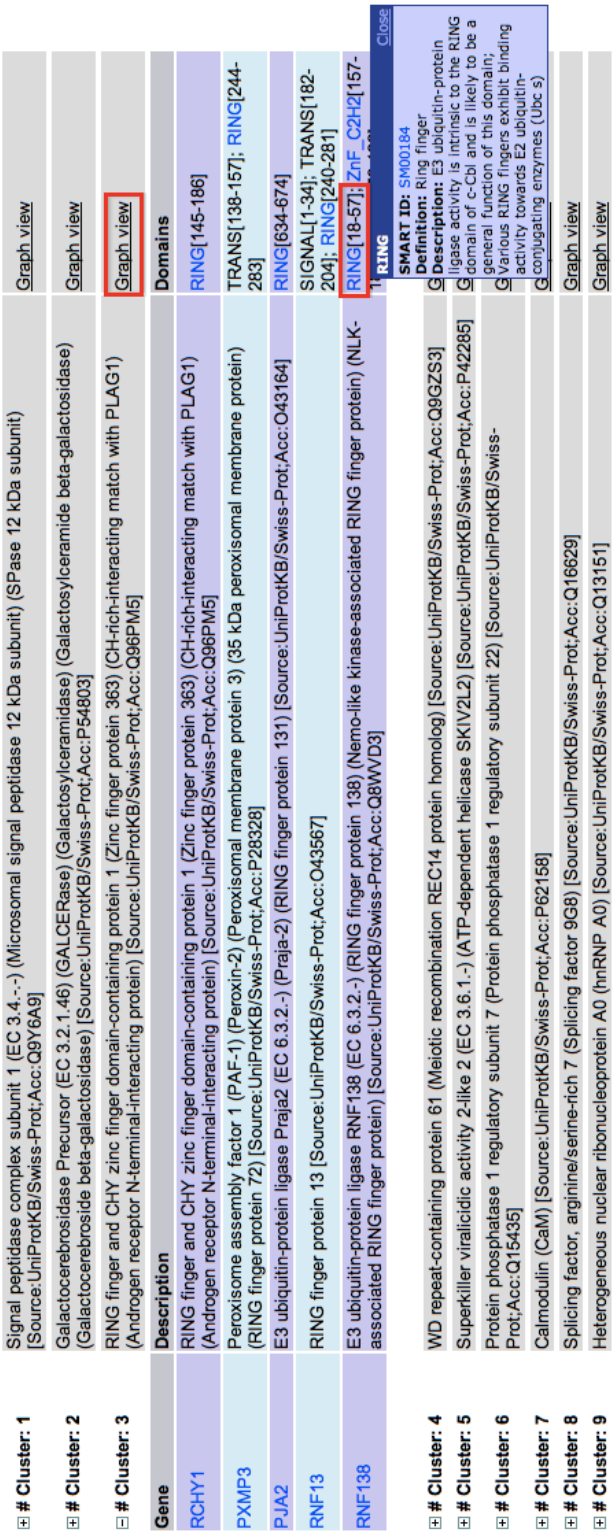


Figure 2.17: **Domain clustering**. The figure shows the example of a domain clustering results for the GSE25724 type 2 diabetes dataset. One of the clusters is showing its elements - all mainly contain 'RING finger' domains. It shows more details of each domain by having mouse over its name and provides graphical view for each cluser.

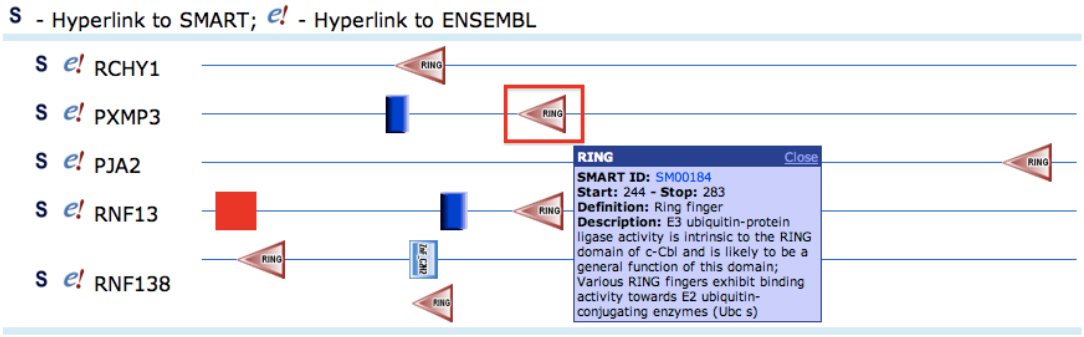


Figure 2.18: **Domain clustering graphical view**. The figure shows the graphical view for the selected cluster with SMART domain representations and provides more details for each domain by having mouse over it (highlighted with red rectangle).

databases for a given list of genes/proteins. Each gene/protein was mapped to each of the above mentioned pathway databases separately in order to find out which pathways are enriched for the selected pathway database. In order to calculate the enrichments, we need a background data set to calculate how significant the enriched pathways are compared to the random selection of the same number of genes/proteins from the background set. In bioCompendium, users can analyze a single list of genes or a few lists of genes. In addition the user can select either whole genome or other uploaded gene list(s) as background (in the case of more than one gene list as input). Two-tailed Fisher's exact test (Fisher, 1922) is used to calculate the p-values from the 2x2 contingency table for each enriched pathway and the q-values (adjusted p-values) were calculated by controlling the False Discovery Rate (FDR) by applying BH (Benjamini-Hochberg) FDR correction method (Benjamini and Hochberg, 1995).

The enriched pathway results are provided in a tabular view. Each pathway is accompanied by a KEGG pathway identifier, corresponding pathway name and a list of proteins involved in that pathway, an example of such results shown in Figure 2.19. Each pathway is hyperlinked to a red flag, which ultimately displays the pathway diagram, an example shown in Figure 2.20 with the proteins from the selected category highlighted in red from KEGG database.

Description							
No of pathways mapped	53						
Information from KEGG pathways - Mapping of given input genes							
KEGG Pathway ID	KEGG Pathway Name	Adjusted P-Value	Gene Name	KEGG Gene	Ensembl Gene		
hsa05110	Vibrio cholerae infection	5.7521e-22	ATP6V0A1	535	ENSG00000033627		
			ATP6V1H	51606	ENSG00000047249		
			ATP6V1B1	525	ENSG00000116039		
			ATP6V1F	9296	ENSG00000128524		
			ATP6V1E1	529	ENSG00000131100		
			ATP6V0D2	245972	ENSG00000147614		
			ATP6V1G3	127124	ENSG00000151418		
			ATP6V1C1	528	ENSG00000155097		
			ATP6V0D1	9114	ENSG00000159720		
			ATP6V1E2	90423	ENSG00000171142		
			ATP6V0A2	23545	ENSG00000185344		
			ATP6V0C	527	ENSG00000185883		
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	3.8980e-21	ATP6V0A1	535	ENSG00000033627		
			ATP6V1H	51606	ENSG00000047249		
			ATP6V1B1	525	ENSG00000116039		
			ATP6V1F	9296	ENSG00000128524		
			ATP6V1E1	529	ENSG00000131100		
			ATP6V0D2	245972	ENSG00000147614		
			ATP6V1G3	127124	ENSG00000151418		
			ATP6V1C1	528	ENSG00000155097		
			ATP6V0D1	9114	ENSG00000159720		
			ATP6V1E2	90423	ENSG00000171142		
			ATP6V0A2	23545	ENSG00000185344		
			ATP6V0C	527	ENSG00000185883		
hsa00190	Oxidative phosphorylation	1.1379e-18	ATP6V0A1	535	ENSG00000033627		
			ATP6V1H	51606	ENSG00000047249		
			ATP12A	479	ENSG00000075673		
			ATP6V1B1	525	ENSG00000116039		
			ATP6V1F	9296	ENSG00000128524		
			ATP6V1E1	529	ENSG00000131100		
			ATP6V0D2	245972	ENSG00000147614		
			ATP6V1G3	127124	ENSG00000151418		
			ATP6V1C1	528	ENSG00000155097		
			ATP6V0D1	9114	ENSG00000159720		
			ATP6V1E2	90423	ENSG00000171142		
			ATP6V0A2	23545	ENSG00000185344		
			ATP6V0C	527	ENSG00000185883		
			ATP4B	496	ENSG00000186009		

Figure 2.19: **Pathway enrichment analysis results**. The figure shows the tabular view of enriched KEGG pathways for given input list of genes, each enriched pathway can be visualized (an example shown in Figure 2.20) by clicking on the red flag (highlighted with red rectangle).

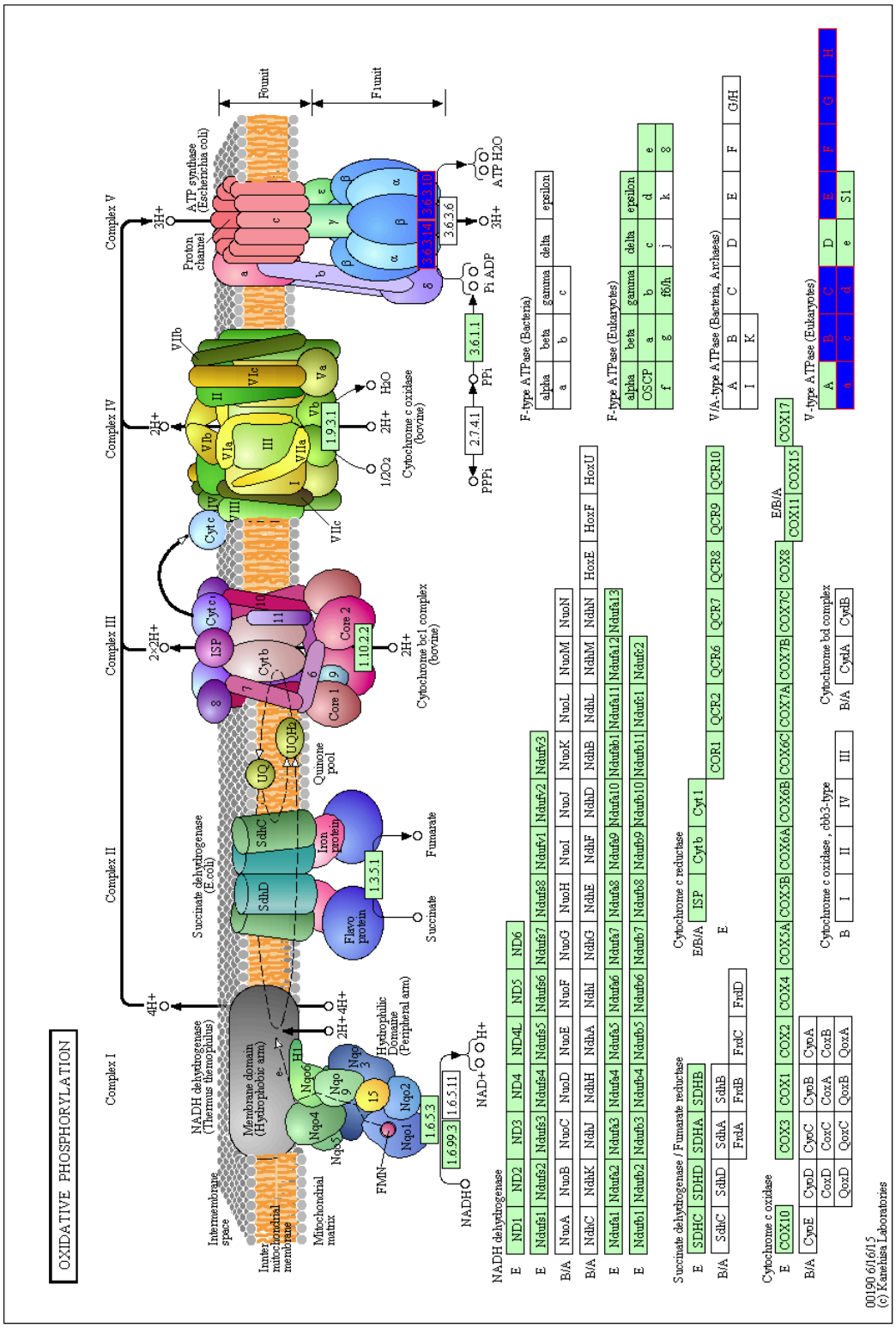


Figure 2.20: **An example enriched KEGG pathway**. The figure shows an enriched KEGG pathway - "Oxidative phosphorylation". The enriched genes from the input gene list are highlighted in blue in the pathway diagram.

2.3.8 Chemistry information

Chemistry information is collected from several public databases DrugBank, HMDB, PubChem, chEBI, STITCH, MATADOR, PDBLigand, and from literature as well by using Alma Knowledge Server 2 (AKS2). All this heterogeneous chemical information is parsed and processed to extract the target proteins and the chemicals are segregated into four sub-categories (i) Drugs, (ii) Ligands, (iii) Metabolites and (iv) Other chemicals. For each chemical, a Structure-Data File (SDF) is obtained from PubChem database, which is one of the largest collection of publicly available chemical databases. Each SDF file contains information about the atoms, bonds and the coordinates of a given chemical (Dalby et al., 1992). A perl script 'mol2png' (Haider, 2012a) is used to generate the 2D chemical structure in PNG format from the SDF files. This is achieved by processing the data through 'mol2ps' (Haider, 2012b) utility program and gs (Ghostscript). Due to the limited availability of chemical and related target (gene/protein) information in the public domain, I have also collected the chemical and related target (gene/protein) information from a text mining resource, Alma Knowledge Server 2 (AKS2). This data has been treated carefully by taking the text-mining relevance score based on number of occurrences into account in order to avoid the false positives. This relevance score in percentage was provided for the corresponding entries and higher the score the better. An example of chemistry information for a target protein 'ATP12A - Potassium transporting ATPase alpha chain 2 (EC 3.6.3.10) (Proton pump)' shown in Figure 2.21.

2.3.9 Gene ontology

Gene Ontology (GO) (Ashburner et al., 2000) enrichments have been obtained for all the three sub-categories of GO - 'Molecular function', 'Cellular component' and 'Biological process'. Methodology of pertinence selection of GO terms as described in the publication (Barriot et al., 2007) have been implemented. It uses a file of Ensembl genes GO annotations and the GO database of GO terms. I have used MySQL dump of GO term database. This program looks for GO terms enrichment or depletion between a given list of genes and background. The background is similar to the one mentioned in the pathway enrichment section

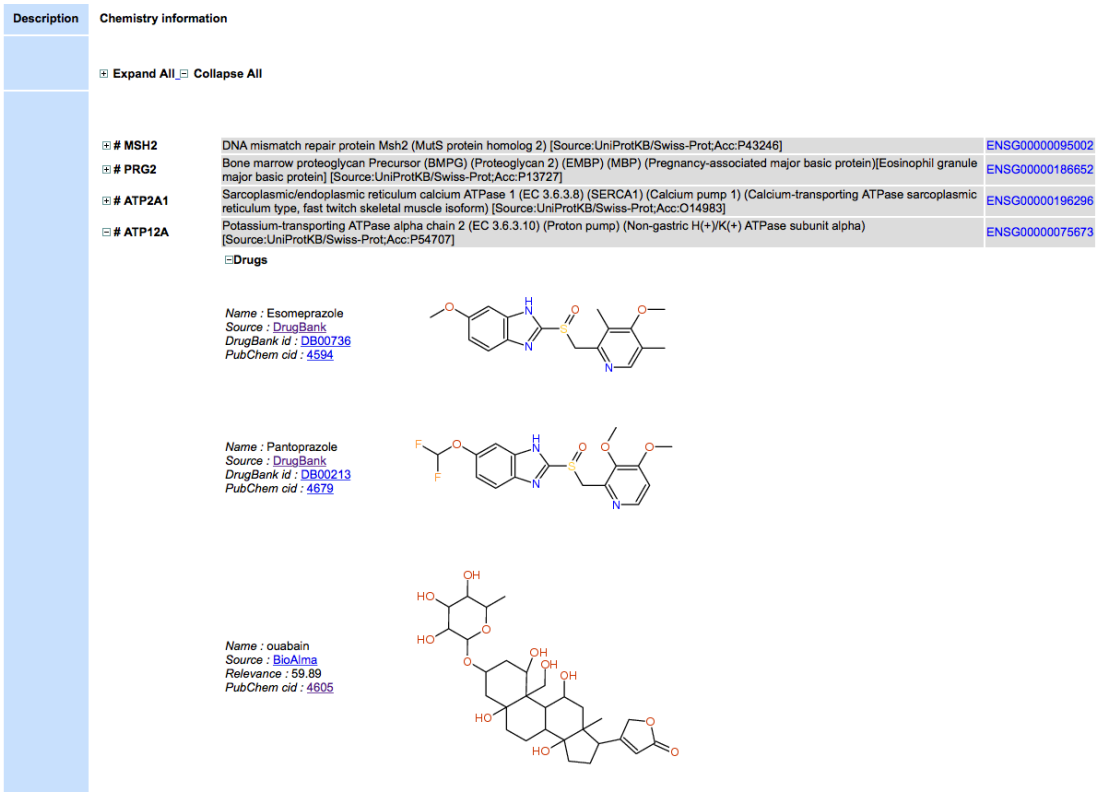


Figure 2.21: **The chemistry information from bioCompendium**. The figure shows the Drugs and Ligands targeting a protein ATP12A - Potassium transporting ATPase alpha chain 2 (EC 3.6.3.10) (Proton pump)'. Here the drugs Esomeprazole, Pantoprazole are highly effective inhibitors of gastric acid secretion used in the therapy of stomach ulcers and zollinger-ellison syndrome.

2.3.7, which is either whole genome or other gene lists. Because GO is an acyclic graph, the child terms were mapped back to the parent terms with the help of GO database. Each node in the graph is mapped to the corresponding genes by using Ensembl genes GO annotations file.

The outline of the analysis is depicted in Figure 2.22 through an artificial example where a query set of genes (b, c, d, e) is searched for enrichment in GO biological process and is adapted from publication (Barriot et al., 2007). To process the query, the search engine (program) converts the annotations into target sets. For example, the term *RNA splicing* will be converted to the target set including the gene d which is directly annotated with this term and also the

genes that are annotated with the more specialized terms i.e. the genes b and c annotated with *regulation of RNA splicing* and the genes c and e annotated with *nuclear mRNA splicing* resulting in the target set (b, c, d, e). During the search, the query set is compared to the target sets by the means of a similarity measure and the system returns the similar target sets as hits with their annotations ordered by decreasing similarity up to a certain threshold. From the enriched features, our query set can be characterized by the *RNA splicing process*.

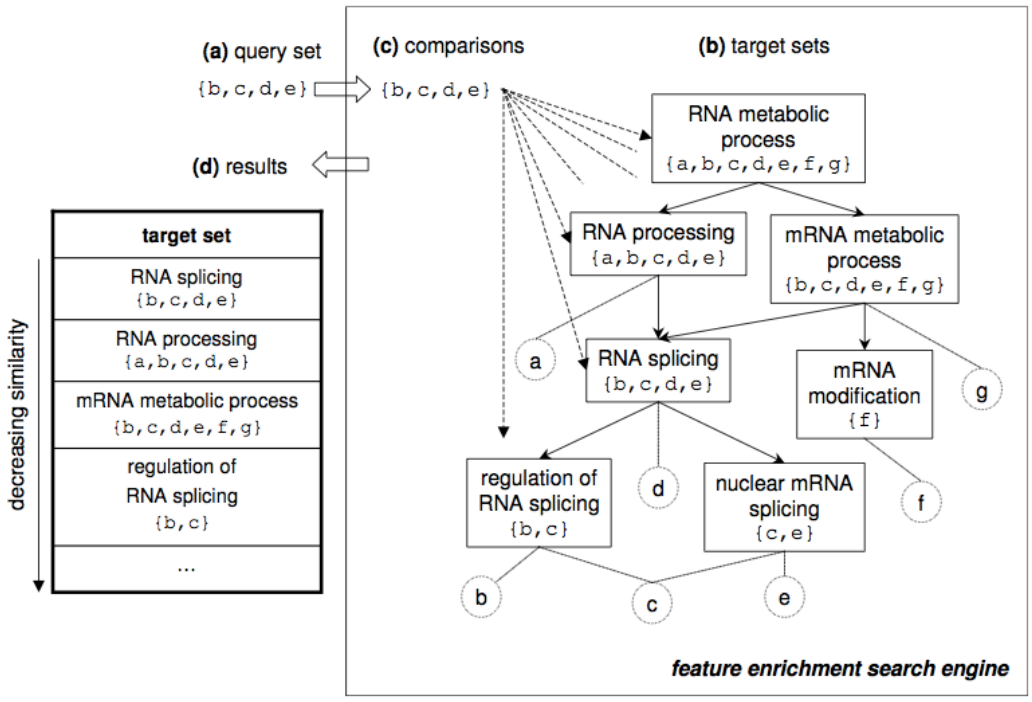


Figure 2.22: **The processing of a query gene set in order to enrich the most pertinent over-represented GO terms**. The figure shows (a) a query set that is submitted to search for similar sets in (b) a set of target sets. Sets can include each other: this is represented by a graph in which nodes represent sets and edges indicate the inclusion of a set into another. (c) the query set is compared to all the target sets based on a similarity model. (d) target sets found similar are returned ordered by decreasing similarity. ©Used with permission from the *BMC Bioinformatics* (Barriot et al., 2007).

Similar to pathway enrichment, in this method Fisher's exact test (Fisher, 1922) is also used to calculate the p-value and the q-values were calculated by controlling the false discovery rate (FDR) by Benjamini and Hochberg method (Ben-

jamini and Hochberg, 1995). The enriched terms were shown in a table view with adjusted p-values and ratio of genes from input gene list and from whole genome related to that GO term for each sub-categories of GO (Molecular function, Cellular component and Biological process). It also provides the list of genes from the input related to each enriched term.

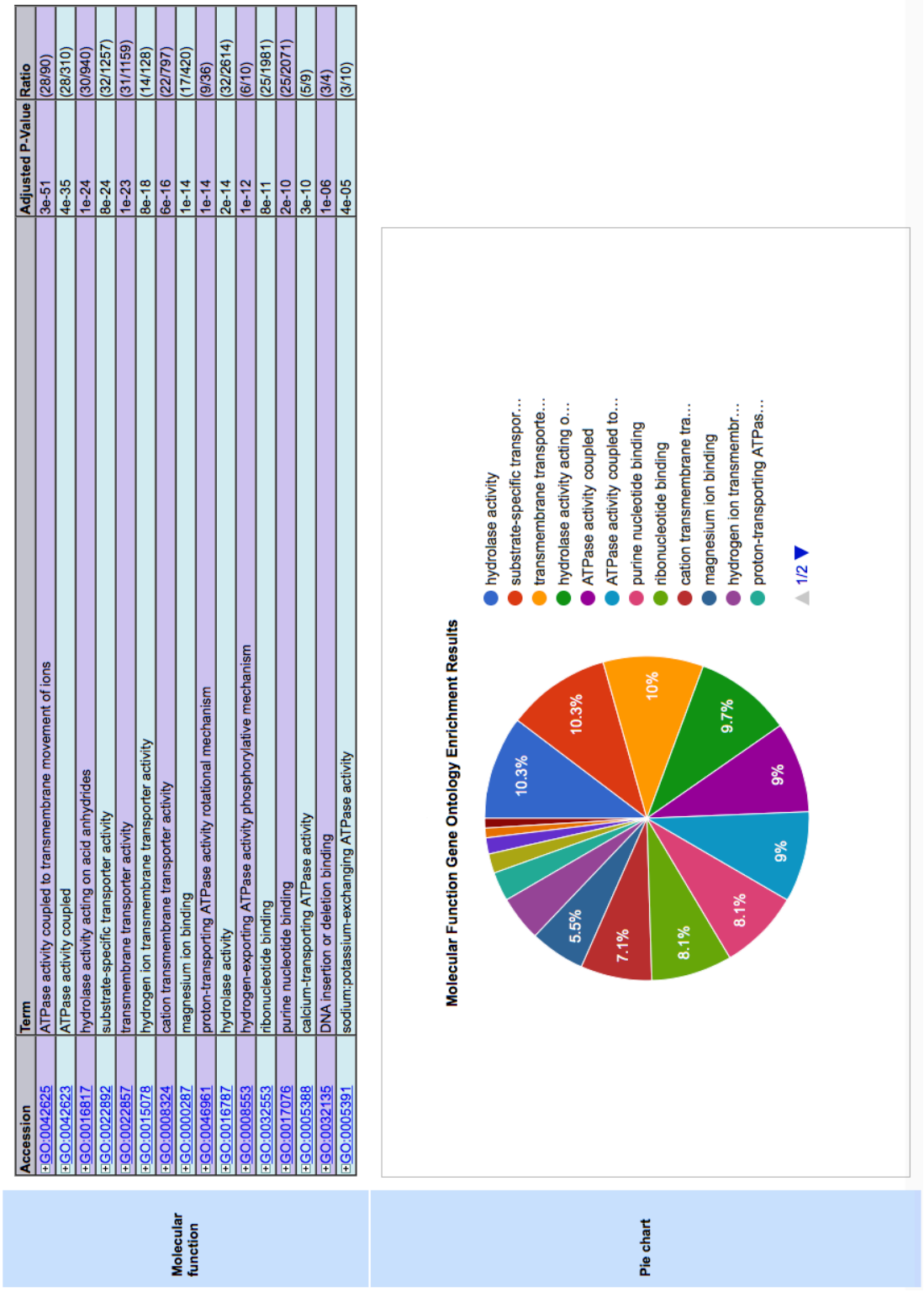
The pathway enrichment component of bioCompendium offers the user the opportunity to filter the enriched terms with the help of p-value. The filtered terms distribution is visualised as a Pie chart using Google chart tools (Google, 2015b). An example of GO molecular function enrichment results and corresponding Pie chart is shown in Figure 2.23.

2.3.10 Transcription factor binding site profiling

The binding of transcription factors (TFs) to specific locations in the DNA is integral to the transcriptional regulation in cells (Whitfield et al., 2012). In order to obtain the Transcription Factor Binding Site (TFBS) profile for a given list of genes, a 5kb upstream region starting from the first exon as shown in Figure 2.24 is selected for each gene. This region contains the $5'$ UTR (untranslated region), promoter, upstream enhancers and other gene regulatory machinery. Transcription factors (TFs), also called the master regulators, tend to bind in this region to regulate the gene expression (Gotea et al., 2010).

This 5kb genomic region is scanned with JASPAR core Position Weight Matrices (PWMs, also know as PSSMs - Position Specific Scoring Matrices) (Bryne et al., 2008). JASPAR is a collection of transcription factor DNA-binding preferences, modeled as matrices. Initially Position Frequency Matrices (PFMs) were collected from JASPAR core database and stored in local MySQL database. Later with the help of $TFBS :: DB :: JASPAR4$ Perl package, these PFMs were converted to PWMs and the above selected 5kb nucleotide sequence was scanned with each matrix with relative profile score threshold - 80% as parameter and the resulted hits were collected in GFF format.

The number of hits for each TF for each gene is collected as a matrix shown in Figure 2.25 and the elements in the matrix are clustered by using an R script (Team, 2011). The user can get more details of the matrix elements like



Accession	Term	Adjusted P-Value	Ratio
+GO:0042625	ATPase activity coupled to transmembrane movement of ions	3e-51	(28/90)
+GO:0042623	ATPase activity coupled	4e-35	(28/310)
+GO:0016817	hydrolase activity acting on acid anhydrides	1e-24	(30/940)
+GO:0022892	substrate-specific transporter activity	8e-24	(32/1257)
+GO:0022857	transmembrane transporter activity	1e-23	(31/1159)
+GO:0015078	hydrogen ion transmembrane transporter activity	8e-18	(14/128)
+GO:0008324	cation transmembrane transporter activity	6e-16	(22/797)
+GO:0000287	magnesium ion binding	1e-14	(17/420)
+GO:0046961	proton-transporting ATPase activity rotational mechanism	1e-14	(9/36)
+GO:0016787	hydrolase activity	2e-14	(32/2614)
+GO:0008553	hydrogen-exporting ATPase activity phosphorylative mechanism	1e-12	(6/10)
+GO:0032553	ribonucleotide binding	8e-11	(25/1981)
+GO:0017076	purine nucleotide binding	2e-10	(25/2071)
+GO:0005388	calcium-transporting ATPase activity	3e-10	(5/9)
+GO:0032135	DNA insertion or deletion binding	1e-06	(3/4)
+GO:0005391	sodium:potassium-exchanging ATPase activity	4e-05	(3/10)

Figure 2.23: **Gene Ontology (GO) enrichment analysis**. The figure shows a sample GO - Molecular function enrichment analysis result in a tabular view and in Pie chart after applying p-value cut off <1e-04.

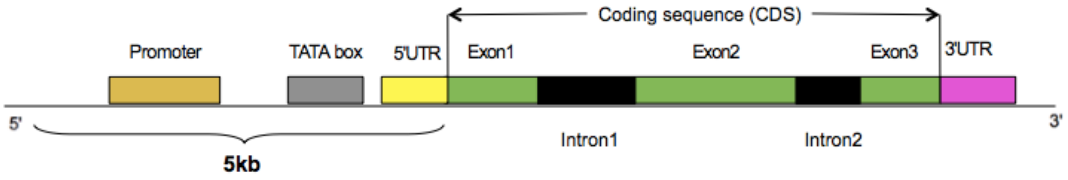


Figure 2.24: **The typical eukaryotic protein coding gene structure**. The above figure represents a eukaryotic gene with 3′ UTR (untranslated region), coding sequence consisting of both introns and exons, 5′ UTR, TATA box, promoter region and the selection of 5kb upstream region.

binding site start and stop position, score, strand etc., in GFF (General Feature Format) format as shown in Figure 2.26 by clicking on the corresponding element. Hierarchal clustering (hcluster) method from 'amap' (Another Multidimensional Analysis Package) (Lucas, 2010) R package is used to get the clustered matrix. Matrix2png tool (Pavlidis and Noble, 2003) is used to visualize this clustered matrix as a heat map as shown in Figure 2.27, from which one can see the significant TFs in the given gene list. The user can also download the matrix details before and after clustering.

2.3.11 Patent information

It is very important to know in advance if the researchers intend to patent the significant genes/proteins from their experiments and they are not already patented yet. Patented sequences were collected from European (www.epo.org), USA (www.uspto.gov), Japan (www.jpo.go.jp), Korean (www.kipo.go.kr) patent offices and the sequences for each organism were separated. Each organism specific patented sequences were mapped to the corresponding proteome by using WU-BLAST (Gish, 1996; Lopez et al., 2003). The results were stored in a local MySQL database and it helps to find if any of the sequence(s) or part of the sequence(s) from the given list of input proteins are already patented.

For each gene/protein in the given dataset, bioCompendium queries the local MySQL database and shows the relevant patent records with hyperlinks to the original data in the respective patent office and an example of such records shown

Gene name	Arnt	Arnt-Ahr	T	Pax5	En1	Evi1	Gata1	Klf4	Nkx2-5	Pax2	Pax4	Prrx2	Sox17	Sox5	Hand1-Tcfe2a	Fos	Myb	Mycn	Spz1	Bapx1	Nobox	Pdx1	Lhx3	ELF5
ENSMUSG00000030302/Atp2b2	45	100	2	1	75	1	94	145	47	96	3	19	30	39	37	34	22	44	11	38	21	34	4	49
ENSMUSG00000048489/8430408G22Rik	60	52	0	1	65	4	123	130	60	111	0	32	34	20	35	40	32	60	13	50	19	44	3	46
ENSMUSG00000030730/Atp2a1	21	67	2	3	65	1	89	191	47	98	0	19	29	18	20	27	39	20	28	42	17	20	0	42
ENSMUSG00000013160/Atp6v0d1	38	48	3	6	71	3	106	134	78	75	0	50	28	51	46	28	29	38	18	33	38	62	9	48
ENSMUSG00000029368/Alb	26	76	1	0	75	0	149	159	123	88	1	81	44	79	28	43	18	26	9	43	48	76	6	45
ENSMUSG00000021469/Msx2	19	39	2	1	70	3	94	174	65	74	0	38	31	50	16	26	33	18	12	24	33	61	3	47
ENSMUSG00000029467/Atp2a2	33	48	1	2	53	4	106	177	81	81	1	36	41	37	35	14	21	32	14	43	30	39	13	54
ENSMUSG00000018893/Mb	50	73	0	5	63	1	130	136	53	104	0	35	30	49	43	30	41	48	17	41	19	37	7	45
ENSMUSG00000033792/Atp7a	40	55	3	4	80	1	128	114	113	72	1	70	41	66	40	27	32	40	12	34	47	79	15	54
ENSMUSG00000006269/Atp6v1b1	32	65	4	2	66	2	102	199	68	98	3	45	43	38	38	47	29	32	18	38	31	48	9	47
ENSMUSG00000027546/Atp9a	45	85	1	1	78	2	114	125	72	92	1	46	25	51	46	44	23	44	12	33	32	52	16	45
ENSMUSG00000014850/Msh3	25	58	3	0	67	2	136	90	113	107	0	62	35	59	41	34	26	24	8	52	36	74	11	39
ENSMUSG00000032498/Mlh1	19	30	1	4	88	1	146	87	115	100	1	63	41	64	37	27	32	18	12	39	34	86	15	57
ENSMUSG00000031618/Nr3c2	37	101	0	1	54	2	106	133	96	82	2	62	36	67	34	21	35	36	4	25	32	65	15	47
ENSMUSG00000030075/Cntn3	12	30	4	1	91	9	127	183	110	78	0	84	55	71	28	26	31	12	17	39	49	94	6	64
ENSMUSG00000079109/Pms2	37	71	5	1	66	3	113	128	79	104	0	28	33	34	31	34	29	36	9	44	21	37	7	48
ENSMUSG00000019302/Atp6v0a1	43	34	2	1	61	24	132	303	104	77	1	79	35	64	32	33	18	42	8	39	44	68	11	71
ENSMUSG00000052459/Atp6v1a	25	48	0	0	81	1	114	125	100	83	1	70	25	65	29	22	23	24	10	39	40	82	12	42
ENSMUSG00000021245/Mlh3	38	45	2	1	83	4	104	210	91	82	0	72	45	56	42	36	20	38	17	37	45	82	19	47
ENSMUSG00000020660/Pomc	38	34	2	4	98	5	119	152	106	87	0	75	44	72	48	33	22	38	15	45	55	77	18	52
ENSMUSG00000024121/Atp6v0c	45	50	5	4	74	3	81	180	54	88	1	28	26	39	39	24	33	44	18	44	15	35	5	54
ENSMUSG00000022229/Atp12a	34	45	2	2	78	0	114	134	80	102	1	43	53	54	35	32	29	34	10	39	38	61	2	56
ENSMUSG00000024151/Msh2	46	49	5	2	76	1	98	148	95	78	0	54	40	51	36	30	35	44	8	40	41	68	51	47
ENSMUSG00000033793/Atp6v1h	20	23	0	1	83	3	174	145	101	81	2	72	48	77	40	19	18	20	9	43	39	73	6	62
ENSMUSG00000034218/Atm	33	46	1	0	57	1	109	229	91	100	0	55	40	64	33	23	19	32	16	37	24	60	8	53
ENSMUSG00000031449/Atp4b	46	97	1	2	82	1	89	165	76	88	0	38	38	48	38	20	23	46	13	40	27	35	4	49
ENSMUSG00000032570/Atp2c1	42	84	3	1	53	4	99	192	79	81	0	45	43	51	34	27	21	40	18	40	28	55	4	48
ENSMUSG00000030720/Cln3	26	37	1	2	58	4	105	160	42	95	0	13	27	29	50	29	29	26	12	38	17	23	17	58
ENSMUSG00000005370/Msh6	52	62	3	1	82	2	103	126	82	79	3	61	37	49	27	24	25	52	13	36	38	69	28	50
ENSMUSG00000032412/Atp1b3	34	46	2	1	71	1	110	131	99	89	3	67	42	71	31	29	34	34	11	32	44	74	23	41
ENSMUSG00000041329/Atp1b2	19	78	0	1	71	1	112	245	63	77	2	49	30	44	38	23	18	18	13	31	33	52	9	54
ENSMUSG00000015575/Atp6v0e	18	43	2	0	92	3	126	132	94	89	2	55	32	74	38	44	41	18	8	43	39	63	11	40
ENSMUSG00000019210/Atp6v1e1	20	36	2	0	84	1	121	147	104	97	0	67	45	74	37	31	30	20	10	28	28	64	18	65
ENSMUSG00000000441/Raf1	42	48	1	4	79	6	112	155	86	83	2	59	37	55	32	29	29	42	25	38	37	58	21	38
ENSMUSG00000005553/Atp4a	44	57	2	0	68	2	122	108	75	108	2	46	49	59	29	30	34	44	8	48	28	55	6	25
ENSMUSG00000038023/Atp6v0a2	14	29	1	5	65	1	107	84	65	94	0	33	39	51	34	31	36	14	7	28	28	40	7	24
ENSMUSG00000027073/Prg2	34	40	1	3	73	3	140	168	91	94	1	53	53	53	44	32	24	34	19	46	31	53	7	58
ENSMUSG00000033379/Atp6v0b	37	61	2	3	72	0	110	105	54	99	2	39	40	30	36	38	34	36	16	35	26	46	2	47
ENSMUSG00000021665/Hexb	41	55	2	3	82	1	134	113	92	103	1	66	52	49	40	51	30	40	16	40	43	65	10	50
ENSMUSG00000020766/Galk1	39	48	3	1	69	6	114	271	71	71	0	44	26	52	39	28	31	38	13	43	29	48	22	50
ENSMUSG00000006567/Atp7b	49	49	2	3	86	1	118	112	124	73	0	77	28	67	34	29	35	48	12	47	50	88	23	44
ENSMUSG00000041607/Mbp	45	77	5	2	55	2	98	127	63	106	0	32	33	41	46	41	32	44	18	41	25	42	0	44
ENSMUSG00000006273/Atp6v1b2	44	39	4	1	77	3	114	141	107	72	0	66	45	64	35	37	14	44	10	32	48	77	30	56
ENSMUSG00000033161/Atp1a1	31	46	3	0	84	0	126	161	83	83	1	51	30	54	33	25	34	30	11	53	37	56	5	60

Figure 2.25: **Transcription factor binding site profiling matrix**. The figure shows an example of a matrix of number of transcription factor binding sites (TFBS) for each gene from given input gene list from mouse after applying relative profile score threshold filter of 80%. TFBS number is obtained as a result of scanning each mouse transcription factor (TF) Position Weight Matrices (PWMs) from JASPAR core vertebrata database on 5kb 5′ UTR of each input gene. Here by clicking on each number, one can explore the detailed scan results in GFF (General Feature Format) format as shown in Figure 2.26 for example for gene 'Atp2a1' and TF 'Arnt', which is highlighted with a red rectangle.

Source	Feature	Start	End	Score	Strand	Frame	Attribute	Relative score
TFBS	TF binding site	27	32	8.609	-	0	sequence AACGTG ; TF Arnt ; class bHLH	93.0494347955369
TFBS	TF binding site	27	32	6.112	+	0	sequence CACGTT ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	239	244	6.112	-	0	sequence CATGTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	239	244	6.112	+	0	sequence CACATG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	344	349	6.112	-	0	sequence CACGGG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	344	349	6.182	+	0	sequence CCCGTG ; TF Arnt ; class bHLH	83.3661051128406
TFBS	TF binding site	514	519	6.112	-	0	sequence CACGGG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	514	519	6.182	+	0	sequence CCCGTG ; TF Arnt ; class bHLH	83.3661051128406
TFBS	TF binding site	894	899	6.112	-	0	sequence CACCTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	894	899	6.112	+	0	sequence CAGGTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	1698	1703	6.112	-	0	sequence CACATG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	1698	1703	6.112	+	0	sequence CATGTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	2210	2215	5.362	-	0	sequence AGCGTG ; TF Arnt ; class bHLH	80.0944403663178
TFBS	TF binding site	2280	2285	6.182	-	0	sequence CCCGTG ; TF Arnt ; class bHLH	83.3661051128406
TFBS	TF binding site	2280	2285	6.112	+	0	sequence CACGGG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	2440	2445	6.112	-	0	sequence CACTTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	2440	2445	6.112	+	0	sequence CAAGTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	3080	3085	6.112	-	0	sequence CACCTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	3080	3085	6.112	+	0	sequence CAGGTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	3405	3410	6.112	-	0	sequence CACTTG ; TF Arnt ; class bHLH	83.0868166588692
TFBS	TF binding site	3405	3410	6.112	+	0	sequence CAAGTG ; TF Arnt ; class bHLH	83.0868166588692

Figure 2.26: **Transcription factor binding site scan results**. The figure shows an example of scan results of a mouse transcription factor 'Arnt' Position Weight Matrics (PWM) from JASPAR core vertebrate database on 5kb 5′ UTR of a mouse gene 'Atp2a1' (Ensembl ID:)ENSMUSG00000030730 in GFF (General Feature Format) format.

in Figure 2.28.

2.3.12 Clinical trials information

bioCompendium processed the clinical trials record from the 'clinicalTrials.gov' web resource and mapped the each record to the corresponding disease(s) and genes. This knowledge is queried for the selected gene list in order to know whether any of the drugs in the current clinical trails apart from the chemistry information may be interesting for the selected targets (genes/proteins). These hits were shown in tabular view.

2.3.13 Protein-protein, protein-chemical interactions

Information about protein-protein and protein-chemical interactions is provided by the STITCH database (Kuhn et al., 2008). The collection of this informa-

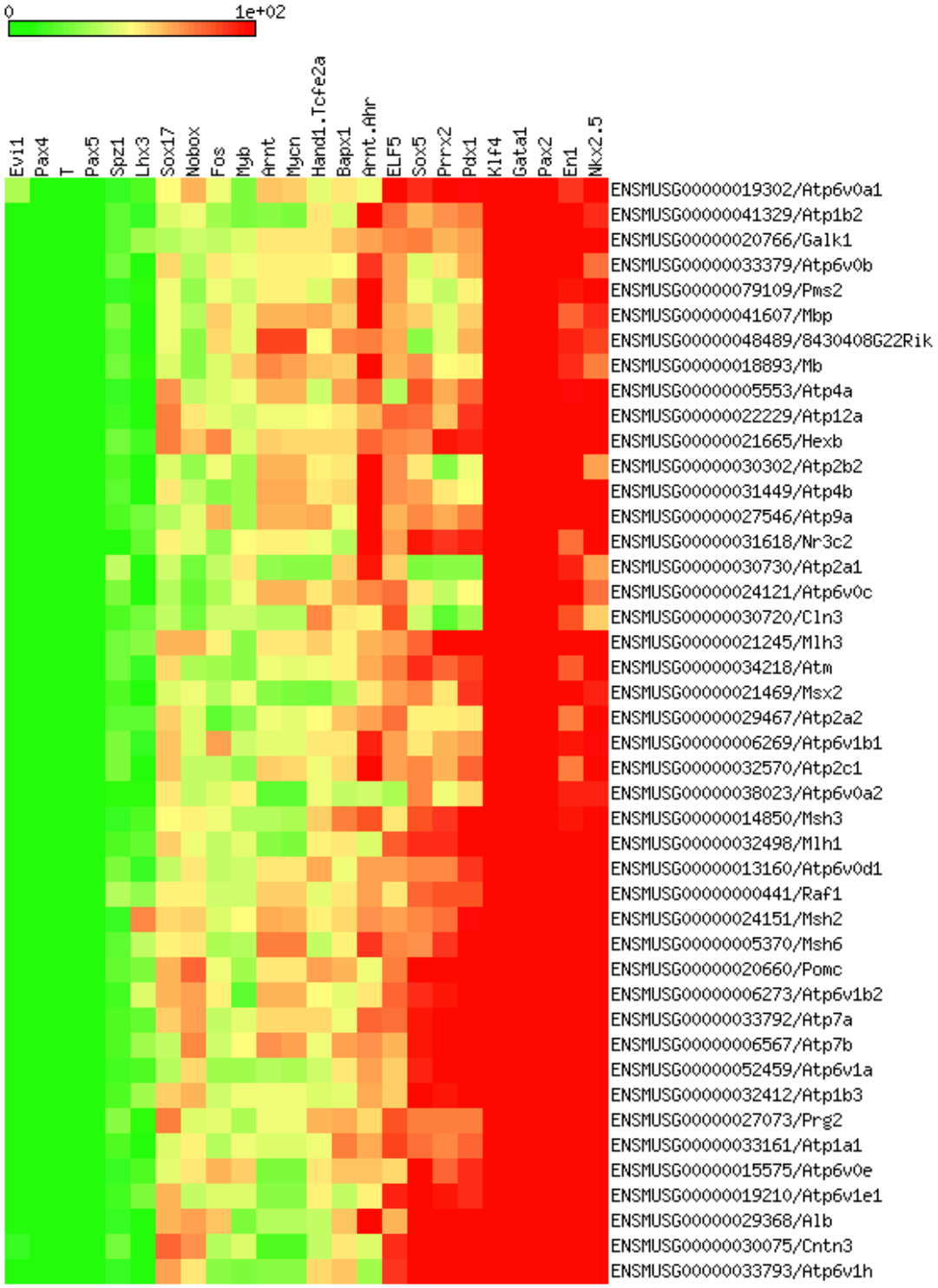


Figure 2.27: **Transcription factor binding site profile heat map**. The figure shows the heat map of transcription factor binding site profile matrix after applying hierarchical clustering (hcluster) method. Here the colour schema is adjusted to the scale 0 to 100, ranging from green(0) to red (100).

Patent Record #1	
Accession:	DD592884
Title:	GENETICALLY ALTERED ANTIBODY-PRODUCING CELL LINES WITH IMPROVED ANTIBODY CHARACTERISTICS
Significance:	1.000
Patent office:	Japan Patent Office
Molecule type:	protein
Patent family:	34221808
Non redundent patent ID:	NRP000F3E29
Publication number:	WO2005023865.
Description:	Sequence 16 from Patent WO2005023865.
Note:	
Patent Record #2	
Accession:	AAM56187
Title:	Chimeric proteins for detection and quantitation of DNA mutations, DNA sequence variations, DNA damage and DNA mismatches
Significance:	1.000
Patent office:	United States Patent and Trademark Office
Molecule type:	protein
Patent family:	26888339
Non redundent patent ID:	NRP000F3E1F
Publication number:	WO0173079.
Description:	Sequence 3 from Patent WO0173079.
Note:	
Patent Record #3	
Accession:	BD952068
Title:	ANTIBODIES AND METHODS FOR GENERATING GENETICALLY ALTERED ANTIBODIES WITH ENHANCED EFFECTOR FUNCTION
Significance:	1.000
Patent office:	Japan Patent Office
Molecule type:	protein
Patent family:	34115492
Non redundent patent ID:	NRP000F3E28
Publication number:	WO2005011735.
Description:	Sequence 11 from Patent WO2005011735.
Note:	

Figure 2.28: **Patent records in bioCompendium**. The figure shows the example patent records related to a human gene 'MSH2; DNA mismatch repair protein Msh2 (MutS protein homolog 2)' (Ensembl ID: ENSG00000095002; UniProtKB/Swiss-Prot;Acc:P43246). Here 'Accession e.g.: DD592884', 'Patent family e.g.: 34221808', 'Non redundent patent ID e.g.: NRP000F3E29', 'Publication number e.g.: WO2005023865' are hyperlinked to the original databases and patent offices.

tion was done through the available API. The STITCH database goes one step further by providing interacting proteins that were found experimentally other than proteins that co-occur in the literature. This is the main reason why this database was selected to provide to the user the relevant protein-protein and protein-chemical interaction networks for the selected list of genes.

This section is also connected to the EnrichNet - Network-based enrichment analysis (Glaab et al., 2012) via API. This is a web-application to identify, prioritize and analyze functional associations between given input gene or protein list and cellular pathways using information from molecular interaction networks.

2.3.14 Visualization of the results

bioCompendium provides interactive visualization functionally. One can visualize all or selected genes/proteins and their relationships to knowledge domains - pathways, diseases, Chemicals (drugs, ligands, metabolites etc) in either 2D representation by using the tool Medusa (Hooper and Bork, 2005), as a java applet; or in 3D representation by using the Arena3D (Pavlopoulos et al., 2008), a java application. Both tools require input data in a specific format as described in the respective publications. bioCompendium CGI script generates the input files in that particular format. In the case of Medusa one can visualize the results in bioCompendium itself, but in the case of Arena3D, bioCompendium provides a downloadable file that fits to the Arena3D input format. One has to download the Arena3D application from http://arena3d.org and upload the downloaded file to visualize the results in 3D representation. Examples of such visualisations using Medusa and Arena3D are shown in Figures 2.29, 2.30 respectively.

2.3.15 bioCompendium API

bioCompendium provides RESTful web services based API apart from the web interface. By using this API users can programmatically access its functionality and it is very practical for programmers and bioinformaticians. I have implemented RESTful web services and they can be accessed via the following URL using HTTP 'post' method.

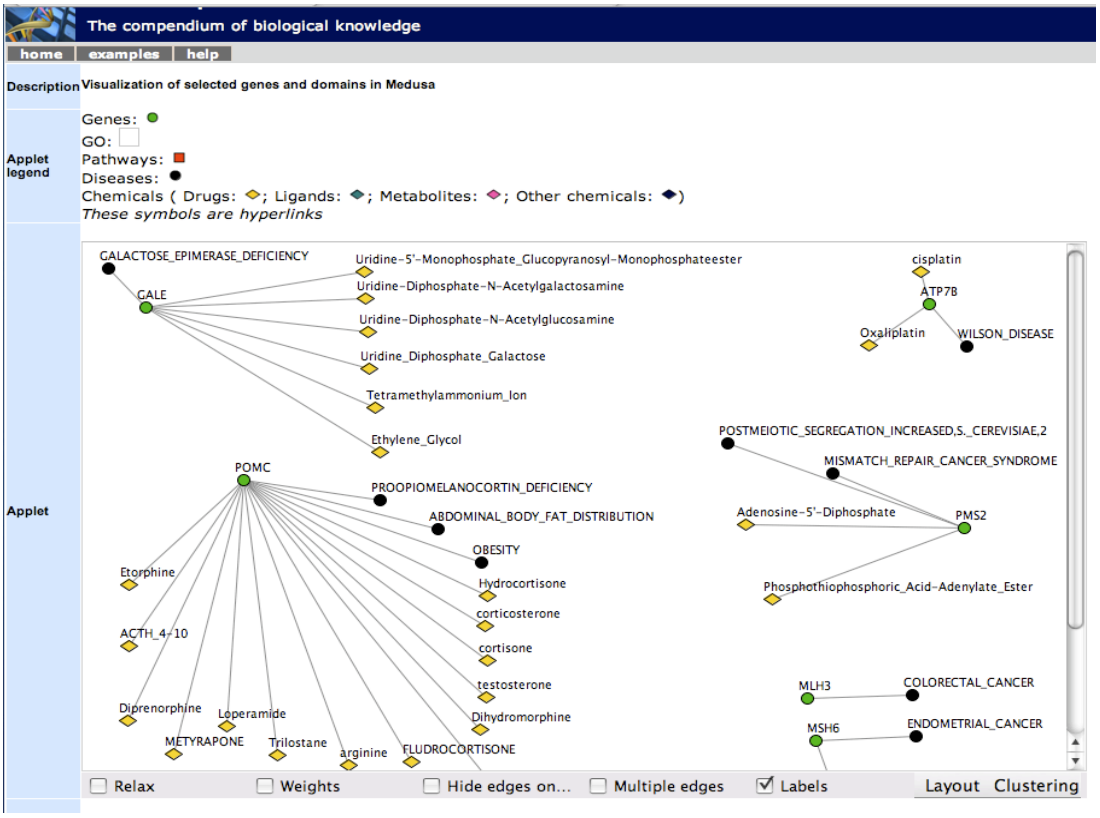


Figure 2.29: **Medusa 2D Visualization**. The figure shows an example of 2D visualization of selected genes represented as green circles and their interactions with diseases (black circles) and drugs (golden rhombuses).

`URL: http://biocompendium.org/REST/[method]/[arguments]`

The following Table 2.2 provides the list of available methods in bioCompendium API and the corresponding arguments provided in the 'Parameters' column of the table. It provides the output in four formats (i) tab-separated values (tsv), (ii) comma-separated values (csv), (iii) XML, (iv) JSON.

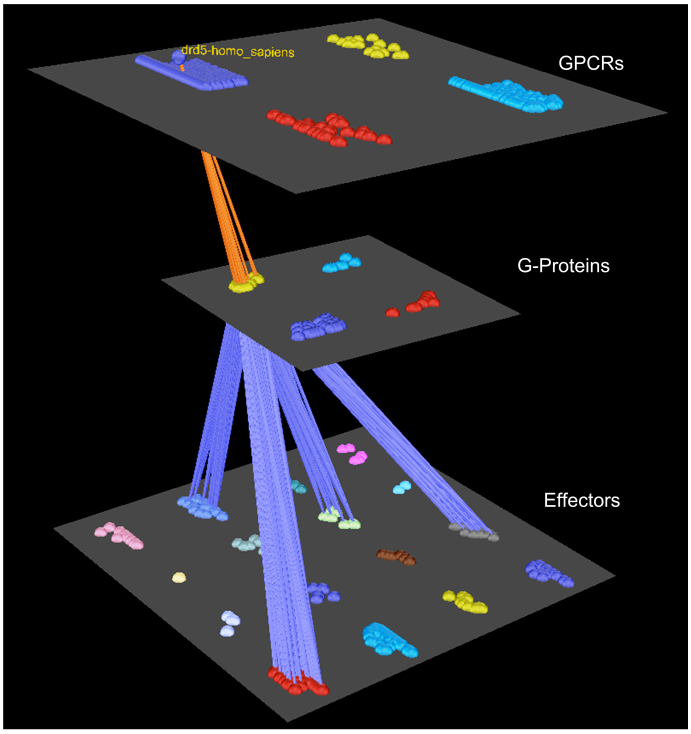


Figure 2.30: **Arena3D Visualization**. The figure shows an example of 3D visualization and organisation of bio-entities in different layers. It provides the clustering of bio-entities on each layer as well as between the layers.

Method	Description	Parameters	Output
GetPathways	Gets enriched KEGG pathways	Organism, GeneList, IdDocType, Format	Returns the pathway ID, pathway name, list of genes mapped to each pathway and p-value in the selected format (tsv, csv, XML, JSON)
GetSummary	Gets annotated gene list	Organism, GeneList, IdDocType, Format	Returns the annotations of gene list

Table 2.2: **The bioCompendium API methods** - The table provides the list of available methods in bioCompendium RESTful API, their descriptions, input parameters and output details.

2.4 Results and applications

As bioCompendium is publicly available, this resource is used by many researchers as shown in web statistics section 2.4.1. I have applied bioCompendium to several inhouse and collaborative projects, some of which are listed in the Table 2.3.

Project	Description	Colloborator	Publication
MitoCheck	Phenotypic profiling of the human genome by time-lapse microscopy reveals genes with functions in cell division, survival or migration	Ellenberg J. EMBL	(Neumann et al., 2010)
Progeria	Defective lamin A-Rb signaling in Hutchinson-Gilford Progeria Syndrome and reversal by farnesyltransferase inhibition	Djabali K., Columbia Univ. NYC	(Marji et al., 2010)
TAMAHUD	Identification of early disease markers, novel pharmacologically tractable targets and small molecule phenotypic modulators in Huntington's Disease	TAMAHUD consortium	(Jimenez-Sanchez et al., 2015a)
GARUDA	Garuda is an open, community-driven, common platform that provides a framework to interface, discover, & navigate through different applications, databases and services in bio-medical research	Garuda alliance (Garuda, 2015)	On going
HIV - Human Interaction	From experimental setup to bioinformatics: an RNAi screening platform to identify host factors involved in HIV-1 replication	Krusslich H.G., Dep. of Infect. Diseases, Univ. Heidelberg	(Borner et al., 2010)
Cancer, inflammatory and neuropathic pain	Study of molecular level difference between cancer, inflammatory and neuropathic pain related targets	Kuner R, Institute of Pharmacology, Univ. Heidelberg	(Bali et al., 2013); (Simonetti et al., 2013); (Bali et al., 2013)
Diabetes	Signaling in insulin-producing cells is altered when cells are grown in islets (3-D) compared to monolayer (2-D)	Bergsten P., Medical Cell Biology, Univ. Uppsala	(Chowdhury et al., 2013)
betaJUDO	Beta-cells of islets of Langerhans function in juvenile diabetes and obesity	Collaborative FP7 project (betaJUDO, 2015a)	(Roomp et al., 2017)
MDSC immunity	Study of mechanism of immune response suppression in Myeloid-Derived Suppressor Cells (MDSCs)	Terness P., Institute of Immunology, Univ. Heidelberg	On going
Systems biology of cancer	Insights into altered regulation in cancer and prediction of strategies for efficient intervention in diseases	Klingmller U., DKFZ, Uni. Heidelberg	On going
SysTec	Functional analysis of non-coding RNAs in living cells	Systec:FANCI consortium (Systec, 2010)	On going
Parkinson's disease	Discovery of biomarkers, understanding disease mechanisms and finding out the therapeutic interventions	In house project	On going

Table 2.3: **Use of bioCompendium** - The table listing the inhouse and collaborative projects in which bioCompendium is used to analyse and annotate the experimental results.

Of these applications, two are discussed below, (section 2.4.2) providing more details on bioCompendium plug-in in GARUDA framework, where several bioinformatics, computational and systems biology tools were integrated in a seamless manner. Section 2.4.3 focusing on betaJUDO, the study of beta-cell function in juvenile diabetes and obesity. And two applications are discussed in detail in the next chapters: Chapter 3 focus on Progeria, an accelerated ageing syndrome, whereas Chapter 4 describes TAMAHUD project, Target and Markers in Huntington's disease.

2.4.1 Web usage statistics

bioCompendium is a publicly available web application and it widely used all over the world. Its usage statistics were monitored by using google analytics (Google, 2015a). The last seven years (from 2010 to 2016)of bioCompendium usage is outlined in Figure 2.31. The upper part of figure (A) shows distribution of sessions over time. There are around 10,228 sessions and 3,968 users, among these 38% are new users and the remaining 62% are returning users. Figure (B) shows geographic distribution of users all over the world.

2.4.2 Integration with Garuda platform

'Garuda - the way biology connects' is an ongoing effort to build an open source, community driven framework to integrate several bio-medical services (bioinformatics, computational and systems biology applications, algorithms, databases etc.,) thorough a common interface (Ghosh et al., 2011). This will facilitate the integration of different applications and services as plug-ins, that communicate and send results from one tool to the other based on the input and output (result) data types and file formats. This provides the possibility to develop and run the workflows for repetitive tasks in bio-medical research. It is intended to be used in both academic and industrial sectors. Domain specific applications such as for oncology, cardiovascular diseases, infectious diseases, metabolic syndromes will be developed using Garuda Common Platform, so that interoperability between different disease domains and re-use of expertise can be enhanced.

Garuda is organising this project by providing language independent API

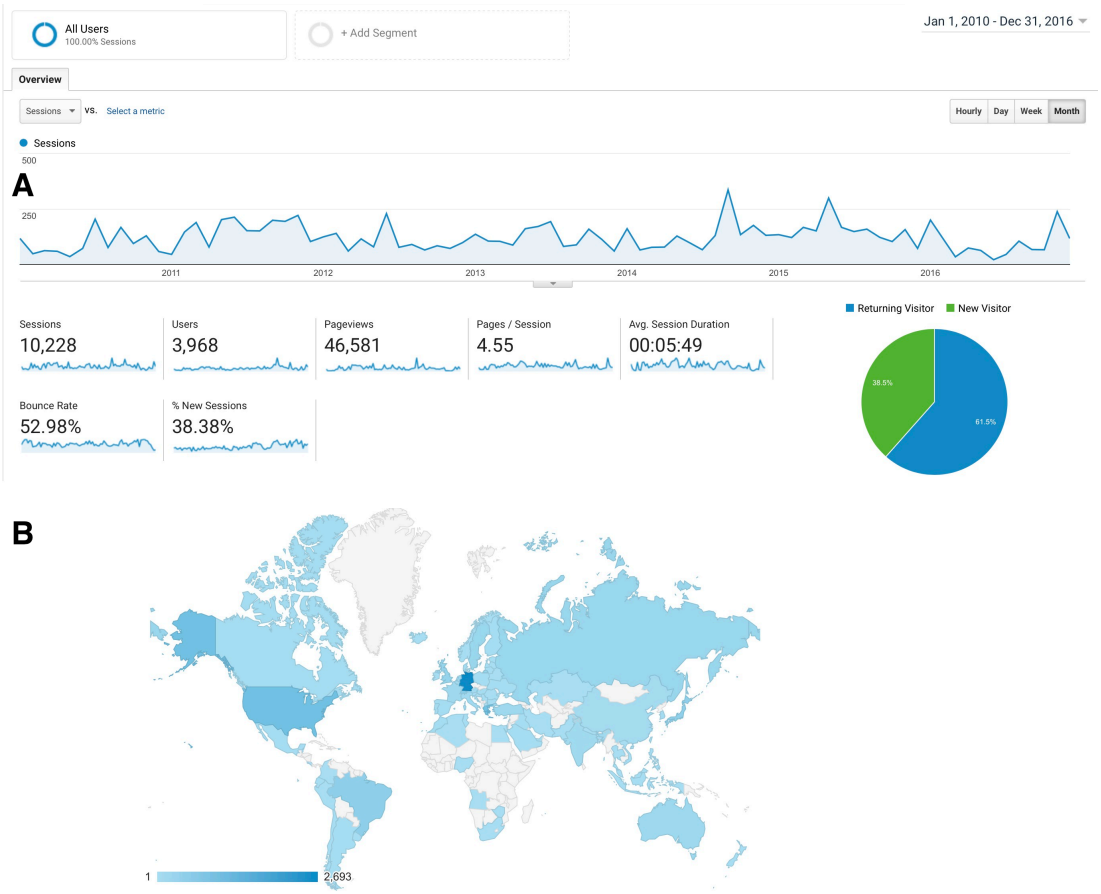


Figure 2.31: **bioCompendium web usage statistics**. The above figure shows the web interface usage statistics of bioCompendium obtained by using google analytics. Here (A) provides different statistical parameters like number of sessions over time, users (new and returning users), session duration etc., and (B) shows the geographic distribution of users all over the world.

to connect software as gadgets, explore them through the gateway and operate them through the dashboard, all the while supported by a global alliance of leaders in computational biology and informatics (Garuda, 2015). The current version, Garuda 1.1 beta is equipped with services meant for annotation, enrichment, modelling and visualisation in Systems Biology, but plans to reach other interested groups and resources.

As part of this initiative, I have integrated two services - bioCompendium analytics and SBML annotation service into Garuda using APIs from respective

services. bioCompendium analytics plug-in within Garuda dashboard is shown in Figure 2.32 (highlighted with red square). It can analyse gene or protein datasets coming from any other tool. Where as SMBL annotation server takes SMBL files as input, for example coming from a modeling tool of biochemical networks, CellDesigner and the annotation service parse the SMBL file and annotate each gene and protein with rich knowledge obtained as a result of integration of several publicly available biological databases. This services is discussed in greater details in Chapter 6.

Upon the selection of bioCompendium analytics service from the Garuda dashboard as shown in Figure 2.32, it will launch the tool as shown in Figure 2.33(A). The user can drag and drop a list of Ensembl genes in the left panel of the Figure 2.33(A) or other possibility gene dataset can be come from another tool in the Garuda, then user can run for example pathway enrichment analysis. This gadget makes a bioCompendium API call and results will be displayed in the right panel of Figure 2.33(A). These results can be downloaded locally by clicking on 'Save' button highlighted with green rectangle, whereas these pathway enrichment results can be connected to other available tools by clicking on the 'Garuda' button highlighted with the red square. This has opened the possibility to launch iPath2 as shown in Figure 2.33(B) to visualise enriched pathways on iPath visualisation web platform.

The integration into Garuda should enable bioCompendium and SBML annotation services to gain access to a wider community of biologists and it will definitely help in increasing the visibility of the tools in the field of Systems biology.

2.4.3 betaJUDO analysis with bioCompendium

2.4.3.1 betaJUDO introduction

betaJUDO is an acronym for beta-cell function in juvenile diabetes and obesity. This is a collaborative project of the European FP7-HEALTH.2011.2.4.3-2 program (betaJUDO, 2015b).

The main aim of the betaJUDO project is to develop innovative therapeutic strategies by increasing pharmacology-based alternatives targeting insulin hyper-

Figure 2.32: **Garuda dashborad**. The above figure shows the different tools and services available in Garuda 1.1 beta version that are covering Garuda core modules, analytics, pathway, modelling, visualisation service etc. bioCompendium analytics service is highlighted with red rectangle.

secretion for the treatment of young obese individuals. Clinical characterization including validation of novel genetic variants of the young obese individuals belonging to different European cohorts and a well-characterized cell model of dysfunction with isolated human palmitate-treated islets will form the basis of the translational work, where principles of reduction of insulin hypersecetion will be mechanistically dissected in the human islets and then tested in the human (betaJUDO, 2015a).

2.4.3.2 betaJUDO experiment setup

In this project biological samples were obtained from six human islet donors. Each sample was split into three equal replicates and each replicate was subjected to one of three treatments: day 0 without palmitate treatment (Control), day 2 with

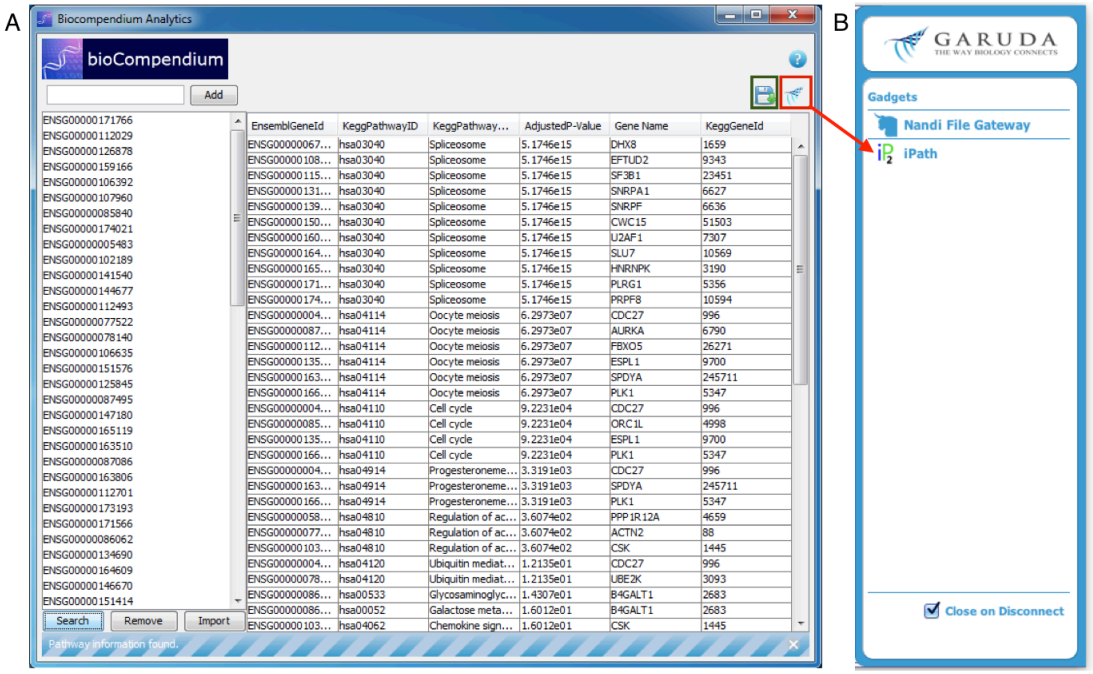


Figure 2.33: **bioCompendium analytics service in Garuda**. The above figure shows the launch and run of bioCompendium analytics from Garuda dashboard. Here A) shows the pathway enrichment widget of bioCompendium analytics, mainly consists of two panels, the left one is the place holder for input gene dataset and right panel to display the enrichment results, that can be saved using 'Save' button highlighted with green rectangle and can be connected to other Garuda gadgets by using 'Gadura Discovery' button highlighted by a red square. Whereas B) shows the compatible tools that can take pathway enrolment wrestles as input, in this case it is iPath2 garuda widget, a KEGG pathway visualisation tool.

palmitate treatment (P2), and day 7 with palmitate treatment (P7).

Two hundred islets were lyophilized and resuspended in 100 µl 0.1% RapiGest TEAB. After a brief sonication, six-plex TMT labelling was performed for protein quantification, according to manufacturers instruction (Thermo Scientific, Waltham, MA, USA). Briefly, 25 µg of each of the 6 samples were digested with trypsin, labeled, and pooled together, according to standard procedures (Dayon et al., 2008). Peptides were separated by off-gel electrophoresis, desalted and solubilized in an appropriate amount of 5% ACN / 0.1% formic acid for mass spectrometry analysis. LC-MS/MS analyses were performed on samples using nanoAcquity system (Waters, Milford, MA, USA).

2.4.3.3 betaJUDO data acquisition and processing

Raw data were converted to mgf files, and peak lists were submitted to EasyProt (Gluck et al., 2013) for research against the human SwissProt database. Selected criteria for identification are <1% False Discovery Rate and a minimum of 2 unique peptides per protein. Isobar (Breitwieser et al., 2011) was used for quantification after isotopic correction, according to manufacturer instructions (Thermo Scientific, Waltham, MA, USA). Slight differences in the total channel intensity are typically observed in each of the channels, as the total amount of protein used for each TMT condition is not the same. Therefore, the total channel intensities relative to each other were normalized; each peptides ion intensity was divided by a channel-specific factor derived from the comparison of the total sum of the ion counts.

Next, the calculation of normalized quantifications was performed at the protein level, where each donor was treated separately. Then Libra methodology from the Trans Proteomic Pipeline (Shteynberg et al., 2015) was applied, only proteins with at least two unique peptides were included. Each peptide channel was normalized by the sum of all channels of the respective peptide. The mean and population standard deviation for all peptides of a protein in each channel were determined. Subsequently, simple outlier removal was performed, where those peptides with intensities two standard deviations outside of the mean in one or more channels were removed, then the mean was recalculated. If a peptide was measured twice due to different retention times and one displayed outlier characteristics, only the outlier peptide was removed. Contaminating proteins were excluded, as defined by Mascot (Koenig et al., 2008). This produced normalized intensities at the protein level.

Kirsten Roomp has processed and analysed the above mentioned mass spectrometry data and generated the significant protein datasets. I have analysed these datasets using bioCompendium. Apart from the classic features of bioCompendium (refer section 2.3), I have added two new domains to the bioCompendium knowledge base exclusively for this project and these are discussed in the following sub-sections. But eventually these new domains will also be made available in the public version of bioCompendium.

2.4.3.4 bioCompendium analysis of betaJUDO

In bioCompendium a permanent session for betaJUDO data has been created and is available at URL http://biocompendium.embl.de/betajudo. The consortia members can always browse the results without upload of the same protein datasets over and over again by different members. This session is shown in the Figure 2.34, and two new modules integrated for this project are highlighted with red rectangle.

2.4.3.5 Tissue expression profiling

The tissue expression information is collected from the Human Protein Atlas (HPA) resource, a tissue-based map of the human proteome (Uhlen et al., 2015). Currently it is covering 213 tissue and cell line samples from major organs of the human body (HPA, 2015). This information has been parsed, mapped to Ensembl gene IDs, and integrated with bioCompendium database using ETL (Extract, Transform and Load) scripts. Upon request from bioCompendium web interface, the server side program runs a join SQL query between the list of the genes and the tissue expression table and results were displayed in tabular view as shown in Figure 2.35. It provides expression information in different tissues and the expression values are represented as low, medium and high and each entry hyper linked to respective entries in Ensembl and HPA databases.

This tissue expression data has been clustered by using hierarchal clustering (hcluster) method 'amap' (Lucas, 2010), a R package. Matrix2png tool (Pavlidis and Noble, 2003) is used to visualize this clustered data as a heat map as shown in Figure 2.36 to visualise the clustering of genes based on their tissue expression. This figure shows, almost all of the significant proteins from betaJUDO mass spectrometry experiment are over expressed in two cell lines Islets of Langerhans and Exocrine glandular cells of pancreas. From this we deduce the inference that these proteins are true and that the selected cell line based approach is a good model to study juvenile diabetes and obesity.

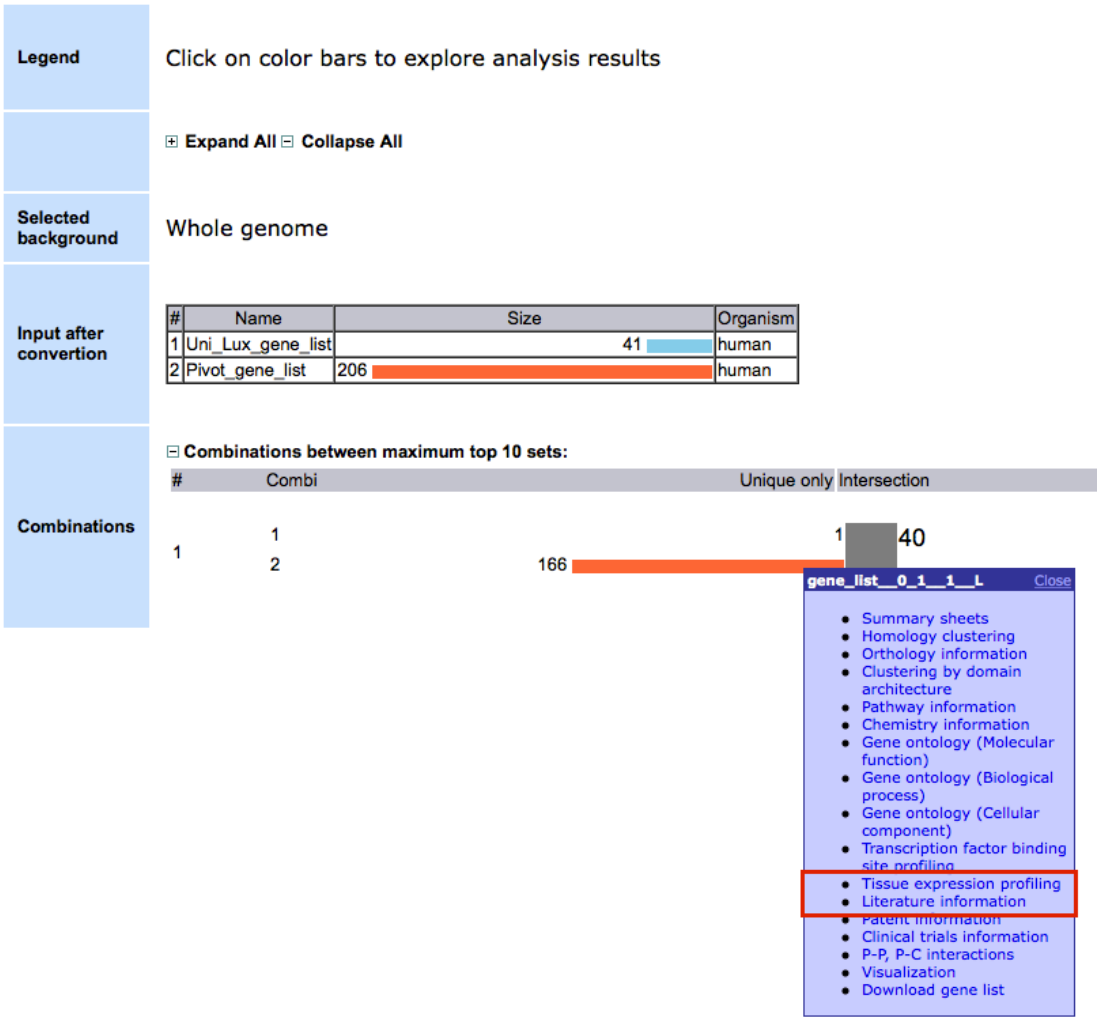


Figure 2.34: **bioCompendium betaJUDO analysis permanent session**. This figure shows the bioCompendium comparative analysis of two protein lists provided by betaJUDO consortium. It shows the available menu and the two new modules integrated for this project (highlighted with red rectangle).

- Hyperlink to HPA - The Human Protein Atlas; e! - Hyperlink to ENSEMBL

Gene name	Ensembl gene	Description	Tissue:Cell type	Expression level
GCG	ENSG00000115263	Glucagon (GRPP) (OXY) (OXM)[Glicentin][Glicentin-related polypeptide] [Oxyntomodulin][Glucagon] [Source:UniProtKB/Swiss-Prot;Acc:P01275]	Pancreas:Islets of Langerhans	High
			Lung:Macrophages	Low
REG1B	ENSG00000172023	Lithostathine-1-beta Precursor (Regenerating protein I beta) [Source:UniProtKB/Swiss-Prot;Acc:P48304]	Small intestine:Glandular cells	High
			Duodenum:Glandular cells	High
			Pancreas:Exocrine glandular cells	High
			Stomach:Glandular cells	Low
GAPDH	ENSG00000111640	Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) (GAPDH) [Source:UniProtKB/Swiss-Prot;Acc:P04406]	Testis:Leydig cells	High
			Epididymis:Glandular cells	High
			Pancreas:Exocrine glandular cells	High
			Tonsil:Germinal center cells	Medium
			Cerebral cortex:Glial cells	Medium
			Placenta:Decidual cells	Medium
			Rectum:Glandular cells	Medium
			Duodenum:Glandular cells	Medium
			Salivary gland:Glandular cells	Medium
			Esophagus:Squamous epithelial cells	Medium
			Skin:Keratinocytes	Medium
			Small intestine:Glandular cells	Medium
			Lateral ventricle:Glial cells	Medium
			Breast:Glandular cells	Medium
			Tonsil:Non-germinal center cells	Medium
			Skin:Melanocytes	Medium
			Cervix, uterine:Squamous epithelial cells	Medium
			Lung:Pneumocytes	Medium
			Uterus:Cells in endometrial stroma	Medium
			Fallopian tube:Glandular cells	Medium
			Skin:Langerhans	Medium
			Adrenal gland:Glandular cells	Medium
			Testis:Cells in seminiferous ducts	Medium
			Appendix:Lymphoid tissue	Medium
			Tonsil:Squamous epithelial cells	Medium
			Cerebellum:Cells in molecular layer	Medium
			Cervix, uterine:Glandular cells	Medium
			Cerebral cortex:Neuronal cells	Medium
			Colon:Peripheral nerve/ganglion	Medium
			Ovary:Ovarian stroma cells	Medium
			Lung:Macrophages	Medium
			Hippocampus:Neuronal cells	Medium
			Kidney:Cells in tubules	Medium
			Heart muscle:Myocytes	Medium
			Pancreas:Islets of Langerhans	Medium
			Liver:Hepatocytes	Medium
			Bronchus:Respiratory epithelial cells	Medium
			Gallbladder:Glandular cells	Medium
			Hippocampus:Glial cells	Medium
			Nasopharynx:Respiratory epithelial cells	Medium
			Spleen:Cells in white pulp	Medium

Figure 2.35: **Tissue expression profiling in tabular view**. The table view provides expression information of tissues and their cell lines for each gene.

2.4.3.6 Literature information

This section of the bioCompendium provides context specific literature that is very relevant for betaJUDO project. This is achieved by using specific MeSH (Medical Subject Headings) (NIH, 2014) terms 'pancreas, pancreatic islets, liver, adipose tissue, insulin, glucagon, carbohydrate metabolism and lipid metabolism'. These MeSH terms are covering relevant tissues, cell types and processes. In addition disease specific terms 'obesity, type 2 diabetes, hyperglycemia, hyperlipidemia, hyperinsulinemia'. These specific MeSH terms were used in the PubMed

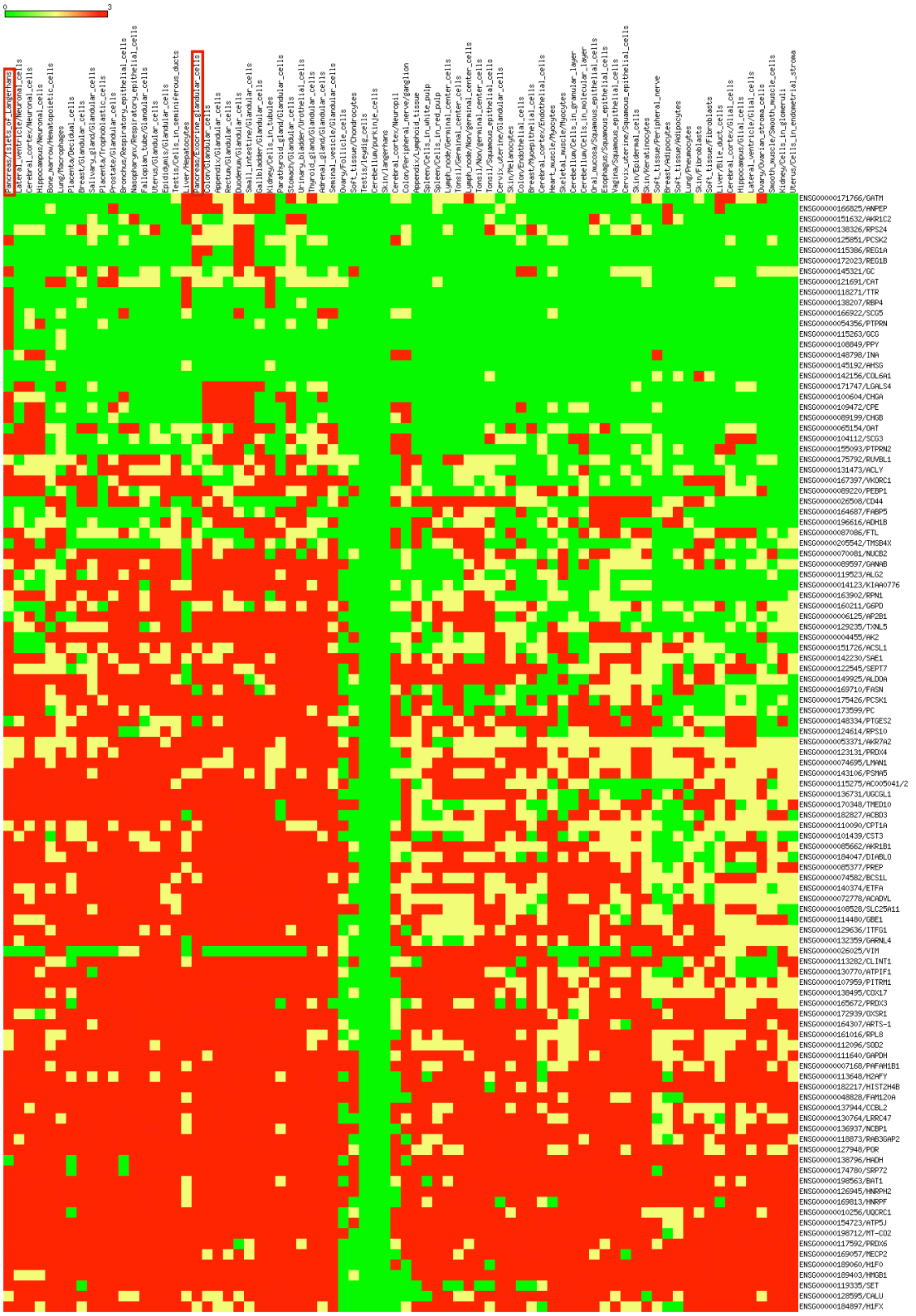


Figure 2.36: **Tissue expression profile clustering**. This figure visualises the clustering of genes based their tissue expression profiles. Here almost all of the proteins are over expressed in two cell lines Islets of Langerhans and Exocrine glandular cells of pancreas that are highlighted with red rectangle. Here color code: green zero to red three high expression, two medium and one low expresson.

queries for the co-mentioning of protein and their synonyms. These queries are restricted only for human, mouse and rat species and run for each and every protein. In order to build these queries, protein names and their synonyms for all the proteins from human, mouse and rat were obtained from UniProt database.

The queries and the retrieved PubMed IDs for each protein were presented in a tabular view as shown in the Figure 2.37 with a possibility to run the query on the fly. The list of PubMed IDs were sorted by the year of publication and each PubMed ID providing tagged version of the abstract by tagging genes, proteins, small molecules and wikipedia terms by using Reflect service (Pafilis et al., 2009). Clicking on a tagged term opens a small popup showing summary information, as shown in Figure 2.38 for a protein e.g. 'insulin'.

bioCompendium provides comprehensive knowledge, annotation and analysis results related to the significant proteins obtained from the betaJUDO mass spectrometry at one place. This is an ongoing project, and the proteomics and knowledge acquired with the help of bioCompendium will be integrated with Lipidomics data.

2.5 Discussion

2.5.1 Summary of results and applications

As we know, complex diseases occur as a result of dysregulation, mutations and the physiological functions, for example, cell division, apoptosis occur as a result of interaction between various bio-entities (e.g., genes, proteins, chemicals, metabolites, regulatory elements, non-coding RNAs) in different biological processes and pathways at different levels - cell, tissue, organ and systems level. These are not local effects, but rather systemic, a network of systematically organised events taking place in organism level. In order to study the systems biology of these events, both pathological and physiological, researchers are employing genome-wide high throughput experimentation using techniques including microarrays, proteomics, metabolomics, next generation sequencing technologies (whole genome, exome, RNAseq). These high-throughput experiments are resulting in different sets of genes or gene products, metabolites, non-coding RNAs (for

Literature information

Download below table as tab delimited file

e! - Hyperlink to ENSEMBL; PubMed - Hyperlink to PubMed

Ensembl ID	UniProt ID	Gene name	Gene alias symbols	Gene full name	PubMed query	PubMed IDs
ENSG00000114480 e!	Q04446	GBE1		1,4-alpha-glucan-branching enzyme (EC 2.4.1.18)	(2.4.1.18[TIAB] OR GBE1[TIAB] OR 1,4-alpha-glucan-branching enzyme[TIAB]) AND (pancreas[MH] OR pancreatic islets[MH] OR liver[MH] OR adipose tissue[MH] OR insulin[MH] OR glucagon[MH] OR carbohydrate metabolism[MH] OR lipid metabolism[MH]) AND (Human[MH] OR Mouse[MH] OR Rat[MH]) AND (obesity[MH] OR type 2 diabetes[MH] OR hyperglycemia[MH] OR hyperlipidemia[MH] OR hyperinsulinemia[MH]) PubMed	[2000]:10868977
ENSG00000118271 e!	P02766	TTR	PALB	Transthyretin	(TTR[TIAB] OR Transthyretin[TIAB] OR PALB[TIAB]) AND (pancreas[MH] OR pancreatic islets[MH] OR liver[MH] OR adipose tissue[MH] OR insulin[MH] OR glucagon[MH] OR carbohydrate metabolism[MH] OR lipid metabolism[MH]) AND (Human[MH] OR Mouse[MH] OR Rat[MH]) AND (obesity[MH] OR type 2 diabetes[MH] OR hyperglycemia[MH] OR hyperlipidemia[MH] OR hyperinsulinemia[MH]) PubMed	[2014]:23873966; [2013]:22849972; [2012]:23129325, 22826435; [2011]:21726512; [2010]:20176725; [2009]:19342603, 19147488; [2008]:18825272, 18780768, 18437353, 17968970, 17961122; [2007]:17618858, 17416795; [2005]:16216936, 15632105; [2004]:14722648; [1997]:9290094, 9013432
ENSG00000124614 e!	P46783	RPS10		40S ribosomal protein S10	(40S ribosomal protein S10[TIAB] OR RPS10[TIAB]) AND (pancreas[MH] OR pancreatic islets[MH] OR liver[MH] OR adipose tissue[MH] OR insulin[MH] OR glucagon[MH] OR carbohydrate metabolism[MH] OR lipid metabolism[MH]) AND (Human[MH] OR Mouse[MH] OR Rat[MH]) AND (obesity[MH] OR type 2 diabetes[MH] OR hyperglycemia[MH] OR hyperlipidemia[MH] OR hyperinsulinemia[MH]) PubMed	

Figure 2.37: **Literature information**. The table view provides literature information relevant to each protein in the context of diabetes and obesity and also provides the specific PubMed queries and the retrieved PubMed IDs hyperlinked to Reflect service that presents a tagged version of the abstract as shown in the Figures 2.38.

example miRNAs, siRNAs) that are playing an important role in the respective disease or healthy condition. There is a pressing need to integrate these heterogenous experimental datatypes and prioritise the bio-entities for either early detection of the disease (as disease biomarkers) or efficient patient stratification and personalised therapies. On the other side there is large amounts of already published literature (>27millions scientific papers in PubMed) and various publicly available biological databases. I want to take advantage of these vast amounts of public knowledge to annotate, validate and prioritise the bio-entries, markers, drug targets from above mentioned high throughput experimental techniques. The main problem with public data resources, they are heterogenous in

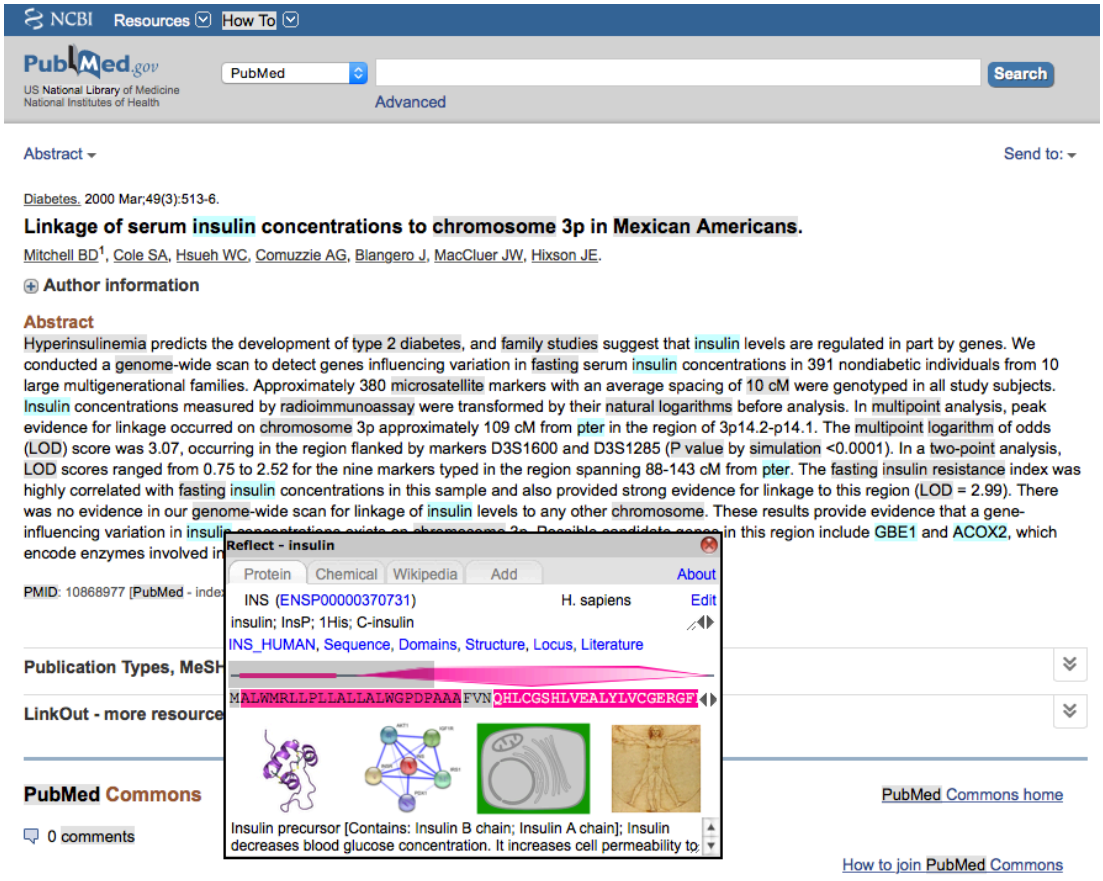


Figure 2.38: **Tagged version of abstract with protein popup**. The figure shows the abstract in which bio-entities: genes, proteins, drugs, other chemicals and wikipedia terms are highlighted. A popup shown in the rectangle opens upon mouse over for example, a protein 'insulin' and it provides corresponding protein's ENSEMBL identifier, its synonyms, description, organism name, sequence, SMART domain architecture, its 3D structure, interaction network, cellular localisation etc.,

both content (e.g. genes, proteins, regulatory elements, chemicals, metabolites, diseases, ontologies, reactions, interactions, pathways, literature etc.,) and format (e.g. flat-file, XML, relational database). This public data needs to be parsed and integrated in order to use this knowledge seamlessly.

In order to address the above mentioned needs, I have developed bioCompendium, which consists of fine grained knowledge stored in a bioCompendium knowledge base. It consists of more than 80 important public biological databases collected locally and integrated into the SRS system. It also uses Ensembl,

BioMart, a text mining resource (AKS2) as bioCompendium data layers (for more details refer the Data integration section 2.2.1). All these databases are scanned for each gene from human, mouse and yeast and the relevant information is stored in a local MySQL database (bioCompendium knowledge base). As scientists use different database identifiers (IDs) to represent their experiment results, in order to compare the results from one format of IDs to another, I have implemented an ID Conversion service. Researchers also do experiments in different model organisms. Therefore to achieve the cross-species comparison and enrichments, I am converting information from one genome to another by using orthology relationships. One more unique feature of bioCompendium is handling of documents (PDF, MS-Word, Excel, PowerPoint, plain text) to extract gene lists. I have implemented an API to access the Reflect (O'Donoghue et al., 2010), a bio-entity tagging service. After converting any of the above mentioned documents to ascii text, that text is sent to the Reflect via an API and the tagged version of the text is received back. The Bio-Entities (genes/proteins) are extracted and processed like a gene set. The information in bioCompendium is Ensembl gene centric, all the information from methods - ID Conversion service, Cross species comparison based on orthology, Handling of documents, boils down to Ensembl genes and it provides several bioinformatics analysis results.

bioCompendium is a 'work horse' in our group to analyse datasets of genes or gene products obtaining from various high throughput experiments from different projects. As this web application is open to the public it became a tool of choice not only for our institute but also for researchers from all over the world as indicated in the web statistics of bioCompendium (section 2.4.1).

It is equipped with an autocomplete-based simple search, one can search the whole database content using this functionality. bioCompendium allows analysis and annotation of one or more gene/protein lists and/or documents from human, mouse or yeast experiments. It provides a menu for different analysis and visualisation of the results, upon clicking on each bar as shown in the Figure 2.9. It offers summary sheets for a given gene dataset. This facilitates the user to browse all the relevant information for a given gene at one place. It saves a lot of time rather visiting each database individually, which is laborious (section 2.3.3 for more details on summary sheet).

bioCompendium offers two clustering workflows - 'Homology' (section 2.3.4) and 'Domain' clustering (section 2.3.6). The first one based on sequence similarity where as the second based on the SMART domains. These clusters were presented in a intuitive way to explore its elements and their features. In addition an interactive 2D graphical view is implemented. It also provides two enrichment workflows - 'Pathway' (section 2.3.7) enrichments have been calculated for KEGG, PANTHER and Reactome. Whereas 'GO' enrichment analysis (section 2.3.9) has been calculated for three sub-categories of GO. It provides transcription factor binding site profiling for each gene as well as patent information.

bioCompendium integrates chemical information from several databases and literature (refer section 2.3.8 for more details) and is mapped to the protein targets. I am also scanning ongoing clinical trials from `clinicalTrials.gov` to cover the drugs which are under current trials and are relevant to the protein targets from the user input. It is also equipped with 2D (Medusa, a java applet) and 3D (Arena3D, a java application) visualisation tools to visualise selected genes/proteins and their relationships to pathways, diseases, chemicals (drugs, ligands, metabolites), GO terms. It also provides tissues expression profiling and highly relevant literature information by running customised queries to PubMed. The abstracts are tagged with Reflect service that provides useful popups for genes, proteins, chemicals and wikipedia terms. Though these two domains are exclusively developed for betaJUDO project, they will be soon made available for public version of bioCompendium.

Apart from webbased functionaries, bioCompendium provides RESTful API. It is very practical for programmers and bioinformaticians to access its functionality programatically. All the available API methods are detailed in the section 2.3.15.

As bioCompendium handles all documents (PDF, MS-Word, Excel, PowerPoint, plain text), one can use this functionally to compare the experimental results with already published data by directly uploading gene dataset(s) and one or more PDFs of published literature. It will help the researcher by avoiding the painstaking step of collecting gene names manually from the literature. This functionality also helpful to just extract the bio-entities from one or more scientific publications.

bioCompendium both front end (the web application) and the backend (the database) are designed in very modular way. It is easy to add new knowledge domains to the existing schema and develop respective web interface views to serve the new knowledge. This service is widely used in in-house projects as well as within the worldwide scientific community.

bioCompendium is a tool of choice for high-throughput experiments data annotation and analysis.

2.5.2 Future development and directions

As there is lot of interest in this resource from the scientific community, I would like to further continue developing the service. Transcription factors binding site profiling section needs to be improved by providing the possibility for the user to select the genomic regions of interest and also apart from the predictions based on PWMs, I would like to integrate ChIP-chip and ChIP-seq experimental data from ENCODE (ENCODE-Consortium, 2011) and other projects. There is lot of room to improve the visualisation part of the application, especially 2D Medusa applet, which will be replaced by interactive javascript based visualisation to visualise the interactions between bio-entities.

More model organisms will be added to the resource, four new ones are in the pipeline - rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), fruitfly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*). New knowledge domains, especially noncoding RNAs are playing a very important role and we have limited knowledge about functionality of these ncRNAs and their interactions with other bio-entities. miRNAs, siRNAs. pseudogenes, circular and long ncRNAs are planned to be integrated into bioCompendium. Another important domain is single nucleotide polymorphism or directed mutagenesis, we will systematically mine literature for the mutations and their effects will be integrated into the resource.

Scientists working in plant science research would like to have a similar resource for plant species. The bioCompendiumPlants will be developed in collaboration with the plant scientific community.

Contributions: The development of bioCompendium was conceived by Reinhard Schneider and me. The programming of the bioCompendium database, web application, bioinformatics pipelines, analytics, API and application to various projects was carried out by me.

Chapter 3

Progeria

3.1 Introduction

Hutchinson-Gilford Progeria Syndrome (HGPS) is a rare premature aging disorder caused by a de novo heterozygous point mutation G608G (GGC.GGT) within exon 11 of LMNA gene encoding A-type nuclear lamins (Marji et al., 2010). The children affected with this disease age 10 times faster than normal humans. Due to this accelerated aging progeria children have the same cardiovascular, respiratory, and arthritic conditions as senior citizen. On average, these children will die at the age of 13 due to progressive coronary atherosclerosis (Hennekam, 2006). There are roughly 100 identified children living with progeria (http://progeriaresearch.org). Apart from progeria, mutations of lamin A/C (LMNA) cause a wide range of human disorders, including lipodystrophy, neuropathies and autosomal dominant Emery-Dreifuss Muscular Dystrophy (EDMD) (Bakay et al., 2006).

The gene structure of LMNA gene is shown in the Figure 3.1 and is adapted from (Pollex and Hegele, 2004). It is 57.6kb in size and contains 12 exons, due to alternative splicing it encodes 4 proteins. Two are minor products, lamin A delta 10, lamin c2 and two major products lamin A and lamin C. These two major proteins form a heterodimer through their rod domains that create the filamentous structures found in the nuclear lamina. There are a couple of mutations causing progeria, among these G608G silent mutation is the most common (Pollex and Hegele, 2004).

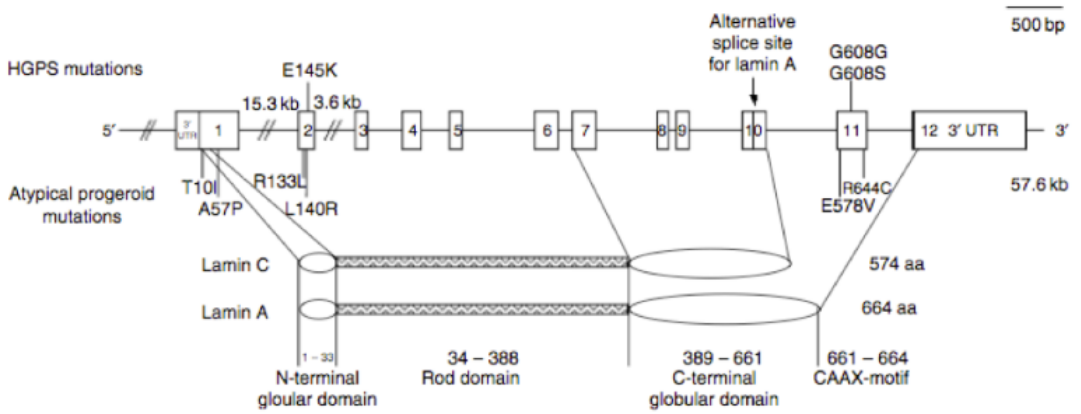


Figure 3.1: **Mutations in HGPS**. The figure shows the LMNA gene structure, mutations, the lamin A, C protein isoforms and locations of the mutations that causing progeria disease. ©Used with permission from the *Clinical Genetics journal, John Wiley and Sons publisher* (Pollex and Hegele, 2004)

.

Lamin A is synthesised from its precursor, prelamin A, which has a CaaX motif at its carboxyl terminus. Figure 3.2 shows the pictorial representation of prelamin A processing in normal and progeria condition. In normal condition the CaaX motif initiates series of catalytic reactions that lead to farnesylation and carboxymethylation of the carboxy termical cysteine. This farnesylated and carboxymethylated prelamin A is normally cleaved near its carboxyl-terminus in a reaction catalyzed by endoprotease ZMPSTE24, leading to the removal of farnesylated cysteine (Worman et al., 2009). It maintains the integrity of nuclear envelope. But in the case of HGPS condition, the LMNA G608G mutation creates a cryptic splice site within exon 11, generating an mRNA that encodes a prelamin A with a 50 amino acid deletion at its carboxyl-terminal domain (De Sandre-Giovannoli et al., 2003; Eriksson et al., 2003), which ultimately deletes the ZMPSTE24 endoproteolytic cleavage site and hence retains the farnesylated and carboxymethylated cysteine at its carboxyl terminus (Worman et al., 2009).

Expression of progerin, the truncated protein induces severe abnormalities in nuclear morphology, heterochromatin organization, mitosis, DNA replication and

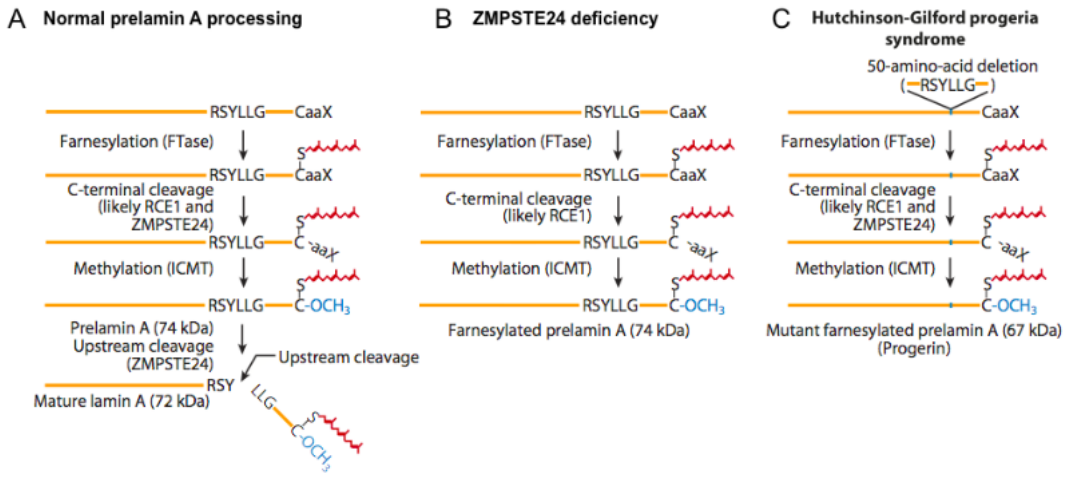


Figure 3.2: **Lamin A vs Progeria biogenesis**. Biogenesis of lamin A and the failure to generate mature lamin A in the case of ZMPSTE24 deficiency that leads to restrictive dermopathy and Hutchinson-Gilford progeria syndrome (HGPS). Prelamin A (664 amino acids) normally undergoes four posttranslational processing steps (A). First, the cysteine of the CaaX motif is farnesylated by protein farnesyltransferase (FTase). Second, the -aaX is released. Third, the newly exposed farnesylcysteine is methylated. Fourth, the carboxyl-terminal 15 amino acids, including the farnesylcysteine methyl ester, are clipped off by ZMPSTE24 and degraded, releasing mature lamin A (646 amino acids). In the case of ZMPSTE24 deficiency (B), the last endoproteolytic processing step does not occur, resulting in the accumulation of the farnesylated form of prelamin A. In the case of HGPS (C), a point mutation results in a 50-amino-acid deletion in prelamin A (amino acids 607656), which removes the site for the second endoproteolytic cleavage. Thus, the farnesylated mutant prelamin A (progerin) accumulates in cells, and no mature lamin A is formed. ©Used with permission from the *Annual Review of Genomics and Human Genetics* (Davies et al., 2009). RCE1, Ras converting enzyme 1; ICMT, isoprenyl cysteine methyl transferase.

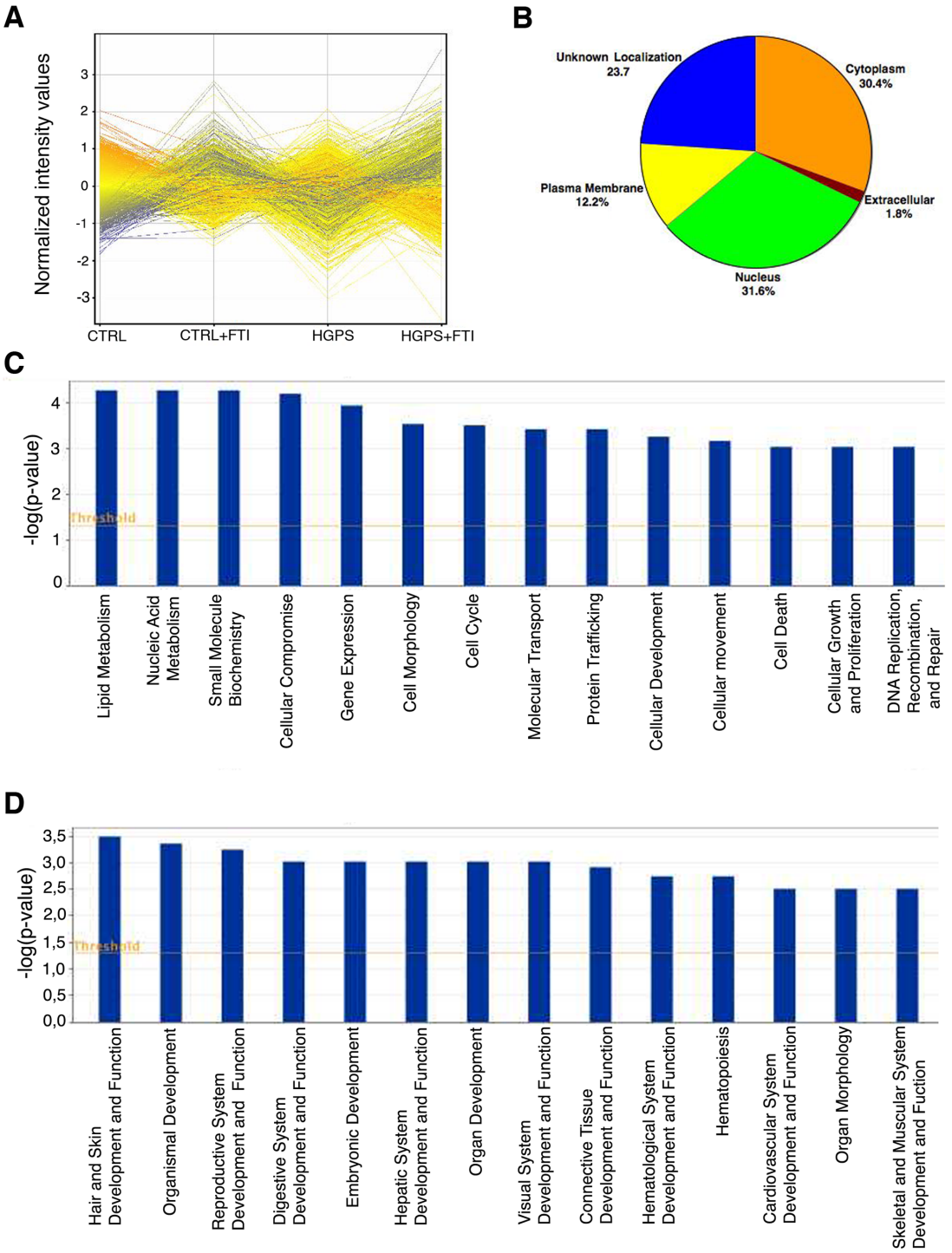
DNA repair (De Sandre-Giovannoli et al., 2003; Eriksson et al., 2003). Progerin toxicity is attributed at least in part to its farnesyl moiety, as chemical inhibitors of protein farnesyltransferase (FTIs) reverse abnormalities in nuclear morphology in progerin expressing cells (Capell et al., 2005; Toth et al., 2005). In addition, FTIs and other chemical inhibitors of protein prenylation partially reverse progeria-like phenotypes in genetically modified mice that express progerin or lack ZMPSTE24, and therefore accumulate unprocessed, farnesylated prelamin A (Capell et al., 2008; Fong et al., 2006; Varela et al., 2008; Yang et al., 2006).

While several studies have clearly implicated farnesylated progerin in HGPS, the precise molecular mechanisms of how it induces HGPS pathology remain to be

understood. Initial gene expression profiling of fibroblasts from human subjects with progeria syndromes and transfected cell models identified changes in sets of genes implicated in diverse pathways that have not always been consistent and have not been shown to be reversed by interventions such as treatment with FTIs (Csoka et al., 2004; Ly et al., 2000; Park et al., 2001; Scaffidi and Misteli, 2008). Therefore, our collaborators at Columbia University Medical Center, Dr. Djabali group (now the group is located at Technical University Munich) carried out additional genome-wide expression studies in cells from children with HGPS to identify alterations in functional groups of genes that define defective signaling pathways and to determine if FTI treatment reverses these defects. I have carried out the bioinformatics analysis and demonstrated a link between progerin and the retinoblastoma protein (Rb) signaling pathway in HGPS.

3.2 Experiment setup

Dermal fibroblasts from subjects with HGPS were obtained from the Progeria Research Foundation (www.progeriaresearch.org). The following fibroblasts were used: HGADFN003 (M, age 2), HGADFN127 (F, age 3), HGADFN155 (F, age 1), HGADFN164 (F, age 4) and HGADFN188 (F, age 2). Age-matched control dermal fibroblasts were obtained from Coriell Institute for Medical Research (Camden, NJ). The following cell lines were used: GM01652C (F, age 11), GM02036A (F, age 11), GM03349C (M, age 10), GM03348E (M, age 10), GM08398A (M, age 8). RNA was isolated from these five subjects with HGPS and five control individuals that were treated or untreated with FTI lonafarnib for three days. Affymetrix U133 plus 2.0 arrays were used for hybridization to get the gene expression profiles of genes in each condition and the data was analysed as described below. Experimental work was done by Jackleen Marji from Prof. Karima Djabali's group, Department of Dermatology, College of Physicians and Surgeons, Columbia University, New York, USA, Department of Dermatology, Technical University Munich, Munich, Germany. I have analysed the data using GeneSpring, MetaCore and bioCompendium.


A
Normalized intensity values
3
2
1
0
-1
-2
-3
CTRL
CTRL+FTI
HGPS
HGPS+FTI
B
Unknown Localization
23.7
Cytoplasm
30.4%
Plasma Membrane
12.2%
Extracellular
1.8%
Nucleus
31.6%
C
-log(p-value)
4
3
2
1
0
Threshold
Lipid Metabolism
Nucleic Acid Metabolism
Small Molecule Biochemistry
Cellular Compromise
Gene Expression
Cell Morphology
Cell Cycle
Molecular Transport
Protein Trafficking
Cellular Development
Cellular movement
Cell Death
Cellular Growth and Proliferation
DNA Replication, Recombination, and Repair
D
-log(p-value)
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
Threshold
Hair and Skin Development and Function
Organismal Development
Reproductive System Development and Function
Digestive System Development and Function
Embryonic Development
Hepatic System Development and Function
Organ Development
Visual System Development and Function
Connective Tissue Development and Function
Hematological System Development and Function
Hematopoiesis
Cardiovascular System Development and Function
Organ Morphology
Skeletal and Muscular System Development and Fuction

3.3 Data analysis

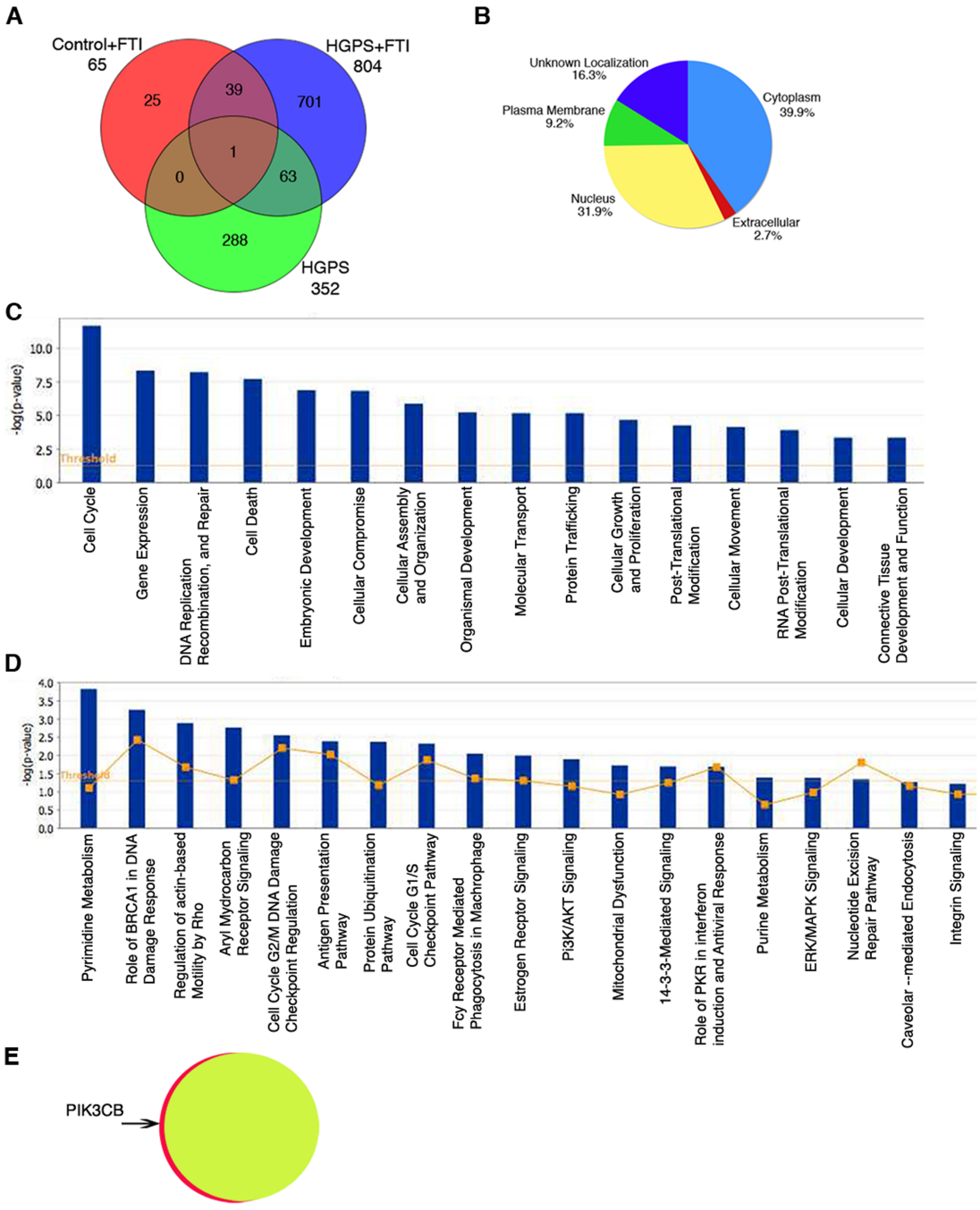
3.3.1 Microarray data analysis

The raw data was processed and analyzed using GeneSpring GX 10.0. I have identified 50,636 probe sets (Figure 3.3A) and selected the differentially expressed genes which are two-fold above or below the baseline and the P-value cutoff ≤ 0.01. As shown in Figure 3.4 from our Progeria publication (Marji et al., 2010), 352 genes were significantly differentially expressed between fibroblasts from subjects with HGPS and controls (gene list 1: `Control_versus_HGPS`).

65 genes were differentially expressed in control subjects after FTI treatment compared to no treatment (gene list 2: `Control_versus_Control+FTI`) and 804 in subjects with HGPS after FTI treatment compared to untreated controls (gene list 3: `Control_versus_HGPS+FTI`). But only one gene, PIK3CB was upregulated in cells from subjects with HGPS treated with FTI compared to cells from normal subjects treated with FTI (gene list 4: `ControL+FTI_versus_HGPS+FTI`).

I have analyzed above mentioned four gene lists with bioCompendium, annotated and enriched each gene list with knowledge obtained from several databases.

Figure 3.3 *(preceding page)*: **Genome-wide expression profiling of HGPS and control fibroblast cultures**. (A) Microarray plot profiles indicate changes in gene expression in control, HGPS, FTI-treated control and FTI-treated HGPS fibroblasts. Each continuous line corresponds to the normalized intensity value of an individual probe set. Line colors denote the intensity of the signal (red: strong and blue: low signal). Probes that satisfied a greater or less than two-fold cutoff and statistically significant difference of p,0.01 are displayed. (B) Pie chart indicates the predicted subcellular localization of proteins encoded by the 352 genes differentially expressed in HGPS. The list of differentially expressed genes in HGPS versus control cells was analyzed using Ingenuity Pathway Analysis (IPA) and encoded proteins assigned a subcellular localization based on information contained in the Ingenuity Knowledge Base. (C) Genes differently expressed in HGPS (352 genes) were assigned to diverse cellular functions using the 'Functional Analysis' tool of IPA software (www. ingenuity.com). Columns represent groups of genes associated with specific cellular functions (x-axis). The significant genes were compared to IPA database and ranked according to a p-values generated with Fisher's exact test. P-values less than 0.05 indicates a statistically significant, non-random association between a set of significant genes and a set of all genes related to a given function in IPA database. The ratio (y-axis) represents the number of genes from the dataset that map to the pathway divided by the number of all known genes ascribed to the pathway. The yellow line represents the threshold of p,0.05. (D) As described above, the 352 genes were assigned to diverse physiological systems according to IPA. ©This figure is from our own Progeria publication (refer section 3.6) (Marji et al., 2010) and used by following terms of use from *Public Library of Science (PLOS)*.


A
Control+FTI
65
HGPS+FTI
804
HGPS
352
25
39
701
1
0
63
288
B
Unknown Localization
16.3%
Cytoplasm
39.9%
Plasma Membrane
9.2%
Nucleus
31.9%
Extracellular
2.7%
C
-log(p-value)
10.0
7.5
5.0
2.5
0.0
Threshold
Cell Cycle
Gene Expression
DNA Replication Recombination, and Repair
Cell Death
Embryonic Development
Cellular Compromise
Cellular Assembly and Organization
Organismal Development
Molecular Transport
Protein Trafficking
Cellular Growth and Proliferation
Post-Translational Modification
Cellular Movement
RNA Post-Translational Modification
Cellular Development
Connective Tissue Development and Function
D
-log(p-value)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Threshold
Pyrimidine Metabolism
Role of BRCA1 in DNA Damage Response
Regulation of actin-based Motility by Rho
Aryl Mydrocarbon Receptor Signaling
Cell Cycle G2/M DNA Damage Checkpoint Regulation
Antigen Presentation Pathway
Protein Ubiquitination Pathway
Cell Cycle G1/S Checkpoint Pathway
Fcγ Receptor Mediated Phagocytosis in Machrophage
Estrogen Receptor Signaling
PI3K/AKT Signaling
Mitochondrial Dysfunction
14-3-3-Mediated Signaling
Role of PKR in interferon induction and Antiviral Response
Purine Metabolism
ERK/MAPK Signaling
Nucleotide Excision Repair Pathway
Caveolar --mediated Endocytosis
Integrin Signaling
E
PIK3CB

A permanent session has been created in bioCompendium with analysis results and is available at http://biocompendium.embl.de/Progeria. We have also analyzed these gene lists with MetaCore (www. Genego.com) and Ingenuity Pathways Analysis (www.ingenuity.com) software.

We first focused on the 'gene list 1' consisting of 352 differentially expressed genes in fibroblasts from controls and subjects with HGPS that were not treated with FTI. Of those genes, 306 were downregulated and 46 upregulated in fibroblasts from subjects with HGPS. The assigned subcellular localizations indicated that at least 31.6% of the gene products were localized to the nucleus, 30.4% to the cytoplasm, 23.7% to the plasma membrane and 1.8% to the extracellular matrix, with the remainder unknown (Figure 3.4B). Molecular function analyses (Figure 3.3C) and physiological distributions (Figure 3.3D) indicated that a significant number of genes were implicated in lipid metabolism, cell growth and differentiation, cell cycle, DNA replication and repair as well as cardiovascular

Figure 3.4 *(preceding page)*: **Genome-wide expression profiling of FTI-treated and untreated fibroblasts from subjects with HGPS**. (A) Venn diagram comparison of microarray datasets of controls versus control-FTI, or HGPS and/or HGPS-FTI. 65 genes were differentially expressed in fibroblasts from control subjects after FTI treatment compared to no treatment (Control+FTI; red); 352 in fibroblasts from subjects with HGPS compared to those from control subjects (HGPS; green); and 804 in fibroblasts from subjects with HGPS after FTI treatment compared to untreated controls (HGPS+FTI; blue). Genes with altered expression common between these different datasets are shown as areas of overlap with different colors in the Venn diagram with numbers of common genes indicated. (B) Pie chart indicates the subcellular localization of the encoded proteins of the differentially expressed genes between control and HGPS-FTI datasets according to information contained in the Ingenuity Knowledge Base. (C) Genes differentially expressed between controls versus HGPS-FTI datasets were assigned to diverse cellular functions according to Ingenuity Pathway Analysis (IPA) and (D) were associated to canonical pathways according to IPA. The significant genes were compared to IPA database and ranked according to p-values generated with Fisher's exact test. P-values less than 0.005 indicate a statistically significant, non-random association between a set of significant genes and a set of all genes related to a given function in IPA database. The ratio (y-axis) represents the number of all known genes ascribed to the pathway. The Yellow line represents the threshold of p,0.05. (E) Comparison of gene expression alterations between FTI-treated control cells (yellow) with FTI-treated fibroblasts from subjects with HGPS (red). Using criteria of corrected p-value from unpaired t-test, 0.01 and two-fold change in expression, only one gene, PIK3CB was upregulated in FTI treated cells from subjects with HGPS compared to cells from normal subjects treated with FTI. Therefore, 99% of the genes in FTI-treated cells from subjects with HGPS were expressed at the same levels as in FTI-treated normal fibroblast.

system development (Marji et al., 2010).

3.3.2 Network analysis

Protein interaction network analysis using MetaCore database (www.genego.com) shown in Figure 3.5 and STRING interaction database shown in Figure 3.6 identified Rb1 as the only one encoding a protein product, Rb, known to interact directly with A-type lamins (Mancini et al., 1994; Ozaki et al., 1994).

3.3.3 Comparison with related studies

As shown in Tables 3.1 and 3.2, the Differentially Expressed Gene (DEG) list in fibroblasts from subjects with HGPS from this study has been compared to the differentially expressed genes lists in studies by Scaffidi and Mitseli (Scaffidi and Misteli, 2008), Csoka et al., (Csoka et al., 2004) and Bakey et al., (Bakay et al., 2006). Control versus HGPS differentially expressed genes established after a statistical analysis using the t test with 1% significance and 2-fold cutoff were compared to the initial microarray analyses performed on three HGPS fibroblast strains derived from patients at age 8 (AG11513), 9 (AG10750) and 14 years old (AG11498) (Csoka et al., 2004). Only two genes: ATR, KIAA0746 are common. This small overlap may be due to variation intrinsic to each cell, in addition to the fact that cells from subjects with HGPS exhibit increased variation in cellular phenotype with cellular age in vitro and with the donor's age. Our study was performed using five fibroblast cultures from five subjects with HGPS at age 2 to 4 years old, kindly provided by Progeria Research Foundation. Cells were collected at an early population doublings stage (PPD <25) when their growth rate remained similar to that of control fibroblast cultures. Phenotypic variations and different levels of progerin expression can contribute to the heterogeneity between fibroblasts from subjects with HGPS. We compared the control versus HGPS gene list with a study from Scaffidi and Mitseli that used a cellular model for HGPS by overexpressing progerin in normal immortalized fibroblasts (Scaffidi and Misteli, 2008). We also compared our results with another study (Bakay et al., 2006). In this study, 125 human muscle biopsies from 13 diagnostic groups including Emery-Dreifuss muscular dystrophy (EDMD) patients with LMNA and emerin

mutations were studied and found very little overlap in both the cases (Bakay et al., 2006).

	Control vs HGPS(352 DEGs)	**(Scaffidi and Misteli, 2008)**(1394 DEGs)	**(Csoka et al., 2004)**(361 DEGs)	**(Bakay et al., 2006)**(232 DEGs)
Control vs HGPS	-	3.41%	0.57%	3.13%
(Scaffidi and Misteli, 2008)	0.86%	-	2.8%	0.79%
(Csoka et al., 2004)	0.55%	10.8%	-	1.66%
(Bakay et al., 2006)	4.74%	4.74%	2.59%	-

Table 3.1: **Comparison of Differentially Expressed Genes (DEGs) from different microarray studies.** Comparison of the DEGs in fibroblasts from subjects with HGPS from this study to the DEGs in studies (Bakay et al., 2006; Csoka et al., 2004; Scaffidi and Misteli, 2008) that have been previously published. The table shows the % overlap between DEGs derived from the different microarray studies.

3.4 Summary of results

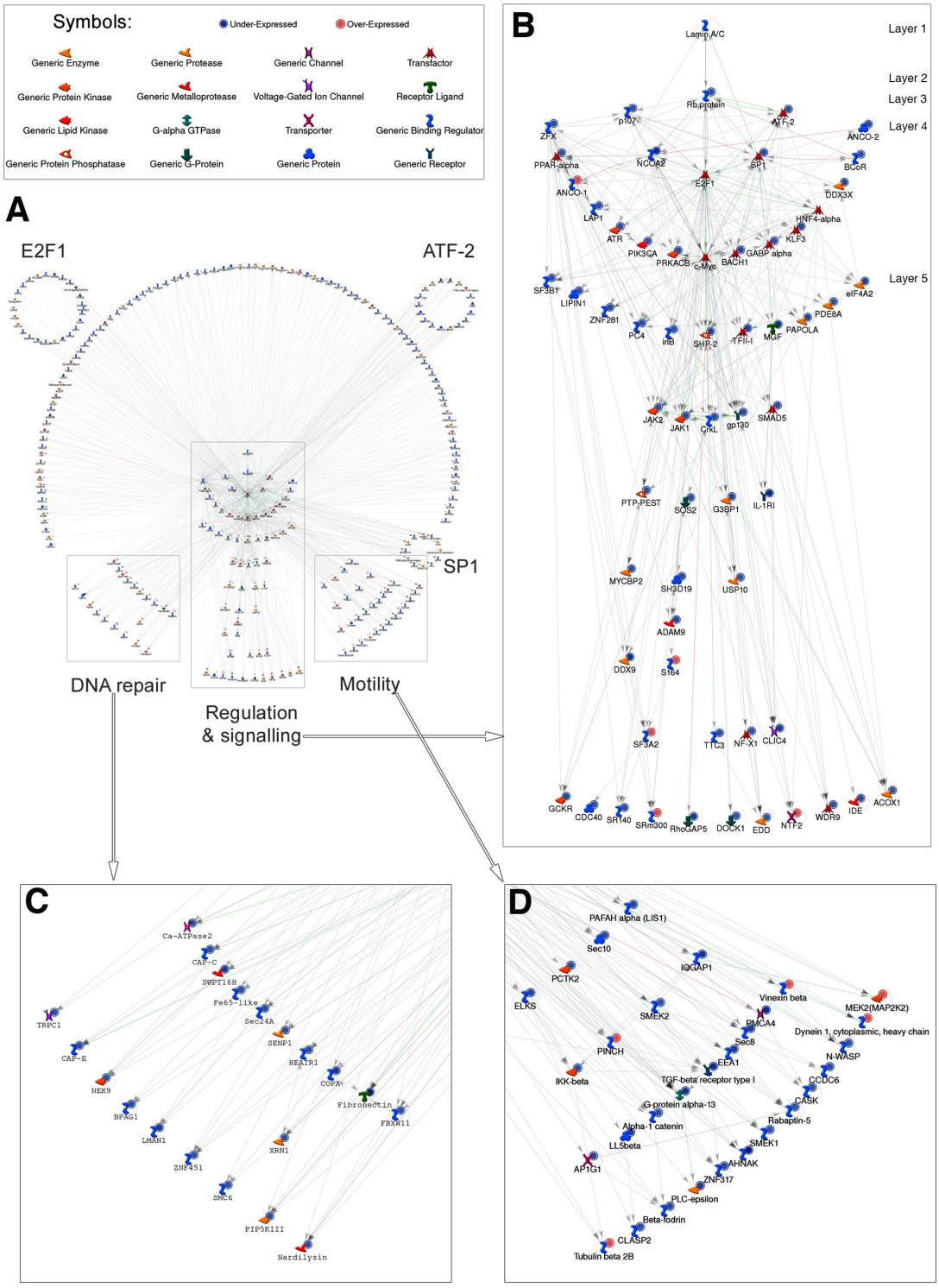
We analyzed global gene expression changes in fibroblasts from human subjects with HGPS and found that 352 genes were significantly differentially expressed between fibroblasts from subjects with HGPS and controls. Most of these genes are located in nucleus. Molecular function analysis and physiological distributions indicated that a significant number of genes were implicated in lipid metabolism, cell growth and differentiation, cell cycle, DNA replication and repair as well as cardiovascular system development. Protein interaction network analysis using MetaCore and STRING database identified Rb1 interact with A-type lamins. The expression of Rb1 was downregulated in HGPS.

The differential expression of most of the identified genes in HGPS can potentially be explained by the hypothesis that an abnormal prelamin A variant causes Rb to differentially interact with or regulate downstream partners. Importantly, the differentially expressed genes in HGPS indicated that Rb is a key regulatory component affected by LMNA mutation and that it is at the center of a signaling network that is abnormally active in the disease.

To unravel the mechanism underlying reversal of the HGPS phenotype by

	Control vs HGPS(352 DEGs)	(Scaffidi and Misteli, 2008)(1394 DEGs)	(Csoka et al., 2004)(361 DEGs)	(Bakay et al., 2006)(232 DEGs)
Control vs HGPS	-	ARFGEF1; BUB1; EIF4A2; MBNL2; RALGPS2; REV3L; SP1; SPRED1; SR140; TGFBR1; USP10; USP34	ATR; KIAA0746	ASPH; ATF2; ATP2B4; BRD7; GNS; KIAA0265; NFASC; NUP160; SF3B1; SOS2; UBE4B
(Scaffidi and Misteli, 2008)	-	-	AHR; BENE; C6orf31; COL13A1; COL4A1; COL4A2; CRTL1; DKFZP586K1520; DKFZp761D221; EMX2; ENH; HIP1; IFI27; IR2155535; ITGA2; KIAA1671; KIAA1912; MAIL; MGC15437; MGC21981; MGC24995; MGC2555; MTLC; NID2; PBX1; PGBD3; PLD1; PRG1; ROR1; SALL1; SGCD; SLC1A4; TBX3; TFPI; TNA; TNC; TP53I11; TRIM5; ZSIG13	EDG2; EPPB9; GALNACT-2; MGC2749; OSBPL1A; PBX1; PDE4DIP; SAV1; SGCD; WAC; ZFHX1B
(Csoka et al., 2004)	-	-	-	AASS; GALNT1; MEST; PBX1; SGCD; STK38L
(Bakay et al., 2006)	-	-	-	-

Table 3.2: **Common Differentially Expressed Genes (DEGs) from different microarray studies.** The table provides list of common DEGs in fibroblasts from subjects with HGPS from this study to the DEGs in studies (Bakay et al., 2006; Csoka et al., 2004; Scaffidi and Misteli, 2008) that have been previously published.


Symbols:
Under-Expressed
Over-Expressed
Generic Enzyme
Generic Protease
Generic Channel
Transfactor
Generic Protein Kinase
Generic Metalloprotease
Voltage-Gated Ion Channel
Receptor Ligand
Generic Lipid Kinase
G-alpha GTPase
Transporter
Generic Binding Regulator
Generic Protein Phosphatase
Generic G-Protein
Generic Protein
Generic Receptor
A
E2F1
ATF-2
SP1
DNA repair
Regulation & signalling
Motility
B
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Lamin A/C
Rb protein
p107
ATF-2
ZFX
ANCO-2
PPAR-alpha
NCOA2
SP1
BCoR
E2F1
ANCO-1
DDX3X
LAP1
HNF-4-alpha
ATR
PIK3CA
KLF3
PRKACB
c-Myc
GABP alpha
BACH1
SF3B1
eIF4A2
LIPIN1
ZNF281
PDE8A
PC4
IntB
SHP-2
TFII-I
MGF
PAPOLA
JAK2
JAK1
CrkL
gp130
SMAD5
PTP-PEST
SOS2
G3BP1
IL-1RI
MYCBP2
SH3D19
USP10
ADAM9
DDX9
St64
SF3A2
TTC3
NF-X1
CLIC4
GCKR
CDC40
SR140
SRm300
RhoGAP5
DOCK1
EDD
NTF2
WDR9
IDE
ACOX1
C
Ca-ATPase2
CAP-C
SWPT16H
Fe65-like
TRPC1
Sec24A
SENP1
CAP-E
HEATR1
COPA
NEK9
Fibronectin
BPAG1
FBXW11
LMAN1
XRN1
ZNF451
SMC6
PIP5KIII
Nardilysin
D
PAFAH alpha (LIS1)
Sec10
IQGAP1
PCTK2
ELKS
Vinexin beta
MEK2(MAP2K2)
SMEK2
PMCA4
Dynein 1, cytoplasmic, heavy chain
Sec8
PINCH
N-WASP
EEA1
IKK-beta
TGF-beta receptor type I
CCDC6
G-protein alpha-13
CASK
Alpha-1 catenin
Rabaptin-5
LL5beta
SMEK1
AP1G1
AHNAK
ZNF317
PLC-epsilon
Beta-fodrin
CLASP2
Tubulin beta 2B

blocking protein farnesylation, we determined the gene expression changes occurring in fibroblasts after FTI administration on both control as well as subjects with HGPS. Comparison of gene expression analysis of DEGs from control and HGPS subjects before and after the treatment of FTI indicated that FTI restores a nearly normal gene expression profile in fibroblasts from subjects with HGPS. Microarray analyses results were validated by real-time RT-PCR by Dr. Djabali's team and the results were available in our publication (Marji et al., 2010).

3.5 Discussion

Inhibition of protein farnesyltransferase activity by FTIs has been shown to improve abnormalities in nuclear morphology in cells from subjects with HGPS and from animal models of the disease (Capell et al., 2005; Dechat et al., 2007; Glynn and Glover, 2005; Mallampalli et al., 2005; Toth et al., 2005; Wang et al., 2008). FTIs also improve the progeria like phenotypes of mouse models of HGPS (Capell et al., 2008; Yang et al., 2006). Presumably, blocking farnesylation of progerin, the truncated prelamin A expressed as a result of the G608G LMNA mutation which is responsible for most cases of HGPS, renders it less toxic. However,

Figure 3.5 *(preceding page)*: **The lamin A-Rb network**. The main network (center left) shows downstream interactions between lamin A/C and the 352 differentially expressed genes in HGPS fibroblasts. The network was built using MetaCore analyses, which finds known interactions between gene products. (A) The network divides into several distinct regions, the main region being the regulatory and signaling network. (B) The detailed view of the main region (top right) shows that the only gene differentially expressed in HGPS directly downstream of lamin A/C (layer 1) is that encoding Rb (layer 2). In turn, most of the immediate downstream partners of Rb included in this dataset are p107, NCOA2, SP1, and ATF-2 and E2F1 (layer 3). Of the remaining genes that occur downstream of layer 3, nearly half of the 280 genes interact with only one or more of these transfactors associated to the main network (center left); these gene products are placed above the centre. From layer 4 and 5, most of the genes can be connected at least to one entity according to GeneGO. In the HGPS dataset, 124 genes that had no known interactions in GeneGO are not shown. From the center regulatory and signaling network (zoom left, panel B)), several groups of genes segregate into six subnetworks, based on mutual interactions: a group denoted DNA repair (bottom left, panel (C)), motility (bottom right, panel (D)), and three circles of genes regulated by E2F1, ATF-2 and SP1, respectively. Symbols associating the genes with functions are indicated. Genes labeled with blue circles were downregulated and red circles were upregulated.

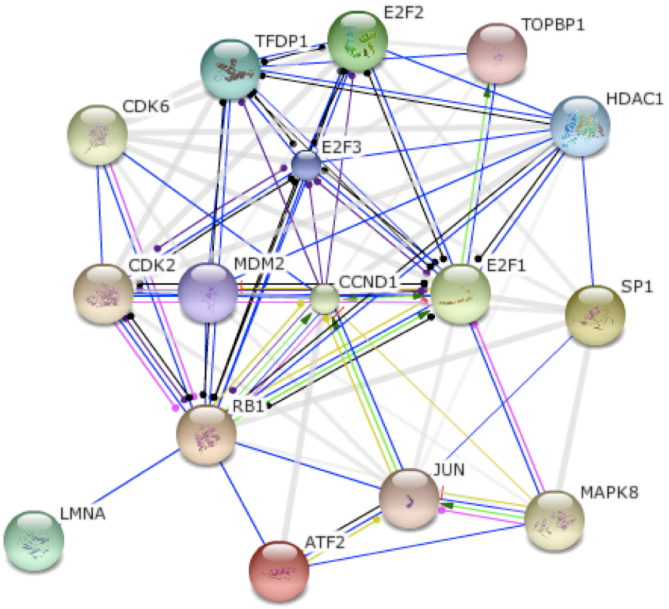


Figure 3.6: **The lamin A-Rb network from STRING database**. The protein-protein interaction network of genes differentially expressed in HGPS fibroblasts. This network was built using STRING database. Similar to MetaCore network shown in Figure 3.5, STRING network also reveals that lamin A - Rb signaling network is a major defective regulatory axis in HGPS.

the molecular mechanisms by which progerin exerts its toxicity and whether FTI treatment reverses specific molecular defects induced by progerin remain largely unknown. We have now identified a defective lamin A-Rb signaling network in cells from subjects with HGPS and demonstrated that treatment of cells with an FTI reverses abnormalities in the expression of genes encoding proteins in this network.

We compared our gene expression profiles to those of previous studies that have examined gene expression in HGPS and found very little overlap (3.5%) of differentially expressed genes with Scaffidi and Misteli (Scaffidi and Misteli, 2008), 3% with Bakey et al., (Bakay et al., 2006) and less than 1% with Csoka et al., (Csoka et al., 2004)(Table 3.1). The corresponding common differentially expressed genes in these comparisons are shown in Table 3.2. (Csoka et al., 2004) used fibroblasts from only three subjects ages 8 to 14 years of age whereas we used cultured fibroblasts from five subjects with HGPS obtained at earlier ages

(age 1 to 4 years). As the growth rate and proliferation potency of fibroblasts from subjects with HGPS decrease with cellular age in vitro and the donor's age, this could partially explain the differences between our results. Importantly, the average age of death in HGPS is 12 - 15 years, which suggests that the cellular phenotype may be severely altered in older children (Merideth et al., 2008). (Scaffidi and Misteli, 2008) generated immortalized normal fibroblasts that overexpressed progerin and examined gene expression in only two lines, making a comparison difficult given the very different system and the low statistical power from using only two biological replicates.

Our results, therefore, suggest that lamin A - Rb signaling network is a major defective regulatory axis. Treatment of fibroblasts with a protein farnesyltransferase inhibitor (lonafarnib) reversed the gene expression defects. This study identifies Retinoblastoma protein (Rb) as a key factor in HGPS pathogenesis and suggests that its modulation could help in premature ageing and as well as physiological ageing.

Contributions: This is a collaborative project between scientists from 1) Department of Dermatology, College of Physicians and Surgeons, Columbia University, New York, USA; 2) EMBL, Heidelberg, Germany; 3) Departments of Medicine and of Pathology and Cell Biology, College of Physicians and Surgeons, Columbia University, New York, USA; 4) Department of Pediatrics, Warren Albert Medical School of Brown University, Providence, Rhode Island, USA and 5) Department of Dermatology, Technical University Munich, Munich, Germany. This research project was conceived and designed the experiments by Karima Djabali. Experimental work was done by Jackleen Marji and bioinformatics analysis was carried out by me under the guidance of Sean I. O'Donoghue and Reinhard Schneider.

3.6 Progeria publication

This research work has been published in PLOS ONE journal (Marji et al., 2010) and the article is provided in the Publications section 9.3.

Chapter 4

TAMAHUD

4.1 Description

Identification of early disease markers, novel pharmacologically tractable targets and small molecule phenotypic modulators in Huntington's Disease (HD) - TAMAHUD (TArgets and MArkers in HUntington's Disease) is a FP6 European Commission funded project. The main objective of this project is to improve strategies for the prevention and cure of HD by using genome information to better understand the molecular mechanisms underlying the pathology. It employed a state-of-the-art technology, a phenotypic screen to identify human genes encoding druggable proteins in a human cell line overexpressing a full length mutated Huntingtin construct.

In cellular and animal models of the disease, as well as in patients, intracellular aggregates containing different proteins including the polyQ-bearing amino-terminus of the mutated HTT protein are detectable. In fact, the proteolytic generation and accumulation of such HTT fragments correlates well with cytotoxicity in cellular and animal models of the disease (Crook and Housman, 2013). Extensive genetic and transgenic data support the idea that the mutation causes disease predominantly by a toxic gain-of-function mechanism, although it is possible that this may be modulated to a minor extent by loss-of-function (Di Prospero and Fischbeck, 2005; Landles and Bates, 2004).

Many different possibilities have been proposed as molecular mechanisms un-

derlying HD, including excitotoxicity, loss of transcriptional activity of genes encoding neurotrophic factors, loss of axonal transport functions, gain of protein-protein interactions leading to transcriptional dysregulation and altered cholesterol metabolism (Bates, 2003; Cattaneo et al., 2001; Li and Li, 2004; MacDonald et al., 2003; Sugars and Rubinsztein, 2003). Although genetic testing can identify individuals carrying the HD allele (Myers, 2004), accurate prediction of disease onset is presently very difficult to achieve.

Current therapeutic approaches in HD are largely unspecific and aimed at addressing the psychiatric symptomatology and reducing motor dysfunction (Leegwater-Kim and Cha, 2004). Some experimental therapeutic approaches have shown a degree of benefit in animal models (Melone et al., 2005), modestly increasing parameters such as life-span, motor performance and weight loss. However, where such approaches have been tested in the clinic, little or no significant benefit could be demonstrated. Today, therefore, individuals diagnosed with HD and their families face the certainty of an incurable, fatal, progressive and devastating neurodegenerative disease, without knowing either when disease onset will occur or how severe and prolonged disease progression will be.

4.2 Experimental setup

Although primary cultures of mammalian striatal neurons, or striatum-derived cell lines with the HD mutation would represent an ideal cellular tool for target identification approaches, the limited accessibility and/or capability to manipulation (robustness in culture and transfection efficiency) afforded by these systems make them better suited as target validation tools (i.e. useful for the validation of previously identified targets) rather than for high-throughput screens aimed at target identification.

Sienabiotech and University of Cambridge developed a stable human embryonic kidney (HEK293) T-REx cell line which can inducibly express either a full length HTT(Q138) under the control of a tetracycline-inducible promoter or a wild-type HTT transgene. Here, we combined two approaches to identify modifiers of mutant HTT toxicity by first performing a cell-based screen to identify genes that, when knocked down, could suppress mutant HTT-induced toxicity,

using a library of 16,869 siRNAs targeting 5,623 genes based on the RefSeq annotation of the human genome. We performed this screen in two different HD models. The primary high-throughput, high-content siRNA screen was carried out by CENIX, an RNAi-based specialist company which provided the consortium with the 'druggable genome' siRNA library, the experience, know-how and specialist technical competencies required for a high throughput RNAi screen (Sachse et al., 2005; Sonnichsen et al., 2005). In a secondary analysis, Dr. Rubinsztein and his team at University of Cambridge validated primary hits in a Drosophila model of HD.

4.3 Data integration and analysis

For the high-throughput screen, we used a strategy consisting of an iterative siRNA screen where positive genes were selected after three consecutive rounds to compensate for the variability of the assay. We eliminated nonpositive siRNAs and added new siRNAs targeting the selected genes in consecutive passes. We assessed rescue of cellular toxicity by each siRNA by fluorescence microscopy and automated image analysis using three independent readouts, (i) number of cell nuclei (#nuclei), (ii) apoptotic index and (iii) aberrant nuclei index, and we used rescue indices to express the effect of each individual siRNA for each parameter analyzed. In an initial screen, we tested three independent siRNAs for each of the 5,623 genes (a total of 16,869 siRNAs), from which we selected 670 primary genes (see Supplementary Note 1 of our TAMAHUD publication (Jimenez-Sanchez et al., 2015a,b) for screen assay and criteria selection) for a second pass 2 validation. After three consecutive rounds of screening, we selected 257 genes and ranked these on the basis of all three rescue indices, using #nuclei rescue index as a primary criterion (see Supplementary Data Set 1 of our TAMAHUD publication (Jimenez-Sanchez et al., 2015a,c)).

To validate these 257 hits obtained in mammalian cells and to focus on targets with potential relevance in vivo, we performed a secondary screen in a Drosophila model that expressed a construct of the HTT protein containing 48 polyglutamines (Q48) that causes eye degeneration. Out of these 257 mammalian genes previously selected, 133 genes have one or more orthologs in flies (check Supple-

mentary Data Sets 1 and 2 of TAMAHUD publication (Jimenez-Sanchez et al., 2015a,c,d)). Of these 133 mammalian genes with fly orthologs, 74 Drosophila genes (corresponding to 66 mammalian genes) rescued the Q48-induced eye degeneration with at least one RNAi line, whereas the others showed no obvious effect (Supplementary Fig. 3a,b and Supplementary Data Sets 1 and 2 of TAMAHUD publication (Jimenez-Sanchez et al., 2015a,c,d,e)).

An important added value to TAMAHUD project is to have expression and functional data on the same disease model and a comparison with public data. In order to support this siRNA screen target identification and validation efforts, I have selected and analysed 17 publicly available Huntington's Disease studies from human, mouse and rat species and the list of these studies is shown in Table 4.1. HEK293 T-REx cell line siRNA hits 670 obtained in pass 1, then 257 in pass 2 and 74 genes from Drosophila model have been integrated with microarray results obtained from same cell line before and after the treatment of doxycycline. In addition the transcriptome results from selected already published HD datasets were also integrated into the same database. The hits from TAMAHUD screen both functional as well as expression and from public HD sets were extensively annotated with the knowledge from publicly available databases. This consolidated database and corresponding web application facilitate cross study analysis, verification of experimental findings and considerably strengthen them. Some of my contributions are mentioned below.

Annotation pipeline: An annotation pipeline has been set up to annotate the TAMAHUD screen targets using a wide range of biological databases (gene, protein, chemistry, pathway, disease, gene expression, domain, gene-ontology, protein-protein, protein-chemistry interaction databases).

Web server: A secure interactive web server has been developed using MySQL database as backend, CGI-Perl and JavaScript with access control. It provides easy access to all the project data to the collaborators of TAMAHUD to the hits information from both experimental data as well as public microarray datasets and their annotations. These datasets are accessible via distinct logical sections as shown in Figure 4.1. In additional several published Hungtinton's disease experimental datasets shown in Table 4.1 have been selected. These datasets were analysed by using in house developed R scripts and integrated with TAMAHUD

screen to facilitate cross study analysis and poolig of the data. A plugin is developed to connect TAMAHUD web application to bioCompendium resource (refer to Chapter 2 for more details on bioCompendium) with the help of bioCompendium API. Using this plugin, the selected datasets were further analysed with bioCompendium. Gene expression dataset selection and connection to bioCompendium is shown in Figure 4.2. The URL of this web server is `https://schneider2.embl.de/tamahud`.

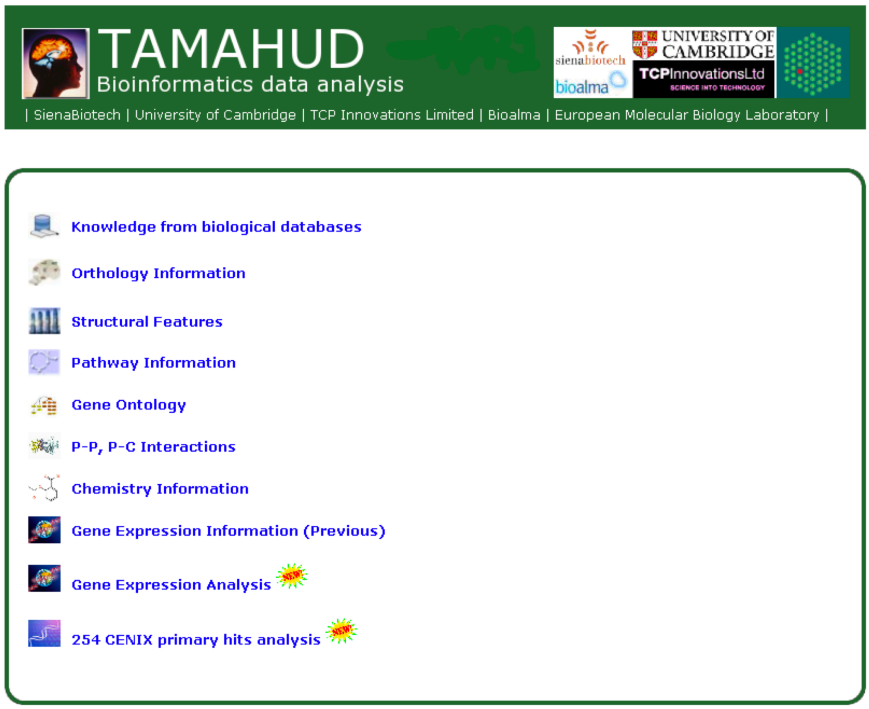


Figure 4.1: **TAMAHUD web interface**. The figure shows the landing page of the TAMAHUD web application. Here TAMAHUD siRNA screen hits were annotated with knowledge from several biological databases and organised the annotations and enrichments in different sub sections. It also provides the entry point to gene expression analysis of TAMAHUD and public HD datasets shown in Table 4.1.

Summary of known information from biological databases: The different target groups (Cenix selected, passed 1-4 passes) are mapped to Ambion siRNAs and provided a comprehensive summary sheet for each gene.

Orthology information: In order to facilitate the work on more than one animal model, the orthology information from important model organisms are

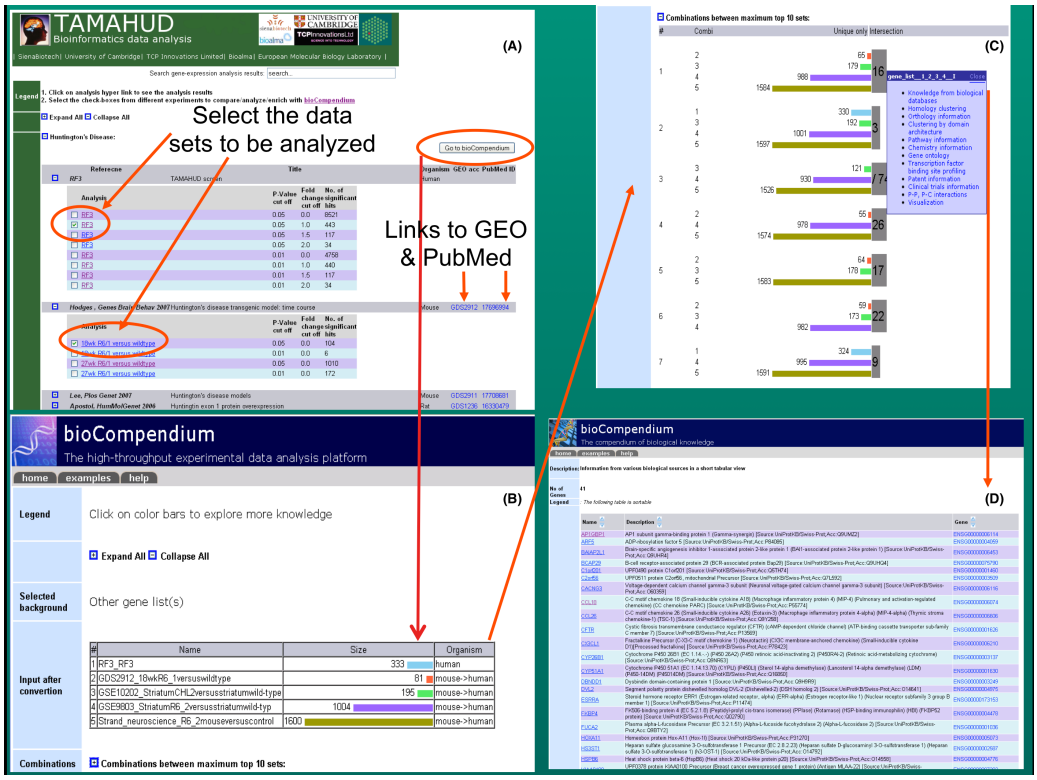


Figure 4.2: **TAMAHUD and Public Expression Data Analysis Interface**. In this figure panel (A) shows the organisation of different Huntington's Disease microarray analysis results with different p-value, fold-change cut offs and number of significant transcripts with a hyper-link. This link opens a table view of significant hits with detailed annotations. This panel also provides link to the corresponding GEO accession and the publication. From here the user can select one or more transcript lists and send them to the bioCompendium service by selecting the corresponding checkboxes and clicking 'Go to bioCompendium' button and are highlighted with red ellipses. Panel (B) shows bioCompendium data exploration and analysis interface. Here the table 'Input after conversion', shows the names of the selected gene lists and their sizes after mapping them to human genes by following orthology relations. In this particular example RF3 is the differentially expressed dataset from human embryonic kidney (HEK293) T-REx cell line before and after the treatment of doxycycline and its corresponding organism is human, where as other four selected public HD datasets are from mouse models and bioCompendium converts them from mouse to human by following orthology relations. At this point all the gene lists are converted to human specific genes, easy to pool or compare and further explore with the help of exhaustive menu provided by bioCompendium by clicking on the coloured bars. The 'Combinations' expand, collapse button provides union or intersection of the gene lists as shown in panel (C). The bioCompendium menu shown in panel (C) provides summary sheets, clustering based on sequence and domain architecture similarity, KEGG and GO enrichments etc., (refer to Chapter 2 Bar diagrams and menu section for more details). Panel (D) shows a tabular view of the selected list of genes and a summary sheet will be open upon selection of a gene. This is an example to explore bioCompendium menu items.

provided.

Structural features: This section contains protein domain features from SMART (Letunic et al., 2009) and Pfam (Punta et al., 2012), and 3D structures from PDB (Berman et al., 2007).

Pathway analysis: Hits are analysed with KEGG (Kanehisa et al., 2012), PANTHER (Mi et al., 2005), Reactome (Croft et al., 2011), IPA (Ingenuity Pathway Analysis) resources and are enriched for each pathway, hits are overlaid in KEGG pathways. The pathway analysis with IPA shown in Figure 4.3.

Gene ontology: Detailed gene ontology information (function, process and cellular component) is provided for each hit.

P-C, P-P interactions: Protein-protein, protein-chemistry information from STRING (Szklarczyk et al., 2011) and STITCH (Kuhn et al., 2008) are provided for each hit with a graphical network visualization.

Chemistry information: Chemistry related data are provided from PubChem (Bolton et al., 2008), DrugBank (Wishart, 2008), HMDB (Wishart et al., 2009), Bioalma. This information is further categorized into drugs, ligands and metabolites.

Gene expression information: Gene expression related data to TAMAHUD hits were collected from EMBL-EBI Expression Atlas (`https://www.ebi.ac.uk/gxa`), which provides normalized data over different gene expression experiments. Here I have analyzed the data related to our hits and provided the inference in different sections as follows: a) Brain related - more precisely, the brain parts where our TAMAHUD hits are expressed, b) Disease related - to point out other diseases associated to our hits, c) Huntington's disease related - to know if there are any other proteins of interest in this set or to detect how many of our hits are matching genes that are significantly expressed in these gene expression experiments.

Bioalms novoseek: An information extraction system for searching the published knowledge in biomedical literature from Bioalma is incorporated into the TAMAHUD web server.

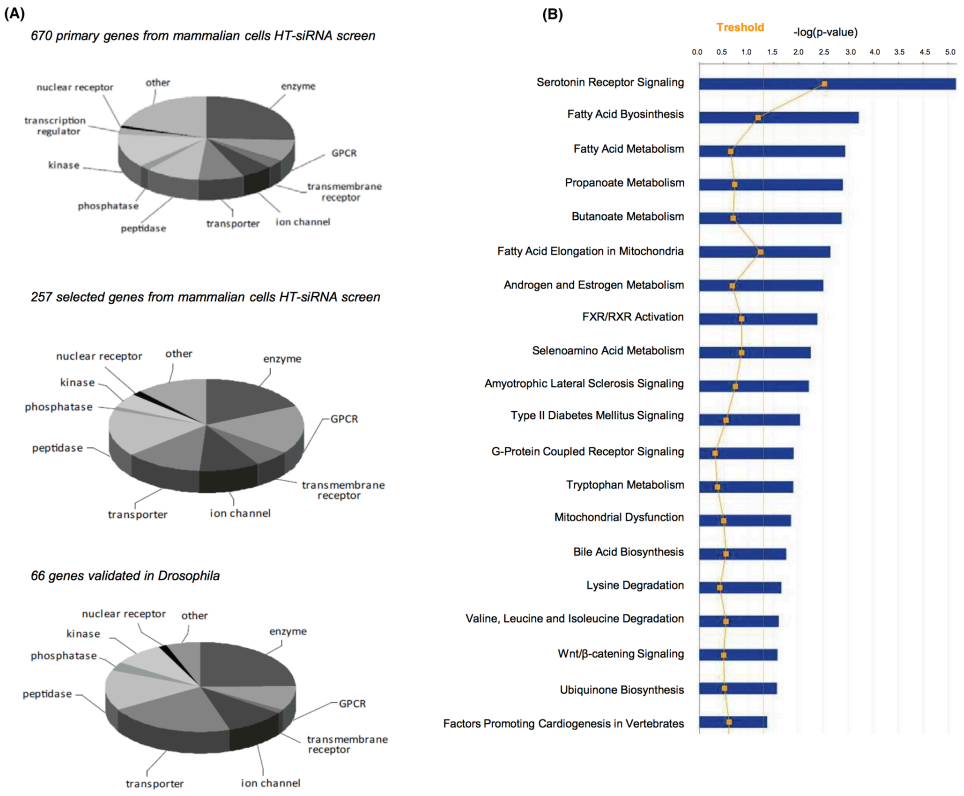


Figure 4.3: **Functional categorization of mutant HTT toxicity modifiers**. (A) Pie chart representation of the main molecular functions of genes obtained from siRNA screen in HEK293 cells (670 primary and 257 further selected genes) (top and middle) and 66 validated genes in Drosophila (bottom). (B) Top functional pathways associated with the 66 hits validated in Drosophila. Canonical pathways were determined by an Ingenuity Pathway Analysis and were ranked by log(p-value). Fisher's exact test was used to calculate a p-value (blue bars). The threshold line represents a p-value of 0.05, canonical pathways below the yellow line are not statistically significant. Number and name of the genes associated with each of the top 30 pathways are shown in Supplementary Table 1b. of TAMAHUD publication (Jimenez-Sanchez et al., 2015a,e).

GEO Acc	Title	Summary	Organism	Platform	Reference
GSE3621	Huntington's disease transgenic model: time course	Analysis of brains of R6/1 transgenic animals at 18 to 27 weeks of age. The R6/1 transgenic animal expresses exon 1 of the mutant human huntingtin gene that contains 115 CAG trinucleotide repeats. Results provide insight into the early pathogenesis of HD	*Mus musculus*	Affymetrix Mouse Genome 430 2.0 Array	(Hodges et al., 2008)
GSE3583	Huntington's disease models	Comparison of striatal cells harboring a mutant huntingtin protein with 111 glutamines to wild type striatal cells treated with 3-nitropropionic acid (3-NP). The size of the polyglutamine tract in huntingtin is correlated with mitochondrial function. 3-NP is a respiratory chain inhibitor	*Mus musculus*	Affymetrix Mouse Genome 430 2.0 Array	(Lee et al., 2007)
GSE2602	Huntingtin exon 1 protein overexpression	Analysis of Htt14A2.5 cell line expressing a truncated form of huntingtin (Htt) exon 1 protein controlled by a promoter inducible by ponasterone (PA). Expression examined 48 hours after treatment with 5 uM PA. Htt expression in Htt14A2.5 results in protein aggregation	*Rattus norvegicus*	Affymetrix Rat Genome U34 Array	(Apostol et al., 2006)
GSE857	HD and combination drug therapy	Whole brain hemispheres from Huntington's disease model R6/2 transgenic mice treated with creatine, tacrine and moclobemide from 5 weeks of age. Drug therapy was aimed at boosting neurotransmitter levels and improve cognitive functions	*Mus musculus*	Affymetrix Murine Genome U74A Version 2 Array	(Morton et al., 2005)
GSE10581	PC12 cell model of HD	Nrf2-related oxidative stress response and impaired dopamine biosynthesis in a PC12 cell model of Huntington's disease	*Rattus norvegicus*	Illumina ratRef-12 v1.0 expression beadchip	(van Roon-Mom et al., 2008)
GSE10202	Genetic mouse models of HD	Striatal gene expression data from 22-month-old CHL2 mice and control mice	*Mus musculus*	Affymetrix Mouse Genome 430 2.0 Array	(Kuhn et al., 2007)
GSE9803 GSE9804	Genetic mouse models of HD	Striatal gene expression data from 12 weeks-old R6/2 mice and control mice (set 1 and 2)	*Mus musculus*	Affymetrix Mouse Genome 430 2.0 Array	(Kuhn et al., 2007)
GSE9375	Genetic mouse models of HD	Striatal gene expression data from 12 months-old Hdh4/Q80 mice and control mice	*Mus musculus*	Affymetrix Murine Genome U74A Version 2 Array	(Kuhn et al., 2007)
GSE7958	Genetic mouse models of HD	Striatal gene expression data from 3- and 18-month-old Q92 mice and control mice	*Mus musculus*	Affymetrix Mouse Genome 430 2.0 Array	(Kuhn et al., 2007)
GSE11139	Mouse cell model of HD	Elucidating a normal function of huntingtin by analysis of huntingtin-null mouse embryonic fibroblasts	*Mus musculus*	Sentrix Mouse-6 Expression BeadChip (2)	(Zhang et al., 2008)
GSE3790	Post mortem human brain tissue model of HD	Human cerebellum, frontal cortex [BA4, BA9] and caudate nucleus HD tissue experiment	*Homo sapiens*	Affymetrix Human Genome U133A and U133B Arrays	(Strand et al., 2007a,b)
GSE1767	HD blood expression profile	Analysis of peripheral blood samples of 5 presymptomatic and 12 symptomatic Huntington's disease (HD) patients	*Homo sapiens*	GE Codelink Human Uniset I, II, and 20K	(Borovecki et al., 2005)
GSE11358	HD expression profile	Histones associated with downregulated genes are hypo-acetylated in Huntington?s disease models	*Mus musculus*	Affymetrix Mouse Genome 430 2.0 Array	(Sadri-Vakili et al., 2007)
GSE12481	HD expression profile	TRE-Htt-N853 Huntington's Disease in vitro model	*Rattus norvegicus*	Affymetrix Rat Genome 230 2.0 Array	(Runne et al., 2008)
	HD transgenic mice expression profile	The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice.	*Mus musculus*	Illumina Mouseref-8 Expression Beadchips v1	(Thomas et al., 2008)
	HD transgenic mouse models	Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects	*Mus musculus*	Affymetrix Mu 11K GeneChip	(Luthi-Carter et al., 2002)

Table 4.1: **Selected and Analysed Public Huntington's Disease Studies** - The table listing the manually selected and in house analysed Huntington's disease datasets. For most of the studies, the corresponding raw data has been downloaded from NCBI Gene Expression Omnibus (GEO), only for few studies data has been downloaded from the publications. These datasets were analysed by using in house developed R script and integrated with TAMAHUD screen to facilitate cross study analysis and poolig of the data.

4.4 Summary of results

Cenix, an RNAi-based specialist SME performed the primary cell screen for suppressors of mutant HTT toxicity using a stable human embryonic kidney (HEK293) T-REx cell lines expressing full-length human HTT bearing 138 polyglutamines (Q138) under the control of a tetracycline-inducible promoter, which we refer to as HTT(Q138). The expression of HTT(Q138) confirmed after inducing the cells with doxycycline using antibodies recognizing the N terminus of human HTT and quantitative RT-PCR using primers spanning different areas of the human HTT cDNA. This cell line had reduced cell viability after expression of mutant HTT, which was reverted through treatment with a known reference compound (Y27632) (Li et al., 2009b), suggesting that this model could be used to identify potential modulators of mutant HTT cellular toxicity in a large-scale screen.

We have selected 'druggable genome' human siRNA library, a total of 16,869 siRNAs covering 5,623 human genes encoding pharmacologically tractable proteins including G-protein coupled receptors, ion channels, enzymes and nuclear receptors (Hopkins and Groom, 2002). The first pass of the screen provided 670 significant genes and after three consecutive rounds of screening, we selected 257 genes for further validation. To validate these hits the University of Cambridge team performed a secondary RNAi screen in a Drosophila model of HD and came up with 74 Drosophila genes that are corresponding to 66 mammalian genes. And the same time Sienabiotech SpA performed the microarray analysis of HTT(Q138) cells before and after the treatment of doxycycline (TAMAHUD RF3 dataset).

To gain further insight into the biological relevance of the data generated, TAMAHUD knowledgebase based on both public and experimental data has been developed by me at EMBL under the supervision of Reinhard Schneider by following data integration and knowledge management approaches. Here I have performed bioinformatics analysis of these significant druggable targets obtained from siRNA screen through passes 1-4 that are 670 genes, 257 and 66 genes respectively. As shown in Table 4.1, I have also selected 17 published Huntington's disease studies related to human, mouse and rat species and analyzed them using

R & BioConductor along with TAMAHUD RF3 dataset. I have applied data integration and knowledge management approaches to integrate TAMAHUD experimental results with public high throughput experimental datasets. All these datasets and bioinformatics results were systematically labelled and stored in a MySQL database and a Perl-CGI based web application have been developed with rich JavaScript functionality. The TAMAHUD hits as well as DEGs (Differentially Expressed Genes) were annotated with information from several biological databases such as gene, protein, chemistry, pathway, disease, gene expression, domain, gene-ontology, protein-protein, protein-chemistry interaction databases. The annotated TAMAHUD hits and their enrichments were organised in different logical sections as depicted in Figure 4.1. The gene expression datasets were organised in an expandable tableview along with number of significant DEGs and their corresponding p-value and fold-change cut offs. The user can select one or more DEG lists and send them to the bioCompendium service on the fly for further exploration and analysis as shown in Figure 4.2.

There is a lot of literature on Huntington's disease and it's difficult to follow this vast amount of growing information manually. Bioalma mined the relevant literature by using their text-mining methods and stored the text-mining results in Alma Knowledge Server (AKS). These text-mining results were also integrated with above web application by using AKS API.

With the help of this centralized TAMAHUD database and the web server, we categorized the different sets of HD toxicity modulators according to their molecular function. Suppressors were enriched for certain classes of proteins such as GPCRs or transporters compared to the initial library, whereas the number of positive kinases in the screen was reduced, and no cytokines, growth factors or translational regulators were represented. We observed similar functional categorizations after selection from the cell and Drosophila screen. An Ingenuity Pathway Analysis of the hits obtained in the primary screen in cells shown in Figure 4.3 revealed that the majority of these proteins participate not only in general processes such as GPCR- or cAMP-mediated signaling but also in canonical pathways related to neurodegeneration such as apoptosis, mitochondrial dysfunction, amyloid processing or protein ubiquitination. Notably, ten of these proteins have been previously related to HD signaling, including subunits of the

succinate dehydrogenase complex and HTT-associated protein 1 (Supplementary Table 1a of TAMAHUD publication (Jimenez-Sanchez et al., 2015a,e)). Many of the genes validated in Drosophila are also involved in processes related to neurodegeneration but are enriched in mitochondrial metabolic pathways, especially those associated with fatty acid biosynthesis and metabolism.

With the help of TAMAHUD knowledgebase, a gene that had one of the strongest and most consistent effects in rescuing mutant HTT-induced toxicity in the cell-based siRNA screen has been prioritized for further experimental validation. The gene product has glutaminyl cyclase activity and is named QPCT. The University of Cambridge team headed by Dr. David C Rubinsztein further validated the functionality of QPCT and screened the library of compounds that inhibit QPCT. Refer to TAMAHUD publicaiton in section 4.6.

4.5 Discussion

In this collaborative project, our approach using a two-step screen, starting with an initial large-scale analysis in human cell models followed by validation in Drosophila, has yielded a number of potentially druggable targets which may be suitable for HD. A variety of high-throughput RNAi screens have identified genetic suppressors of phenotypes mediated by mutant HTT N-terminal fragments in Drosophila, Caenorhabditis elegans and mammalian (mouse and human) cells (Lejeune et al., 2012; Miller et al., 2010, 2012; Yamanaka et al., 2014). In most cases, aggregation was the primary readout, often measured with C-terminal GFP fusions. Differences in the nature of the previous screens (species, cellular context, HTT fragment length, length of the polyglutamine expansion, primary readout, and differences in siRNA or shRNA sequences) complicate cross-screen comparisons. Also, virtually no screens in this area have examined their false negative rates owing to inefficient knockdown. Additionally, the screen presented here was biased toward the druggable component of the human genome, and a further selection was made in the course of sorting toward specific protein target classes. This most likely contributes to the relatively poor overlap of hits in the present and previous screens. A comparison with a screen performed in HEK293T cells to identify genetic suppressors of inducibly expressed mutant HTT exon 1 toxicity (Miller

et al., 2012) revealed an overlap of only four genes (CPA1, GRIN2A, NR3C2 and USP21) when considering the top 257 hits (check Supplementary Data Set 1 of TAMAHUD publication (Jimenez-Sanchez et al., 2015a,c)). However, matrix metalloproteases, identified in HEK293T cells as modulators of fragmentation and toxicity of N-terminal portions of mutant HTT (Miller et al., 2010), were also identified in our data set together with PAK1, which we previously identified as a kinase promoting mutant HTT self-association and toxicity (Luo et al., 2008), thus validating the effectiveness of the screen.

With the help of TAMAHUD knowledgebase, given the reproducible and clear rescue that QPCT inhibition exerts on mutant HTT toxicity in cells and in Drosophila, we focused on this target. We identified and characterized a series of compounds that efficiently reduce mutant HTT aggregation in mammalian cell lines and also in primary mouse neurons, fly eye and in zebrafish. Although the levels of rescue obtained varied between compounds depending on the model used, this may be as a result of differences in absorption routes and bioavailability. Nevertheless, our data showed that pharmacologic inhibition of QPCT can rescue HD phenotypes and provides proof of principle for QPCT as a potential therapeutic target for HD and possibly other related intracellular proteinopathies by modulating the formation of oligomeric forms, which have been proposed as the most toxic species in these diseases (Lajoie and Snapp, 2010; Takahashi et al., 2008). Clearly, further work, most likely including additional drug development, is required before we can consider whether this will be clinically relevant. Nevertheless, in a broader perspective, our data suggest that a discovery pipeline from druggable genome screen to drug development may be tractable for neurodegenerative diseases.

TAMAHUD knowledgebase with both TAMAHUD experimental results as well as public high-dimensional Huntington's disease datasets has been developed as a result of this collaborative project. This is useful for many researcher working in the area of Huntington's disease and more broadly neurodegenerative area.

Contributions: This is a collaborative project, the consortium is composed of partners with complementary expertise and specific know-how. Sienabiotech SpA, University of Cambridge and TCP Innovations carried out the experimen-

tal work: High throughput RNAi (HT-RNAi) screen to establish a druggable genome for HD; validation of primary hits; validation and interrogation functional role of selected hits (targets) in fly and zebrafish models; small molecule screening etc., where as Alma Bioinformatics S.L (Bioalma) mined the relevant literature by using their text-mining machinery, Alma Knowledge Server (AKS). At EMBL Heidelberg, I have collected and integrated all the experimental data generated with in the project as well as some of the selected, publicly available HD datasets and analysed and built a comprehensive TAMAHUD knowledgebase under the supervision of Reinhard Schneider and contributed to the TAMAHUD publication.

4.6 TAMAHUD publication

This research work has been published in Nature Chemical Biology journal (Jimenez-Sanchez et al., 2015a) and the article is provided in the Publications section 9.3.

Chapter 5

Human-gpDB

5.1 Description

G-protein coupled receptors (GPCRs) are a major family of membrane receptors in eukaryotic cells. They play a crucial role in the communication of a cell with the environment. Ligands bind to GPCRs on the outside of the cell, activating them by causing a conformational change, and allowing them to bind to G-proteins. Through their interaction with G-proteins, several effector molecules are activated leading to many kinds of cellular and physiological responses. The great importance of GPCRs and their corresponding signal transduction pathways is indicated by the fact that they take part in many diverse disease processes and that a large part of efforts towards drug development today is focused on them.

Human-gpDB is a publicly accessible, relational database which currently holds information about 713 human GPCRs, 36 human G-proteins and 99 human effectors. The collection of information about the interactions between these molecules was done manually and the current version of Human-gpDB holds information for about 1663 connections between GPCRs and G-proteins and 1618 connections between G-proteins and effectors. Major advantages of Human-gpDB are the integration of several external data sources and the support of advanced visualization techniques. Human-gpDB is a simple, yet a powerful tool for researchers in the life sciences field as it integrates an up-to-date, carefully curated collection of human GPCRs, G-proteins, effectors and their interactions. The

database may be a reference guide for medical and pharmaceutical research, especially in the areas of understanding human diseases and drug discovery.

The initial step was to collect the sequence information for each GPCR, G-protein and effector from UniProt/SwissProt database. This information was obtained by parsing respective database entries for the description (DE), the gene (GN) and the database cross references (DR). The next step was to isolate and keep those proteins that have at least one connection with other proteins, e.g., a GPCR with a G-protein and a G-protein with an effector and vice versa. In order to achieve this, an extensive manual literature search was performed to detect proteins that co-occur in the same abstract and are biologically relevant. The corresponding PubMed IDs were also recorded along with the co-occurrence information. This manual literature search was done my Margarita C. Theodoropoulou, University of Athens. I have build the rest of the knowledgebase including backend database design and implementation, development of ETLs (Extract, Transform and Load) scripts and the web application with rich JavaScript and exploration tools. All the three types of proteins (GPCRs, G-proteins and effectors) were categorized into families, subfamilies and types. G-proteins were classified based on their α-subunits sequence homology, whereas effectors classification was based on their function. For GPCR many different classifications exist, we have used IUPHAR classification (Harmar et al., 2009).

UniProt identifiers were used as starting points to integrate the Human-gpDB with various external data sources. The systems that were used to help us with this integration were similar to the ones mentioned in the bioCompendium data layer, i.e. ENSEMBL, BioMart and SRS. For each protein, information about the name, the sequence, the description, the family and the subfamily it belongs to, together with the full record coming from the Dasty2 DAS client (Jimenez et al., 2008) was collected.

In oder to develop this resource I have applied data integration and knowledge management approaches to integrate GPCRs, G-proteins, effectors and their interactions. These proteins and their interactions are annotated with rich knowledge from publicly available biological databases. The overview of Humap-gpDB web application is shown in Figure 5.1 The construction of this service, its content, data integration, visualization technologies used, its implementation details

and utility of the service were well described in my publication of this resource, Human-gpDB publication (Satagopam et al., 2010) and it is available in the Publications section 9.3.

Currently, two Human-gpDB servers are set up, one running at LCSB/EMBL (http://schneider.embl.de/human_gpdb) and the other running at the Department of Cell Biology and Biophysics of the University of Athens (http://bioinformatics.biol.uoa.gr/human_gpd/). Both servers hold the same copy of the Human-gpDB database. Concerning the linking to Human-gpDB from external sources, other databases can link to our database by using, for example, the following URLs: http://schneider.embl.de/cgi-bin/human_gpdb.cgi?search=P21918 or http://schneider.embl.de/cgi-bin/human_gpdb.cgi?search=DRD5_HUMAN, based on Uniprot ID or Accession Number.

5.2 Summary of results

The database currently holds information about 713 human GPCRs, 36 human G-proteins and 99 human effectors. The interactions between these molecules were collected manually and the current status of Human-gpDB reveals information about 1663 interactions between GPCRs and G-proteins and 1618 interactions between G-proteins and effectors. GPCRs are categorized in four classes. Table 5.1 shows the number of families and subfamilies in each GPCR class, while Table 5.2 shows the distribution of GPCRs' subfamilies based on the number of Gα families with which they interact. G-proteins are categorized in Gα, Gβ and Gγ groups. Gα consists of four respective families. From the 36 human G-proteins, 17 are characterized as Gα, 7 as Gβ and 12 as Gγ. Effectors are categorized in 20 families, 29 subfamilies and 63 types based on their biological function. The two most highly populated effectors families are Ion Channels and Tubulins.

Visualization of the interactions between GPCRs, G-proteins, effectors and drugs together with the rich data integration part is one of the main features of Human-gpDB. Medusa (Hooper and Bork, 2005) application was used for 2D representation of the networks of interactions. Arena3D (Pavlopoulos et al., 2008) was chosen for 3D and more efficient representation of either the whole network of

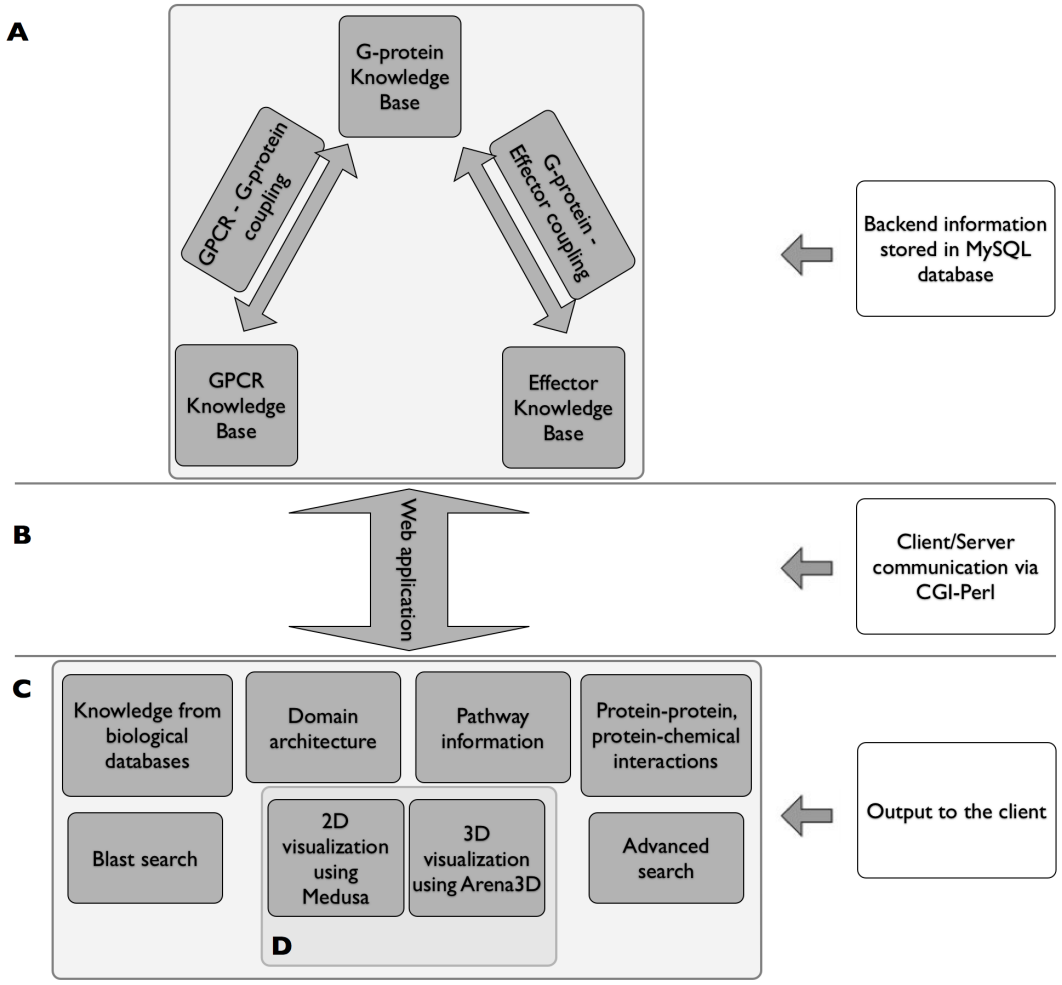


Figure 5.1: **Overview of Human-gpDB web application**. (a) Backend of the application consisting of manually collected information regarding GPCRs, G-proteins, effectors and their interactions as well as a wide range of publicly available information for each of these proteins stored in a MySQL database. (b) A CGI-Perl script handles the communication between the client and the server and (c) provides a wide range of information as output to the user. The system can be coupled with (d) a 2D visualization tool (Medusa) and a 3D visualization tool (Arena3D), which allows the easy visualization of the relationships between GPCRs, G-Proteins, effectors and the drugs. ©This Figure is from my own publication and is used with permission from the *Oxford University Press* (Satagopam et al., 2010).

GPCRs' Class	#GPCR Families	#GPCR Subfamilies
Class A	55	640 (422 subfamilies of olfactory receptors)
Class B	6	16
Class C	4	41 (29 subfamilies of taste receptors)
Frizzled/Smoothened	2	11
Total	67	708

Table 5.1: **Number of families and subfamilies in each GPCR class** - Class A is the largest and consists of 55 families and 640 subfamilies (422 subfamilies of olfactory receptors). Class B consists of 6 families and 16 subfamilies. Class C consists of 4 families and 41 subfamilies (29 subfamilies of taste receptors). Frizzled/Smoothened class consists of 2 families and 11 subfamilies.

Couples with	#GPCR Subfamilies
1 Gα family	623
2 Gα families	48
3 Gα families	15
All 4 Gα families	1
Unknown coupling	21
Total	708

Table 5.2: **Distribution of GPCR subfamilies based on the number of Gα families with which they interact** - One subfamily of GPCRs, the TSHR family, couples with members of all four Gα families. Most of the GPCR subfamilies couple with members of one Gα family (623 out of the 708 subfamilies of GPCRs). Fifteen GPCR subfamilies couple with members from 3 Gα families, whereas 48 couple with members from 2 Gα families. Twenty-one GPCR subfamilies do not have known coupling.

interactions or dense subparts of it. Medusa, which is a Java applet, offers the user a first glance of the respective network. However Medusa still has disadvantages compared to Arena3D mainly due to the fact that the visualization it offers is in 2D so the space might be a limiting factor for larger or dense networks.

5.3 Discussion

Human-gpDB compared to the previous gpDB databases (Elefsinioti et al., 2004; Theodoropoulou et al., 2008) is now richer and focuses only on human GPCRs, G-proteins and effectors. Human-gpDB is not simply a gpDB subset, since it contains more recent data, and it also contains new information concerning the clas-

sification of GPCRs (11 new subfamilies were added and all existing subfamilies are classified based on the IUPHAR classification) and also contains interactions between all molecules. It is fully integrated with external data sources by bridging information that did not exist in the previous versions (e.g. drugs and chemicals) and it now comes with a new user-friendly environment supported by advanced visualization technologies. The interface makes the navigation friendlier, the exploration of information more efficient and the extraction of new knowledge easier. Human-gpDB database was built to provide a simple but yet a powerful tool for researchers in the life sciences field as it integrates a current, careful collection of human GPCRs, G-proteins, effectors and their interactions. Human-gpDB uses advanced visualization technologies to make the volume of data more informative and the advanced data integration techniques make Human-gpDB a unique tool, a reference guide in pharmaceutical research and especially in the areas of chemical and drug discovery for human diseases. In the future, the expansion of the current version of the database for other organisms starting from the ones that are evolutionarily closer to Humans is essential.

Contributions: The development of Human-gpDB was conceived by Reinhard Schneider and Stavros J. Hamodrakas from the Department of Cell Biology and Biophysics of the University of Athens. The manual literature search was done my Margarita C. Theodoropoulou and I have build the rest of the knowledgebase including backend database design, data integratoin and implementation, development of ETLs and the web application with rich JavaScript and exploration tools. Georgios A. Pavlopoulos worked on visualization tools, Nikolaos C. Papandreou, Christos K. Stampolakis, Pantelis G. Bagos. helped in the development of the application. Reinhard Schneider and Stavros J. Hamodrakas supervised the project. All are involved in the writing of the Human-gpDB publication.

5.4 Human-gpDB publication

The work has been published in Database journal (Satagopam et al., 2010) and the article is provided in the Publications section 9.3.

Chapter 6

SBML Map Annotation Service

6.1 Description

In systems biology scientists often represent the biological pathways information in computer readable SBML (Systems Biology Markup Language) format, which is an XML document. These SBML maps can benefit from comprehensive bioinformatics annotation that may eventually enrich the map with the new elements and links.

The Map Annotation Service has been developed in the context of MINERVA platform (Gawron et al., 2016) to annotate the hosted Parkinson's Disease (PD) map (Fujita et al., 2014) elements - genes, proteins, chemicals, drugs, metabolites.

6.1.1 MINERVA platform

As described in (Gawron et al., 2016), MINERVA (Molecular Interaction NEtwoRks VisuAlization) platform, is a standalone webservice supporting curation, annotation and visualization of molecular interaction networks in Systems Biology Graphical Notation (SBGN)-compliant format. MINERVA provides automated content annotation and verification for improved quality control. The end users can explore and interact with hosted networks, and provide direct feedback to content curators. MINERVA enables mapping drug targets or overlaying experimental data on the visualized networks. Extensive export functions enable downloading areas of the visualized networks as SBGN-compliant models

for efficient reuse of hosted networks. The software is available under Affero GPL 3.0 as a Virtual Machine snapshot, Debian package and Docker instance at http://r3lab.uni.lu/web/minerva-website. MINERVA is an important contribution to systems biology community, as its architecture enables set-up of locally or globally accessible SBGN-oriented repositories of molecular interaction networks. Its functionalities allow overlay of multiple information layers, facilitating exploration of content and interpretation of data. Moreover, annotation and verification workflows of MINERVA improve the efficiency of curation of networks, allowing life-science researchers to better engage in development and use of biomedical knowledge repositories (Gawron et al., 2016).

6.1.2 Parkinson's disease map

Parkinson's disease (PD) is a neurodegenerative disease involving a complex interplay of environmental and genetic factors. It becomes increasingly important to develop new approaches to organize and explore the exploding knowledge of this field. The PD map is a computerbased knowledge repository, representing diagrammatically molecular mechanisms of PD in a structured and standardized way. It can be linked to bioinformatics tools facilitating exploration and updating the contents of the map using bioinformatic annotations (Fujita et al., 2014).

The PD map (http://pdmap.uni.lu) focuses on pathways involved in neurodegenerative processes of the neuronal system, in particular on the degeneration of dopaminergic neurons of substantia nigra pars compacta. Its manually curated interactions are supported by over a thousand publications, resulting in roughly 5,000 elements linked by over 2,000 interactions. The PD map utilizes all MINERVA functionalities to explore a large repository of molecular mechanisms, and enables users to interpret their experimental data and guide content curation via the commenting system (Gawron et al., 2016).

6.1.3 Systems Biology Markup Language

The Systems Biology Markup Language (SBML) is a standard for representing models of biochemical and gene-regulatory networks. It is a computer-readable XML format for representing models of biological processes. SBML is suitable for,

but not limited to, models using a process description approach. SBML is defined in a set of specification documents that describe the elements of the language, its syntax, and provide validation rules (Hucka et al., 2003). Biological processes, gene-regulatory and biochemical network models are built using CellDesigner (`(http://www.celldesigner.org`), a structured diagram editor. Networks are drawn based on the process diagram with Systems Biology Graphic Notation (SBGN) and are stored using SBML (Hucka et al., 2003; Le Novere et al., 2009).

6.2 Implementation

Publicly available biological databases consist of valuable information. The current 2017 Nucleic Acids Research (NAR) database issue, the annual collection of bioinformatic databases on various areas of molecular biology, consists of 1877 biological databases (Galperin et al., 2017). I took advantage of the biological knowledge deposited in these publicly available databases e.g. UniProtKB (Magrane and Consortium, 2011), HGNC (Seal et al., 2011), EMBL (Cochrane et al., 2009) , Ensembl (Flicek et al., 2012), UCSC genome browser database (Fujita et al., 2011), RefSeq (Pruitt et al., 2012), EntrezGene (Maglott et al., 2011), UniGene (Schuler, 1997), GeneCards (Safran et al., 2010), PDB (Berman et al., 2007), Gene Ontology(GO) (Ashburner et al., 2000), KEGG (Kanehisa et al., 2012), PANTHER (Mi et al., 2005), Reactome (Croft et al., 2011), Pfam (Punta et al., 2012), InterPro (Hunter et al., 2012b), PharmGKB (McDonagh et al., 2011) etc., and implemented an annotation service that will enrich the PD map. Each species in the PD map is annotated with information like HUGO official gene symbol, gene description, synonyms, chromosomal location, GO terms, disease associations apart from PD, link outs to several biological databases mentioned above and is available via notes section of the PD map XML scheme. An example of such annotation for a gene SNCA (alpha-synuclein) is shown in Appendix A section 9.3.

All these important databases are parsed, integrated and indexed into Sequence Retrieval System (SRS) (Etzold and Verde, 1997) and this is described in Chapter 2, section 2.2.1. An ETL (Extract, Transform, Load) Perl script has been developed with SRS getz queries & queried these databases systematically and

a comprehensive knowledgebase of bio-entities (e.g., genes, proteins, chemicals, drugs, metabolites), their synonyms, annotations and cross references to other databases for seven model organisms - yeast, worm, fly, zebrafish, rat, mouse and human- have been collected and is named 'SynonymDB'.

In order to process the SMBL files and their contents, especially bio-entities mentioned in the SMBL files (called 'species' in CellDesigner terminology) in realtime, the SynonymDB is stored into the main memory, which accelerates the information retrieval. A TCP/IP (Transmission Control Protocol/Internet Protocol) based Remote Procedure Call (RPC) is implemented to access the bio-entity annotation for the given input of e.g., genes/proteins from the SynonymDB. Two dedicated web-services have been developed using Perl-CGI, HTML and Apache HTTP server and both of them are available under bioCompendium (Chapter 2) as sub-services. The overview of the map annotation service is shown in Figure 6.1.

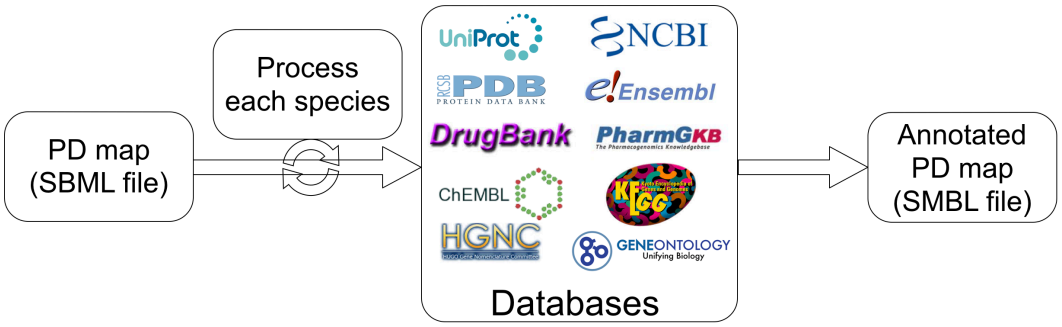


Figure 6.1: **Map annotation service workflow**. The figure outlines the overview of the map annotation service.

6.3 Results

In this project publicly available biological databases have been mined and built into a comprehensive resource, SynonymDB with bio-entities, their synonyms, annotations and cross-references and is stored in the main memory. We have developed two web-applications as sub-services of bioCompendium - (1) SynonymDB

web application, which takes list of bio-entities or their synonyms as input and returns the annotated bio-entities and is available at http://biocompendium.embl.de/synonym_db and its landing page and an example is shown in Figures 6.2, 6.3; (2) Map annotation service which is available at URL http://biocompendium.embl.de/map_annotator. This web application takes SMBL file produced by CellDesigner editor, especially in the case of PD map as input. It parses this XML file, and extracts contents of *<species>..</species>* elements and these elements have an attribute *name*, and its value is the name of the bio-entity. And species element also has a *<notes>..</notes>* child element that is the place holder to incorporate the species annotation information. The species element's name attributes (bio-entities) are collected and the structure information is send to SynonymDB via RPC call. The socket programming method receives the information and returns the annotations for each bio-entity as a structured information with header information. The web server processes the received annotation information, and incorporates the annotations into *<notes>..</notes>* element and provides annotated version of SMBL file for download as shown in the Figure 6.1. For programatic access, RESTful based Application Programming Interface (API) has been developed and it takes a list of bio-entities as input and returns the corresponding enriched annotations as output. As this resource consists of synonyms, it is also serving to harmonise the bio-entities to the standard database identifiers. This part of the API call takes a list of synonyms as input and returns the primary identifiers/accession numbers of the database of user choice.

6.3.1 Applications

Parkinson's disease map annotation: In our PD map project (it is not directly part of this thesis work), annotation, enrichment and analysis of the dysregulated pathways implicated in PD are strongly coupled, and their interconnections need to be represented in an integrated and comprehensive way to be studied efficiently. PD map allows navigation through information on PD associated mechanisms, and constitutes an interface to well established tools and methods for updating, enriching, and analysing its contents (Fujita et al., 2014).

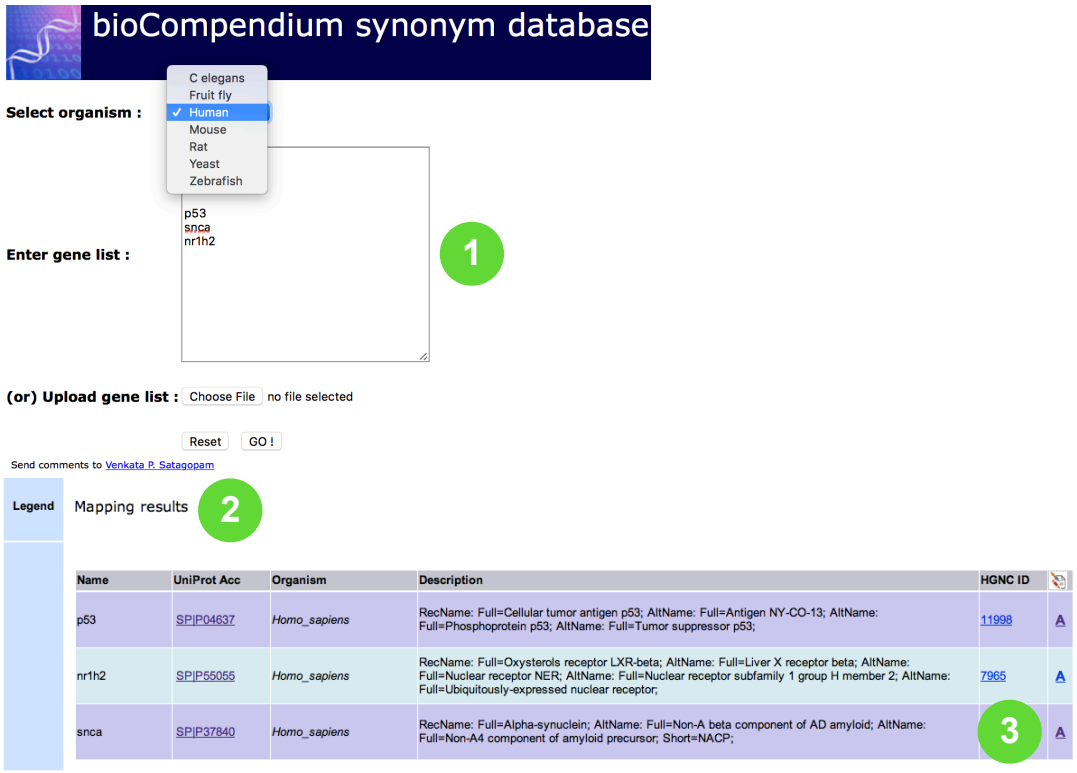


Figure 6.2: **Synonym database web application**. The figure shows (1) the front page of the SynonymDB web application. It consists of selection of organism and a place holder for provide list of bio-entities or upload the file; (2) table view of the mapping results and upon selection of 'A' (3) stands for Annotation, will open the detailed annotation of the bio-entity in a separate page shown in 6.3.

We have enriched the elements of the PD map using the SynonymDB, where the SMBL file of PD map is uploaded to map annotation service and the annotated map is downloaded. Information on official gene symbol, synonyms, description and chromosomal location; association with biological processes and diseases; or molecular interacting partners have been embedded within the map. Annotation of the contents of the PD map facilitates the knowledge exploration by providing additional information about map elements and their interactions, and is easily accessible online. Figure 6.4 shows the PD map in CellDesigner before and after the map annotation.

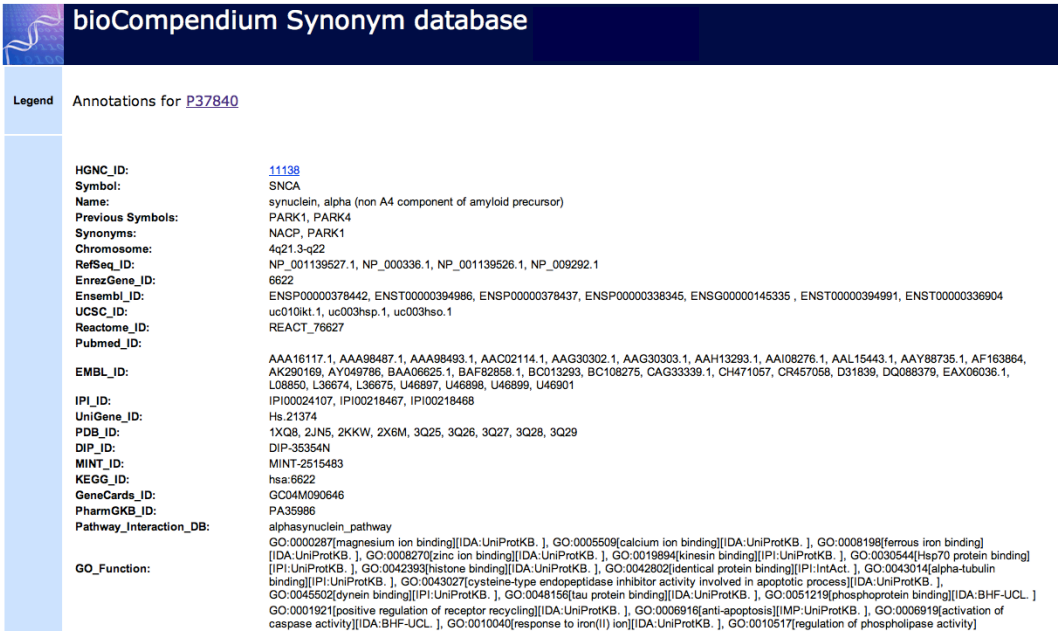


Figure 6.3: **Synonym database web application annotation result interface**. The figure shows annotation for a selected gene of choice for example 'SNCA'.

Map annotation service and MINERVA platform: MINERVA, is the web-based platform hosting molecular interaction networks encoded as process diagrams in SBGN and CellDesigner formats. The platform features configurable automated annotation tools (Gawron et al., 2016). These tools facilitate the automatic annotation of species elements using map annotation service and synonym database RESTful API. The MINERVA platform allows direct exploration of the content, including lookup of multiple terms, hyperlinks to available annotations as shown in the left panel of the Figure 6.5 and visualization of high-throughput datasets on the right panel of the Figure 6.5.

SynonymDB as CellDesigner plugin and Garuda gadget: SynonymDB web service has been integrated into CellDesigner as plugin and it facilitates the

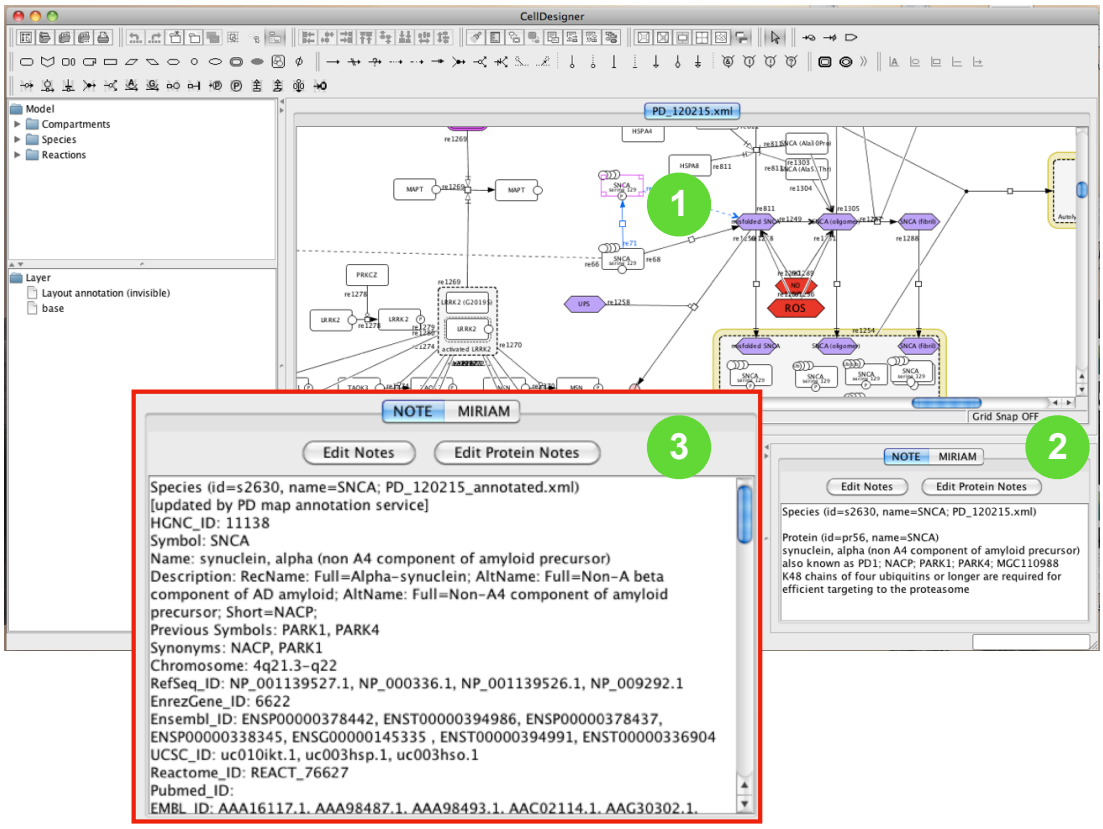


Figure 6.4: **Parkinson's disease map in CellDesigner**. The figure shows (1) PD map in CellDesigner interface and selection of a species 'SNCA'; section (2) is the 'Notes' section of the CellDesigner and is the place holder for annotations. The annotations in section (2) are from SMBL file before application of map annotation service, where as in section (3) after application of the annotation service.

standardisation of the bio-entity (species) names during the development of the contextualized disease maps as shown in Figure 6.6. In order to achieve the similar functionality but from Garuda platform (ref Chapter 2 section 2.4.2 for more details on Garuda), a Garuda gadget has been developed using SynonymDB RESTful API 6.3.1 and integrated into Garuda platform as shown in Figure 6.6. MIRIAM (the Minimal Information Requested In the Annotation of Models) is a standard to annotate and curate computational models in biology `http://www.ebi.ac.uk/miriam`. A MIRIAM widget has been developed using SynonymDB RESTful API and is depicted in the Figure 6.6 block 4.

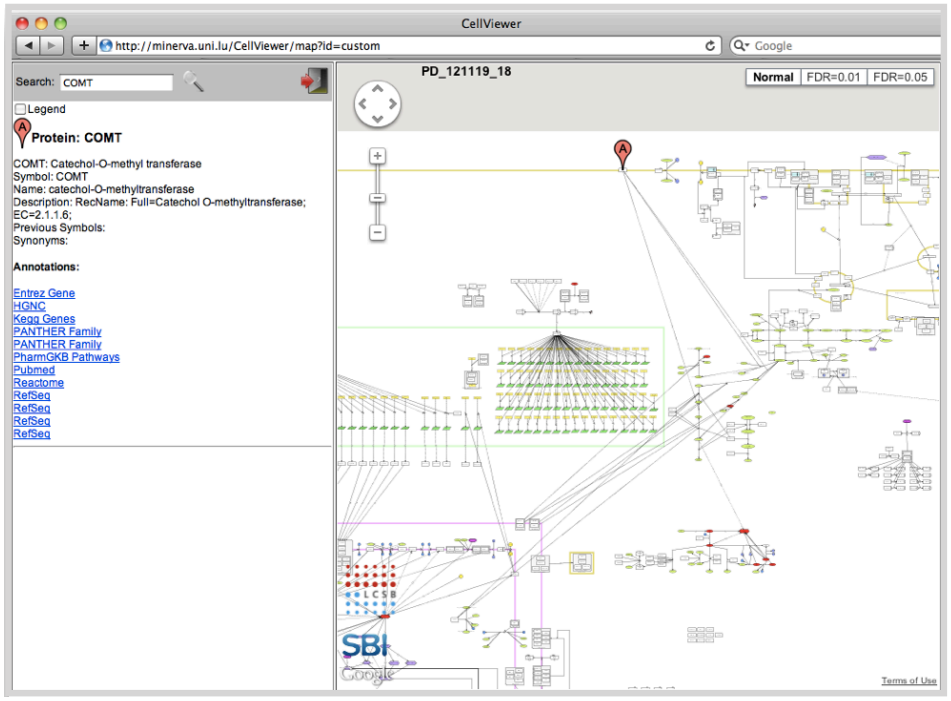


Figure 6.5: **Map annotation service and MINERVA platform**. The figure shows PD map in MINERVA interface. Here the right part of the diagram shows the biochemical interactions involved in PD disease pathology. It consists of various elements including genes, proteins, chemicals, drugs, metabolites. Upon the selection of these elements, the annotations of the selected bio-entity are shown in left panel of the interface and these annotations are retrieved using map annotation service which internally uses API call to Synonym Database.

SynonymDB RESTful API: The SynonymDB is the main service behind the map annotation service and provides RESTful web services based API apart from the web interface. By using this API users can programmatically access its functionality and it is very practical for programmers and tool developers. RESTful web services can be accessed via the following URL using HTTP 'post' method.

```
[http://biocompendium.embl.de]/[synonym_db]/[REST]/[method]/arguments
Method : GetMapping
Organism : human
```

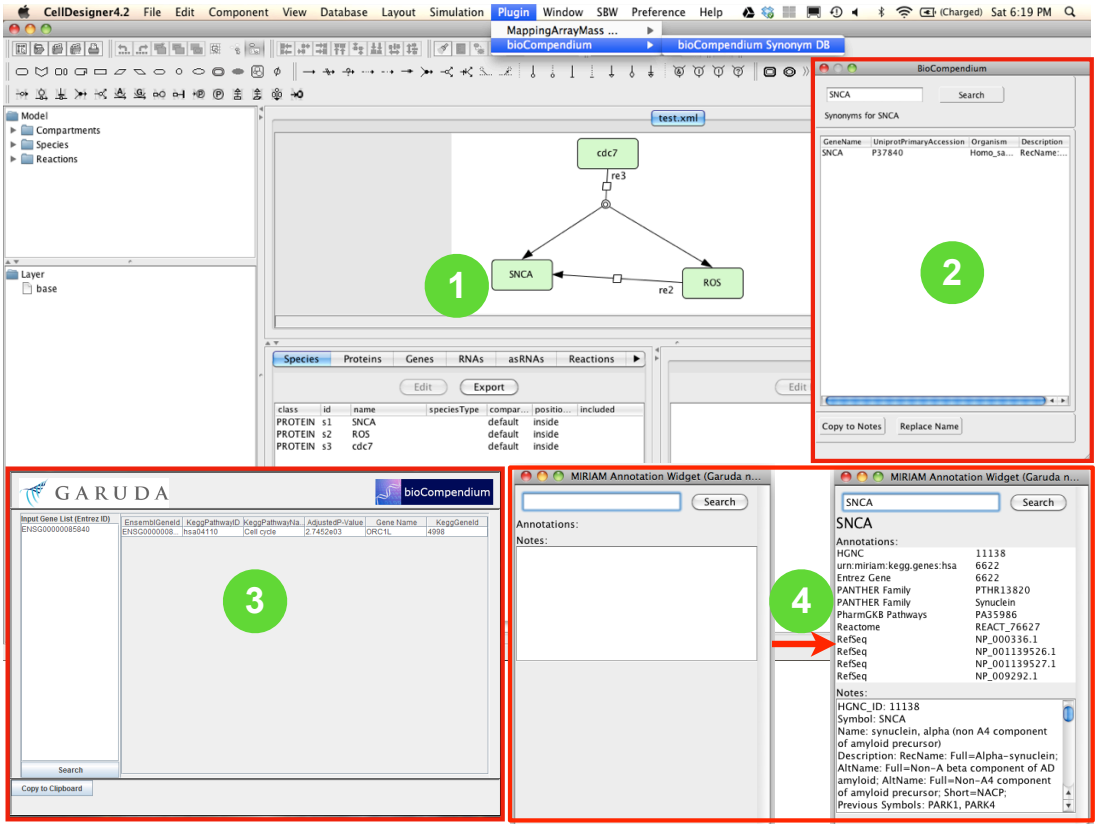


Figure 6.6: **SynonymDB as CellDesigner plugin, GARUDA gadget and MIRIAM widget**. The figure shows (1) CellDesigner interface with example network and integration of SynonymDB as a CellDesigner plugin (2), as GARUDA gadget (3) as well as MIRIAM widget (4) using SynonymDB RESful API and serving all these systems biology tools with the information from various publicly available biological databaes.

```
GeneList :  p53 nr1h2
Format : tsv, csv, xml, json
http://biocompendium.embl.de/synonym_db/REST/GetMapping/Organism=human&
GeneList=p53\%20nr1h2&Format=xml
```

6.4 Discussion

We present the synonym database service, a comprehensive knowledgebase obtained from mining of various publicly available biological databases. It consists of several bio-entities (e.g., genes, proteins, drugs, chemicals, metabolites), their

synonyms, annotations and cross-references to other resources. Another service, SBML map annotation service, enrich the different molecular interaction networks and disease maps e.g., Parkinson's disease map by providing rich annotations to various species elements of the map. Both these services are available as web applications with RESTful APIs. Systems biology tools including - CellDesigner, GARUDA, MINERVA platform - took the advantage of these biological data driven services to develop plugins (gadgets) using provided RESTful APIs to standardise & harmonise bio-entities and/or annotate them. In the future we would like to add more biological databases to this knowledgebase to cover more bio-entities.

Contributions: The development of Map Annotation Service was conceived by myself and Reinhard Schneider. Building of the database, web application and API was carried out by me.

Chapter 7

HIV Mutation Browser

7.1 Introduction

Human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS), infects millions of people worldwide and, to date, has been responsible for over 25 million deaths (Kallings, 2008). More than 35 millions of people of currently living with AIDS (`http://aidsinfo.unaids.org`). The clinical importance of the virus has prompted substantial funding of HIV/AIDS research across many diverse clinical, therapeutic (drug design, vaccine production) and basic research fields. This research has produced an extensive catalogue of HIV literature. Consequently, finding literature pertinent to a particular topic became a difficult task. Researchers are often interested in the phenotypic variation resulting from naturally occurring single nucleotide polymorphism or directed mutagenesis in the HIV genome. Traditionally, mutation data for a particular protein or region must be manually collected by trawling literature repositories such as PubMed using author names, protein/gene names, keywords or a mixture of all three. The scale of the HIV literature (over 275,000 articles) makes such an approach inadequate.

Several valuable online resources have provided mutation data to researchers by manually curating polymorphism and mutagenesis data from HIV studies. These include the Stanford Drug Resistance database (Rhee et al., 2003), which contains curated mutations related to drug resistance, the UniProt knowledge-

base (UniProt-Consortium, 2014), which manually annotates articles describing mutagenesis of HIV proteins and the Los Alamos HIV Database, which annotates various sources of HIV data including epitope variants and escape mutations (http://www.hiv.lanl.gov). However, these resources are limited in scope because manual curation cannot feasibly be carried out on all of the available literature.

Resources such as Reflect (Pafilis et al., 2009) and MutationFinder (Caporaso et al., 2007) are available to quickly scan, tag, annotate the bio-entities computationally and organise large amounts of scientific literature systematically. These techniques should be applied to facilitate the work of HIV researchers. But it is surprising that so few resources are available to access the available literature in an organised and structured way. One such facility is PubMed Central (PMC https://www.ncbi.nlm.nih.gov/pmc/), even though PubMed comprises more than 27 millions publications, only 4.4 millions (16% of PubMed) articles are archived in PMC and are available for scientific community for manual reading. Many of these articles are subject to publishers access licenses and copyright restrictions and are not available for bulk downloading. Only a fraction (1.6 millions, 6%) of these publications are available for bulk retrieval and text mining. This was a major issue in miming full text articles.

Fortunately, recent pressure from government and scientific bodies and the rise of open access publishing has softened the stance of publishers and many are now receptive to waiving these restrictions. Such advances will pave the way for many large scale literature text-mining projects and will likely change the way we access scientific literature. In our case, thankfully, the majority of the publishing companies and societies that we approached granted us permission to text-mine and index HIV mutation information contained in their literature. The list of publishers and journals that have given permission to the HIV mutation browser to access, data-mine and display articles can be found from the HIV mutation browser website (http://hivmut.org/index.php?page=about).

In the case of HIV, currently there are over 275,000 articles in PubMed. The number of articles publishing every year increasing over time and is shown in Figure 7.1. We have the permission to process approximately 45% of these full text articles, which are used for text-miming to build the HIV mutation browser.

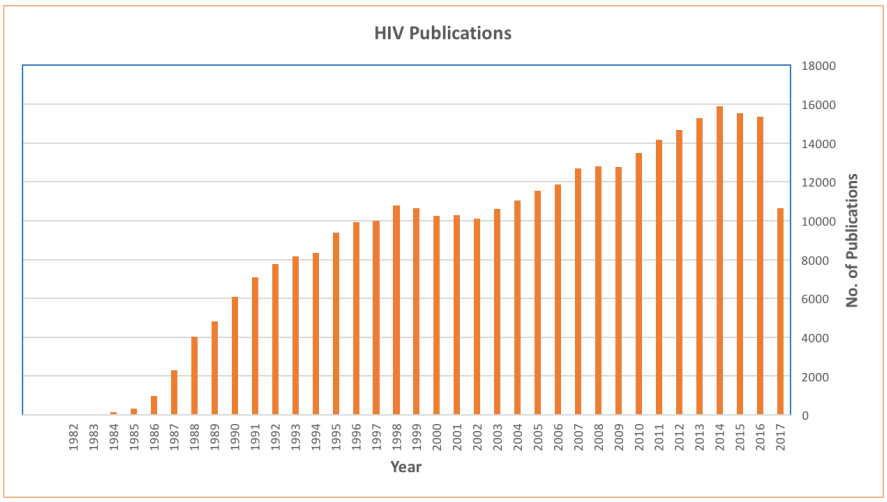


Figure 7.1: **HIV Publications**. The figure shows the number of HIV publications deposited in PubMed annually over last 3.5 decades.

The HIV mutation browser is a database of mutagenesis and mutation data on HIV collected from the scientific literature. The data has been identified and catalogued using computational text-mining methods. A researcher can use the database to find literature describing the phenotype of a mutation, and/or experimental data describing the effect of a mutation. The work reported here is a collaborative project between John Briggs, European Molecular Biology Laboratory (EMBL), Heidelberg and Reinhard Schneider, Luxembourg Centre for Systems Biomedicine (LCSB), Luxembourg. In this project, I have developed the database schema and the prototype of mutation retrieval and mapping software for the HIV mutation database. I have also contributed to the design and

construction of article retrieval software, mutation browser website and negotiating permission for copyrighted articles.

Here we have applied text-mining techniques to extract data on polymorphisms and mutations from the available HIV literature. We have organised this data in a protein and residue-centric way and have made it available through an online resource, the HIV mutation browser (`http://hivmut.org`). This publicly available resource will simplify the task of identifying relevant literature for HIV research, thereby aiding experimental design and reducing replication of efforts.

7.2 Implementation

Development of the workflows for collection of the literature, extraction of the mutations and bio-entities (proteins, HIV strain names) from the literature, standardisation of these bio-entities to gold standards and database identifiers, mapping of extracted mutations on HIV proteins, development of the HIV mutation database and web-server have been outlined in Figure 7.2 and described in detail in our own HIV mutation browser publication (Davey et al., 2014)(refer section 7.5).

7.3 Summary of results

Creation of the HIV mutation browser required a number of steps as shown in Figure 7.2. First, we obtained permission from publishers, identified, and accessed the relevant literature. Second, we established and applied text-mining techniques to retrieve data on mutagenesis and polymorphism from the HIV literature. Third, we associated the mutation data to the appropriate amino acid residues within the HIV proteome. Finally, we developed a web server, through which the data can be accessed in an intuitive and informative way.

As show in the Table 7.1, we identified 275,000 articles containing the search term 'HIV' or 'Human Immunodeficiency Virus' indexed in the PubMed database (from a total of 27 million publications). We retained 126,800 out of 275,000 relevant articles, published across 2,639 journals, representing approximately 45% of

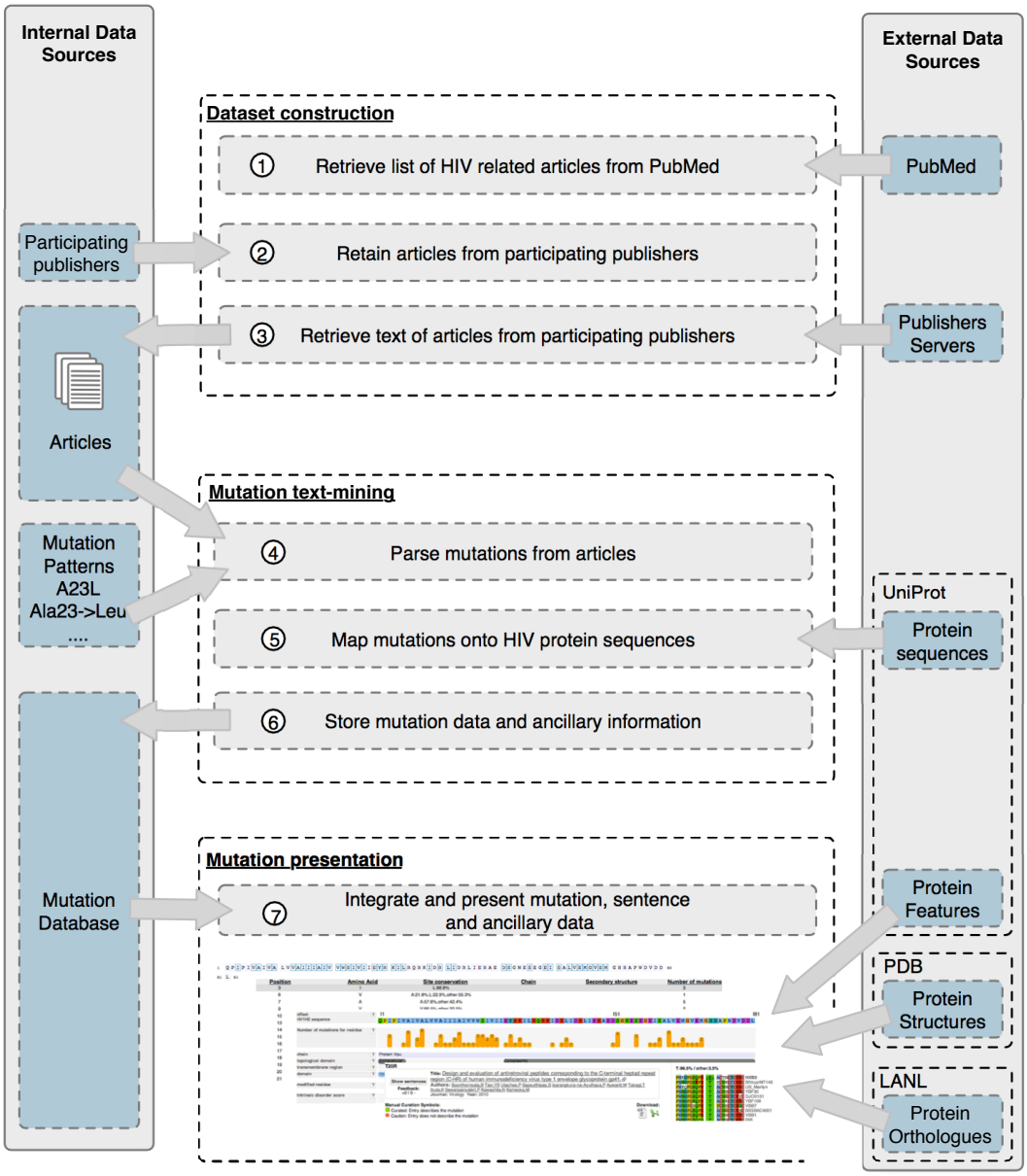


Figure 7.2: **HIV mutation browser schema**. A list of HIV related PubMed article identifiers (PMIDs) are retrieved from PubMed. The publishing journal of each article is compared against a list of participating publishers (i.e. publishers that have given permission for bulk PDF downloading, computational parsing of PDF and display of articles details). Permitted articles are retrieved from the publishers website as PDF files. Retrieved articles are computationally text-mined to parse patterns commonly used in the literature to denote mutations. Each mutation is then mapped onto the HIV proteome. Mutations are stored in a relational database and accessed through a web interface, the HIV mutation browser. The HIV mutation browser organises the data by protein and residue and integrates ancillary information relevant to the users. ©This figure is from our own HIV Mutation Browser publication (refer section 7.5) (Davey et al., 2014) and used by following terms of use from *Public Library of Science (PLOS)*.

the total. For the remaining,˜150,000 citations, permission for computational processing of articles couldn't be obtained from the publishers. The 126,800 articles from participating publishers were text-mined for mutagenesis or polymorphism information, and the mutations were mapped to particular residues within the HIV proteome. This required the development of a method to retrieve the text of these articles, scan the articles for patterns that are widely used to describe directed mutations in mutagenesis experiments or polymorphisms, and to map these mutations to the correct position in the correct protein (refer materials and methods section of the HIV Mutation Browser publication (Davey et al., 2014)). A total of 8,061 distinct mutations (a unique non-wildtype amino acid at a given residue in a given protein) were collected. As each mutation can be described in multiple articles and each article can describe multiple mutations, the 8,061 distinct mutations were defined by 52,281 unique references to 5,875 articles.

The identified mutations shed light on the nature of the HIV research effort of the last decades. On the one hand it has been broad in scope: 2,990 of the 3,118 residues in the HIV proteome have one or more associated references to a mutation in the repository. On the other hand it has been narrow in focus: the coverage is far from uniform and certain regions such as the catalytic sites of the protease and reverse transcriptase, as well as host interaction interfaces, are much more highly studied (Figure 7.3). HIV mutation database protein statistics are shown in the Table 7.2. It shows the distribution of the mutation data across the proteins of the HIV proteome. The number of publications, distinct mutations and distinct mutated positions per protein varies widely. The protein Pol, which contains the 3 enzymatic chains, a protease, an integrase and a reverse transcriptase is by far the most studied protein.

We have processed the publications from from 2,639 different journals. Table 7.3 lists the top 20 journals sorted by the number of mutations annotated.

The above analysis resulted in a database within which each reference to a mutagenesis experiment or polymorphism in a citation is indexed using three pieces of information: the protein that contains the mutation, the position in the protein which has been mutated, and the non-wildtype amino acid to which the wildtype residue has been mutated. To make this data accessible to researchers in a simple, intuitive and informative manner, we designed the HIV Mutation

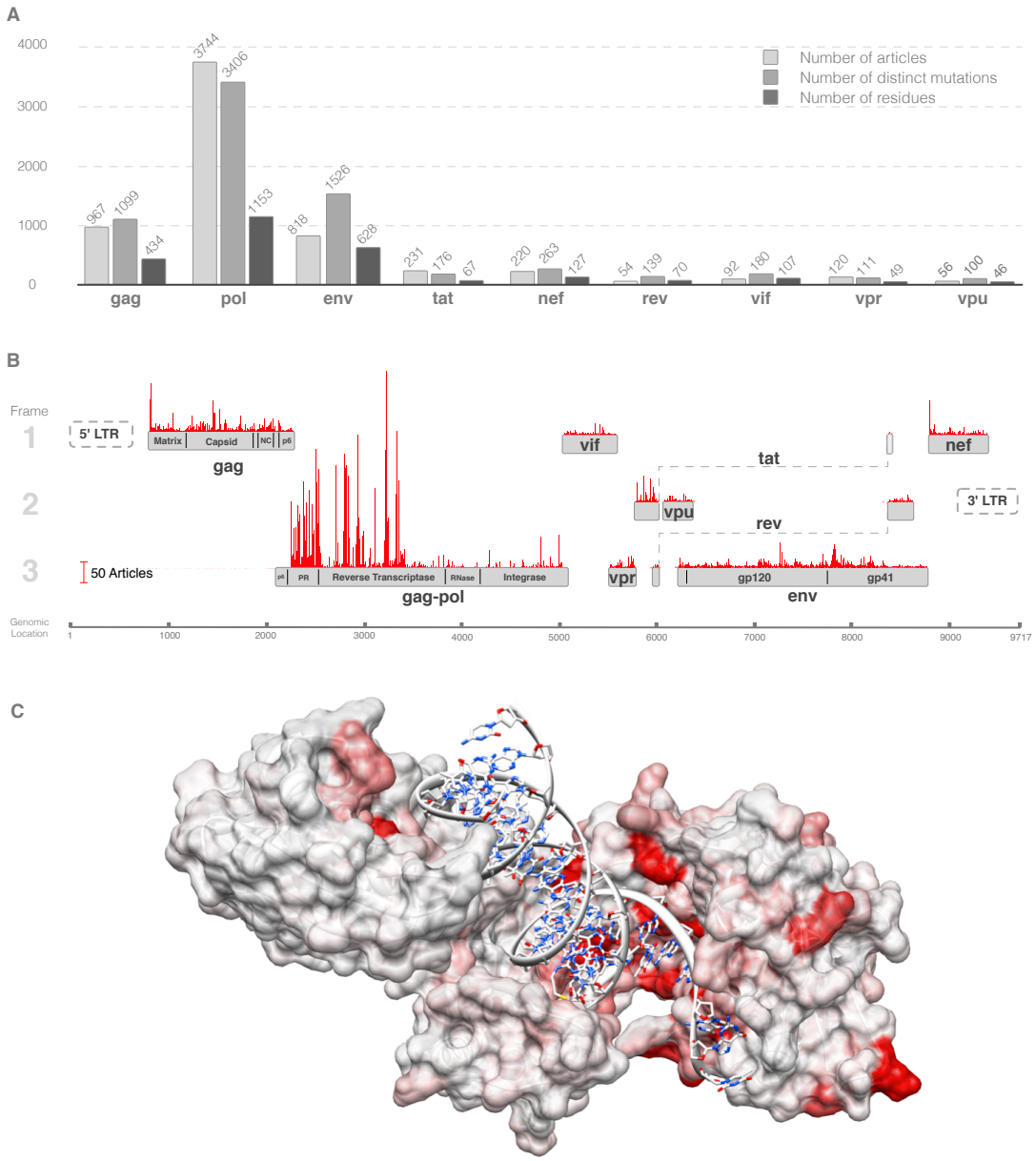


Figure 7.3: **Overview of the distribution of mutation data across the HIV proteome**. (A) Barplot of the counts of (i) the number of articles describing mutations, (ii) the number of distinct mutations and (iii) the number of residues with mutation data in the database for each protein in the HIV proteome. (B) Barplot of the counts of the number of curated articles in the database describing mutagenesis experiments or polymorphisms for each residue mapped onto the HIV proteome/genome. (C) Reverse transcriptase p66 subunit with residues coloured by number of articles referring to them. Most highly cited residues are in contact with the nucleotides or are known drug resistance mutations. White denotes no papers, full red denotes 50 or more papers, colouring is linearly scaled between 0 and 50. ©This figure is from our own HIV Mutation Browser publication (refer section 7.5) (Davey et al., 2014) and used by following terms of use from *Public Library of Science (PLOS)*.

Description	#Total
Number of relevant papers in PubMed	275,000
Number of papers permitted to processed	126,807
Number of papers processed	126,807
Number of papers containing mutational information	5,875
Number of mutations	52,281
Number of distinct mutations	8,061

Table 7.1: **HIV Mutation Database Statistics** - The table shows the statistics of the current version of the database (version 1.0). The content of the database is created by text-mining available HIV literature to find mutagenesis information. A list of articles is retrieved from PubMed using the search terms "HIV" and "Human Immunodeficiency Virus". All articles from this list that are available to us and that we are permitted to analyse computationally are downloaded and processed. We currently have permission to process approximately 40% of the literature, including the majority of basic-science publications on HIV. The database is updated on a monthly basis to add the latest HIV literature.

Protein	#Publications	#Distinct Mutations	#Distinct Positions
gag	1,154	1,503	467
pol	4,234	4,328	1,237
env	990	2,226	720
tat	250	234	75
nef	266	376	162
rev	60	169	75
vif	114	274	144
vpr	145	140	58
vpu	72	127	52

Table 7.2: **HIV Mutation Database Protein Statistics** - The table shows the distribution of the mutation data across the proteins of the HIV proteome. The number of mutations per protein varies widely. Pol which contains the 3 enzymatic chains, a protease, an integrase and a reverse transcriptase is by far the best studied protein.

Browser, a web-interface that acts as a front end for the database. The browser presents the data in a hierarchically organised manner. The user selects a gene of interest, then a position of interest, and the citations relating to this position are presented to the user grouped by non-wildtype amino acid. The web interface is organised in three panels: the navigation panel at the top; the protein panel in the middle; and the residue panel at the bottom (Figure 7.4).

The database has been populated by the results from text-mining, and it is therefore unavoidable that the database contains incorrectly assigned citations. We have therefore incorporated a user feedback system that allows users to flag the quality of an entry either positively or negatively.

To built the resource, we have used MySQL as backend to store the literature, mutation information & their annotations and HTML, PHP, JavaScript have

Figure 7.4: **HIV Mutation Browser interface**. HIV Mutation Browser interface for Vpu residue 51 showing the navigation, protein and residue panels. (1) Options bar for the residue view section of the interface. (2) Mutation information. (3) User feedback buttons. (4) Mutation information download links. (5) Ancillary residue information panel. ©This figure is from our own HIV Mutation Browser publication (refer section 7.5) (Davey et al., 2014) and used by following terms of use from *Public Library of Science (PLOS)*.

been used to develop HIV mutation browser interface. The server is available at `http://hivmut.org`. We are updating the database on a monthly basis to add latest literature automatically using cron jobs and APIs that we have developed for automatic updates. The available mutagenesis and polymorphism data for a residue can be downloaded in both tab delimited text and Excel formats directly from the web interface.

7.4 Discussion and future development

HIV is an important therapeutic target and has been the subject of a major research effort as evidenced by the large catalogue of HIV experimental literature. Appropriate organisation and categorisation of the available HIV literature is necessary to allow efficient and intuitive access to relevant data. In this

Journal	#Mutations	#Papers with mutations	#Papers
Journal of virology	6,533	1,625	11,472
Antimicrobial agents and chemotherapy	2,415	325	1,915
The Journal of biological chemistry	2,045	434	1,882
PloS one	1,917	418	6,165
Virology	1,633	277	2,495
Antiviral research	1,541	121	940
Retrovirology	1,387	142	636
Proceedings of the National Academy of Sciences	1,172	229	3,630
Journal of molecular biology	1,083	160	1,372
Journal of clinical microbiology	957	82	1,919
Journal of clinical virology	880	92	1,085
PLoS pathogens	806	147	895
Viruses	800	29	113
The Journal of antimicrobial chemotherapy	758	75	143
Nucleic acids research	595	99	1,508
Virus research	415	52	716
The Journal of general virology	407	64	597
Journal of virological methods	386	51	823
AIDS research and human retroviruses	373	19	94
AIDS (London, England)	373	21	199

Table 7.3: **HIV Mutation Database Journal Statistics** - Data from 2,639 different journals is curated in the database. The table shows top 20 journals in the HIV mutation browser by number of mutations annotated.

project, we have developed the HIV Mutation Browser, a residue-centric resource of HIV mutagenesis and polymorphism literature designed for use by those carrying out basic and applied HIV research. The HIV Mutation Browser is one of the first resources to computationally text-mine mutagenesis and polymorphism data (Doughty et al., 2011; Krallinger et al., 2009; Laurila et al., 2010), and the first to apply such methods to the extensive corpus of HIV literature. As such the HIV Mutation Browser will complement the available manually annotated and curated HIV resources such as the Stanford Drug Resistance database (Rhee et al., 2003), the UniProt knowledgebase (UniProt-Consortium, 2014) the Los Alamos HIV Database (`http://www.hiv.lanl.gov`). In the coming years, we expect this method or similar methods to be applied to other viral or cellular systems.

The resource will continue to evolve in the following ways. Firstly, HIV literature is produced continuously at a rate of approximately 1,500 articles a month and consequently the HIV Mutation Browser resource will be updated on a monthly basis. Secondly, while the resource does contain the majority of important HIV and general interest journals, it is still incomplete, as we did not

receive permission from all publishers to text-mine their HIV related articles. Journals from additional publishers will be added when possible. Thirdly, not all mutations can be correctly identified and assigned by the text-mining methods. There are various reasons for this. Many mutations are annotated in an article using non-standard patterns that are not widely used to describe directed mutations in mutagenesis experiments or polymorphisms. For example, consider the following excerpt taken from an article by Mitchell et al., "The phenotype of the combination mutant VpuD51A-S52/56N was indistinguishable from that of Vpu-D51A and Vpu-S52/56N" (Mitchell et al., 2009). The pattern "S52/56N" is a non-canonical construct for describing a mutagenesis experiment and currently will not be discovered by the text-mining method. Furthermore, the position of a mutation in a paper can be ambiguous and as a result mapping of the mutation information to the correct residue and protein can be a difficult task. We will continue to improve the methods for text-mining and assignment. We encourage the community to utilise the feedback system for misannotated mutations in the resource and contact us about mutation data that should be in the resource that is not present yet. This community input will improve the quality of the annotated data and will pinpoint parts of the text-mining method that require improvement (Davey et al., 2014).

In summary, the HIV Mutation Browser is a valuable addition to the currently available HIV resources that will allow researchers to quickly and intuitively access data on mutagenesis and phenotypic variation. We expect the database to aid the process of experimental design and be a key resource for the HIV community (Davey et al., 2014).

As the pipelines and software modules developed to build the HIV Mutation Browser generic enough to apply for any other model organisms. We are in the process of building similar resource for Hepatitis C virus (HCV). Recently scientist from the in the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) contacted us to build the similar resource for Ebola virus and at this moment we are collecting the literature related to Ebola virus.

Contributions: This is a collaborative project between Reinhard Schneider (RS) Group at the Luxembourg Centre for Systems Biomedicine (LCSB) in Lux-

embourg and John Briggs (JB) Group at European Molecular Biology Laboratory (EMBL) in Heidelberg. The development of the HIV mutation browser was conceived by RS and JB, experimental work carried out by myself, Salvodor Santiago-Mozos (SSM), and Norman Davey (ND). Data analysis, designing and construction of the HIV mutation database were done by myself and SSM. RS, myself and SSM conceived and constructed the mutation retrieval and mapping software, where as article retrieval software designed and constructed by myself, SSM and ND. RS, JB, myself, ND negotiated permission for copyrighted articles. Myself, SSM, ND equally contributed to this work.

7.5 HIV Mutation Browser publication

This research work has been published in PLOS Computational Biology journal (Davey et al., 2014) and the article is provided in the Publications section 9.3.

Chapter 8

Clinical and Translational Medicine Data Integration and Visualization

8.1 Introduction

The healthcare domain is going through a revolution that will transform the practice of medicine in virtually every way. Personalized medicine, which is predictive, preventive, personalized and participatory (also called P4 medicine) aims at establishing links between biomolecular characterizations, patient conditions, treatment effectiveness and adverse effects, and thus providing patients with the best individual treatment (Hood and Flores, 2012).

The European Society for Translational Medicine (EUSTM), defines Translational Medicine(TM) as "an interdisciplinary branch of the biomedical field supported by three main pillars: benchside, bedside and community. The goal of TM is to combine disciplines, resources, expertise, and techniques within these pillars to promote enhancements in prevention, diagnosis, and therapies" (Cohrs et al., 2015).

Translational medicine is a domain turning results of basic life science research and investigations in humans into new tools and methods in a clinical environment, for example, as new or improved diagnostics or therapies. It also includes,

non-human or non-clinical studies conducted with the aim to advance therapies to the clinic or to develop basis for application of therapeutics to human diseases. Nowadays, the process of translation is supported by large amounts of heterogeneous data ranging from medical data to a whole range of -omics data. It is not only a great opportunity but also a great challenge, as translational medicine big data is difficult to integrate and analyze, and requires the involvement of biomedical experts for the data processing. The rise of translational and personal medicine have been made possible due to the advancement in many high-throughput technologies to study the cellular processes and molecular functions of organisms at different levels - cellular, tissue, organ and system as a whole. These technologies such as genome, transcriptome, proteome, lipidome, metabolome, epigenome, and microbiome collectively called -omics of both single as well multi-cell samples, their integration and systems biology, have greatly advanced our understanding of human health and diseases (Canuel et al., 2015; Hawkins et al., 2010). However, the progress comes at a cost - translational research data sets nowadays include genomic, imaging, and clinical data sources (Bender, 2015; Topol, 2015), making them large and heterogeneous. In effect, important steps of the data life cycle in discovery - collection, integration, analysis, and interpretation - are a challenge for biomedical research. Moreover, enabling biomedical experts to efficiently use big data processing pipelines is another challenge.

As translational medicine data become more and more rich and complex, their potential in informing both clinical and basic research grows (Regan and Payne, 2015). With constantly increasing presence of high-throughput molecular profiling, it becomes increasingly important to ensure that data interpretation capabilities follow generation of large-scale biomedical data sets (Costa, 2014; Mardis, 2010). Visualization can support greatly the processing of complex data sets on each of the steps of the data life cycle. This opportunity is actively explored in various domains of biomedical research, including clinical big data (West et al., 2015) or multiscale biomedical ontologies (de Bono et al., 2012).

Modern translational medicine approaches aim to combine clinical and molecular profiles of the patients to formulate informed hypothesis on the basis of stratified data (Tian et al., 2012). Integration of plethora of sources renders

these data sets complex and difficult to process. Visualization of such integrated data sets can aid exploration and selection of key dimensions and subsets for downstream analysis. In turn, visually aided data analysis allows to comprehend even complicated workflows and aids interpretation of resulting data.

In this project, we demonstrate a workflow for analysis and interpretation of high-throughput translational medicine data, in which visualization is an important component at each step of data processing and exploration. In this workflow, three Web services - tranSMART, a Galaxy Server, and a MINERVA platform - are combined into one big data pipeline. We call it TGM (tranSMART - Galaxy - MINERVA) pipeline.

8.2 Implementation

Here I'm providing the outline of the implementation, but for more details refer to my publication (Satagopam et al., 2016) on this project provided in the Publications section 9.3. In the TGM pipeline as shown in Figure 8.1, our data integration platform of choice was tranSMART as it is a server-based solution with Informatics for Integrating Biology and the Bedside (i2b2) (Murphy et al., 2009, 2010) data exploration component. We chose Galaxy as a workflow management system, considering its flexibility and the availability of tools. Finally, to provide informative interpretation of analytical results, we bridged the Galaxy Server with MINERVA platform, allowing overlay of exported data on disease-related mechanisms.

We approached this problem in three steps:

1. Data integration and exploration are handled using tranSMART repository (Szalma et al., 2010)

2. Analysis of tranSMART-provided data is supported by Galaxy Server workflows (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010)

3. Visualization of Galaxy-provided results is enabled via domain-specific knowledge repositories (Fujita et al., 2014).

Figure 8.1: **tranSMART-Galaxy-MINERVA pipeline**. The figure shows the 3 different services integrated in the pipeline for clinical, translational and associated molecular data integration, analysis and interpretation of analytical results using contextualized knowledge repository e.g., Parkinson's disease (PD) map (Fujita et al., 2014; Gawron et al., 2016).

8.2.1 Integration of clinical and molecular data in tranSMART

Translational medicine data sources are heterogeneous and of various granularities (Bender, 2015; Martin-Sanchez and Verspoor, 2014), and visually aided data exploration (Shneiderman et al., 2013) is an important enabling technology for biomedical experts. The powerful visualization and interoperability functionalities of i2b2 are coupled together with omics integration in tranSMART (Athey et al., 2013; Szalma et al., 2010) platform. tranSMART is a well-established platform enabling translation of preclinical research data into meaningful biological knowledge (Scheufele et al., 2014). It supports integration of low-dimensional clinical data and high-dimensional molecular data in a data warehouse architecture. Although tranSMART by default relies on a relational database technology, it extends toward storing the high-dimensional biological data using NoSQL technology HBase (Wang et al., 2014). The platform features data interoperability connectors, including clinical information collection (Blond and de Bruijn, 2015), imaging data (Vast, 2015), visual analytics (Herzinger, 2015; Herzinger et al., 2017), or bioinformatics workflow management (Bierkens et al., 2015). Finally, tranSMART features builtin data mining and analysis applications based on open-source systems, such as i2b2 and GenePattern (Szalma et al., 2010), and provides plugins to external tools, such as Dalliance Genome Browser (Down et al., 2011), or APIs for statistical packages, such as R.

In order to take advantage of the multiple functionalities of tranSMART, the

target data sets have to be integrated following strict rules of data harmonization, semantic alignment, and quality checking. The data sets are curated following three common steps:

1. Data extraction: Source raw data files are extracted from either public or private data repositories. This could be a simple FTP transfer from a Web repository or a database dump from a database management system, such as MySQL or Oracle.

2. Data retrieval: Target information from the raw source files is identified and converted as Standard Format Files as defined by tranSMART curation guidelines. At this step, subject-level to sample-level data mapping is established.

3. Data annotation: Completing and standardizing annotations of metadata are also expected for guaranteeing data provenance.

The final product of the abovementioned steps is a set of Standard Format Files, which are used as input by tranSMART's native ETLs (Extract, Transform, and Load) scripts. After data curation and loading to tranSMART, features collected for subject and samples become variables of the corresponding data set. These variables, as well as the relationships among them, are represented as a hierarchical parent - child tree control structure (i2b2 tree). One can explore this tree and use these variables in filtering the cohort and creation of sub-cohorts for further analysis.

8.2.2 Analysis using Galaxy server

Galaxy as a bioinformatics workflow management system is available as both Web server and cloud workbench, offering flexibility in terms of data interoperability and allocation of computational resources (Abouelhoda et al., 2012; Afgan et al., 2011; Goecks et al., 2010). The Galaxy environment keeps track of every detail of the analysis, allows the building of complex workflows, and allows the results to be documented, shared, and guaranteeing reproducibility (Afgan et al., 2011). Galaxy Tool Shed (Lazarus et al., 2012) is a repository of more than 3000

community-developed tools, allowing easy and versatile establishing of bioinformatics workflows (Blankenberg et al., 2010; Giardine et al., 2005). Such workflows may combine different aspects of expert knowledge required in subsequent analytical steps. Basic knowledge about the system is sufficient to use default elements in the workflow construction. These default methods can be modified, where the user has sufficient expertise. Once the workflow step is done, users can easily share and modify it. Analytical results can be directly visualized using embedded functionalities or exported for downstream processing.

8.2.3 Interpretation of analytical results using MINERVA platform

High-dimensional translational medicine data sets are difficult material to draw conclusions relevant for human health. Data sets exported preselected from tranSMART database and analyzed using Galaxy will either, in many cases, remain multidimensional datasets or will be reduced to the list of prioritized molecules. Interpretation of such results remains challenging and requires both contextualization and visualization. Galaxy Server allows various export options. As the last step of our pipeline, we propose to interpret the results of analysis of Galaxy Server in the context of dedicated knowledge repositories supported by MINERVA platform, such as Parkinson's disease (PD) map (Fujita et al., 2014; Gawron et al., 2016). In particular, molecules prioritized by the constructed pipeline are automatically visualized on molecular interaction networks hosted by MINERVA platform (Gawron et al., 2016).

8.3 Summary of results

8.3.1 tranSMART-Galaxy-MINERVA (TGM) pipeline

The TGM pipeline as shown in Figure 8.1 is a workflow combining three server based platforms - tranSMART server, Galaxy server and MINERVA platform - to enable data integration, visually-aided exploration, bioinformatics workflow construction, analysis and interpretation of high-throughput translational medicine

data in the disease context.

tranSMART server with cohort data: The tranSMART part of the pipeline allows for exploration of integrated translational medicine data and for expert based selection of important subgroups. These are pipelined directly to the associated instance of Galaxy Server. A tranSMART instance has been setup with two Alzheimer's and 15 Parkinson's Disease studies and is available at `http://tgm-pipeline.uni.lu/transmart`. The raw data for these studies, both clinical and corresponding expression data is downloaded from Gene Expression Omnibus (GEO) (Barrett et al., 2012). In the tranSMART instance dataset explorer interface, user can navigate terms from the 'Navigate Terms' panel on the left side of the interface and click the folder structure (or i2b2 data tree) to explore the studies and clinical variables (concepts) for each subject in the study use them in slicing and dicing of cohort data to build sub-cohorts and export them to Galaxy for analysis as shown in Figure 8.2. In this use case, we worked with the GSE7621 PD study data (Lesnick et al., 2007) for defining two focused cohorts using tranSMART dataset explorer.

Study-related variables in tranSMART can be assigned to two broad categories: low- and high-dimensional data. Low-dimensional data correspond mostly to clinical, patient-centric data (e.g., systolic blood pressure) and low-throughput diagnostic measurements (e.g., quantification of a disease-related blood biomarker). The corresponding values of low-dimensional data are usually available as text or numeric values. High-dimensional data, in the majority reflecting 'omics' data, are structured as a numeric matrix. In this example two distinct subsets are defined based on the variables 'disease state' and 'gender'.

Analysis of a selected subset on Galaxy: High-throughput data provided by tranSMART contains gene expression in samples from the two selected cohorts: males with PD (four samples) and age-matched healthy males (eight samples). We have designed a dedicated Galaxy workflow and it is subdivided into steps from incorporation of the input files taken from tranSMART through performing the marker selection (differential expression) analysis and uploading the obtained results, list of significant transcripts to the PD map hosted on the

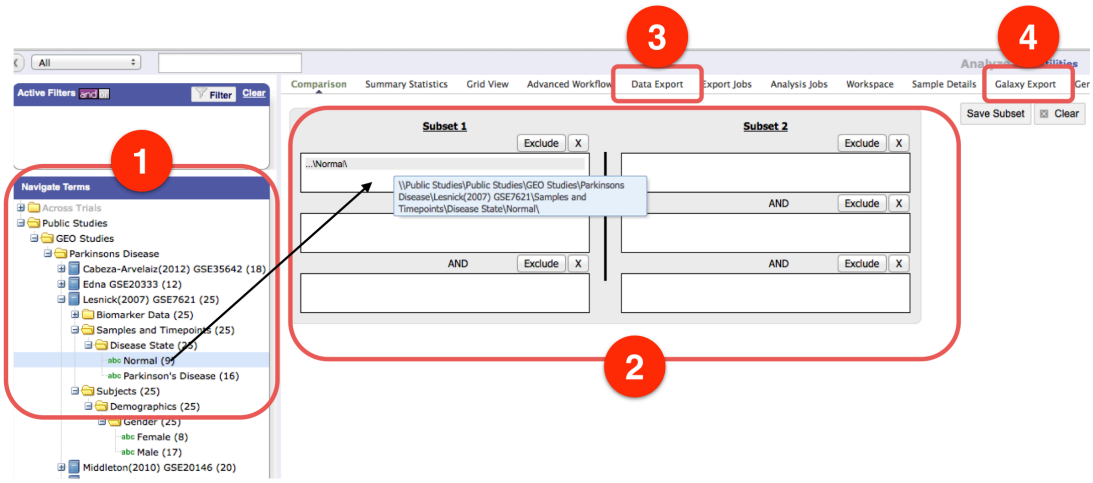


Figure 8.2: **tranSMART instance with cohort data**. The figure shows the tranSMART instance dataset explorer interface and steps to follow to create the sub-cohorts and export the data to Galaxy server. In this interface, users can navigate terms from the 'Navigate Terms' panel on the left side of the interface and click the i2b2 tree structure to explore the studies and clinical variables for each subject in the study as shown in Step 1 part of the figure. Then one can drag and drop each concept to the right panel into either 'Subset 1' or 'Subset 2' to build customized cohort for comparison based on the filtering of the chosen concepts shown in Step 2. After customized cohort is defined, then using the 'Data Export' function in on the top of the interface, user can export the selected dataset (Step 3). Once the export is finished, user can click the 'Galaxy Export' to export the dataset to the Galaxy server. In this step a name should be given by the user as a identifier of the dataset in the Galaxy server (Step 4).

MINERVA platform and making them accessible for interpretation in the disease specific context. Galaxy server interface with different sub tasks has been depicted in Figures 8.3, 8.4. In order to facilitate the PD use case, Galaxy instance with 'tranSMART PDMap' workflow has been setup and is available at `http://tgm-pipeline.uni.lu/galaxy`.

A comparison between these two data sets provides insight about disease-related mechanisms that may be cohort specific. This differential gene expression was calculated as predefined method using Bioconductor package 'limma' in Galaxy (Ritchie et al., 2015) (absolute fold change >1.5, p-value <0.05). The resulting list of 3,286 differentially expressed genes (DEGs) was uploaded via MINERVA to the PD map for visual interpretation. This process led to the labeling of 224 different dysregulated genes involved in PD and/or their related protein products in the PD map. The current version (September'17) of PD

map consists of roughly 5,000 elements linked by over 2,000 interactions. Among these 5,000 elements, 1,400 are genes/proteins and they are involved in pathways and neurodegenerative processes of the neuronal system, in particular on the degeneration of dopaminergic neurons of substantia nigra pars compacta. In this example only 224 genes out of 3,286 DEGs mapped on PD map, remaining DEGs are unknown and need further exploration.

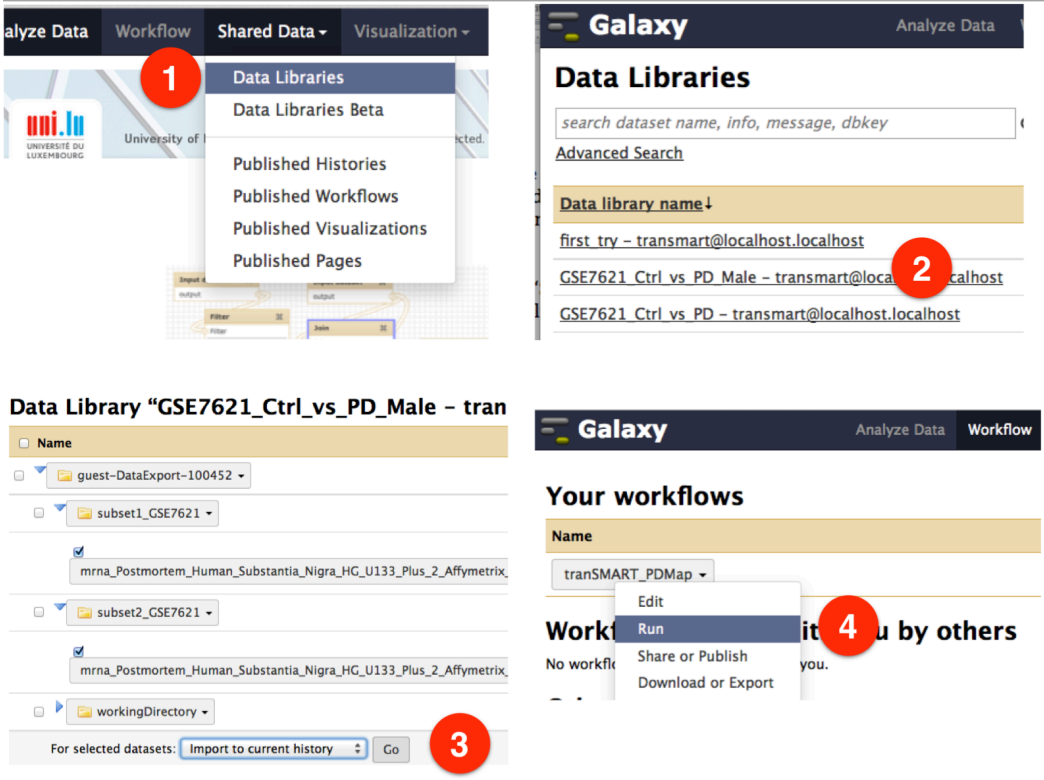


Figure 8.3: **The Galaxy interface to choose dataset and workflow**. The figure shows the Galaxy part of the pipeline to run analytical workflows for data from the tranSMART server. Analytical results (list of molecules) are sent to the associated MINERVA instance, hosting the PD map. In this Galaxy interface, by selecting 'Shared Data' from top navigation bar, then selecting the 'Data Libraries' menu item (Step 1), one can list the available datasets in Galaxy and choose the dataset to be analyzed. The datasets that were exported from tranSMART can be selected (Step 2) and imported into to the current history (Step 3). Then by selecting 'Workflow' tab from top navigation bar, the user can select the workflow of choice (Step 4) e.g., 'tranSMART PDMap' workflow to analyze tranSMART exported gene expression data to find out differentially expressed transcripts and send these list of biomolecules with corresponding p-values and fold-changes to PD map hosted by MINERVA service.

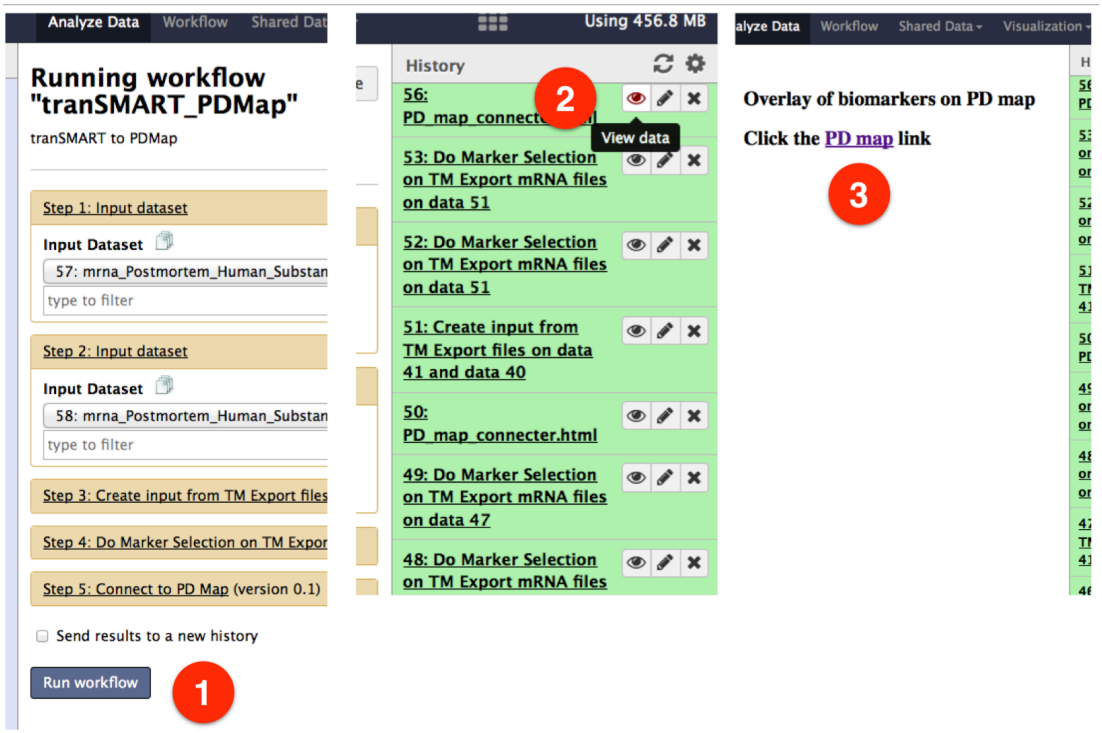


Figure 8.4: **The Galaxy interface to view results and link to PD-Map**. The figure shows workflow steps to run selected workflow upon choosing the imported datasets as inputs (Step 1). After the analysis is finished, one can view the results by click the eye-shape icon ('View data', Step 2) e.g., differentially expressed genes (DEGs). These DEGs can be overlaid on PD-Map hosted by MINERVA service (Step 3).

Interpretation of analysis results in the PD map: The MINERVA part of the pipeline allows interpretation of analytical results from Galaxy server in the context of hosted molecular interaction networks. Here, Galaxy results are projected on the PD map, and Parkinson's disease related mechanisms. A dedicated MINERVA supported Parkinson's disease map has been setup and is available at `http://tgm-pipeline.uni.lu/minerva`. MINERVA platform accepts POST requests, where the user specifies the target molecular network, user, password, and the data set to be uploaded. To ensure seamless data transfer from Galaxy to MINERVA, we created a step in the Galaxy Server Workspace (GSW) called 'PD map connector'. The code explaining the construction of request to MINERVA hosted PD map can be found here (Satagopam and Schneider, 2016). This step generates a POST request to the associated MINERVA instance - PD map in this

case - to generate a custom visualization on the basis of the workflow data.

By seamlessly connecting Galaxy Server to MINERVA platform, the users can securely transfer analysis results obtained from Galaxy workflows to MINERVA platform without leaving the Galaxy system. As shown below in Figure 8.5, visualization of the results on the PD map allows the identification of major molecular pathways perturbed in postmortem brain tissue of male Parkinson's patients, as selected in tranSMART and processed in Galaxy.

From our publication (Satagopam et al., 2016), evaluation of highlighted areas in the PD map shows pronounced alterations in the cell nucleus, in particular a battery of downregulated (red) genes involved in metabolism and secretion of the neurotransmitter dopamine (Figure 8.5, blue circle) (Meiser et al., 2013). Another visible perturbation affects the mitochondria, in particular elements of complex I (Figure 8.5, red circle). This process is essential for energy homeostasis, in particular in high energy demanding neurons. Finally, we observe upregulation (green) of processes involved in neuroinflammation (Figure 8.5, purple circle) (Glass et al., 2010). On the basis of this visual exploration, data analyst may get comprehensive insights in molecular processes potentially involved in the disease of this specific patient cohort supporting new insights for diagnosis, prognosis, and therapy. Another approach for visualization is the drug target interface integrated in the MINERVA platform, enabling the mapping of potential drug interactions with elements of the map, suggesting more precise treatments and possibly an improvement in existing therapies (Poletti and Bonuccelli, 2013)

8.4 Discussion

Clinical and translational medicine projects generate large amounts of data including clinical, molecular (various types of -omics), imaging and kinetic (mobile, sensor) data. Efficient data capturing, curation, harmonization, integration and analysis are challenging and pivotal for stratification of the patients, early detection of biomarkers and drug targets. We show here that visualization and interoperable workflows, combining multiple complex steps, can address at least parts of the challenge. In this project, we present an integrated workflow for exploring, analysis, and interpretation of translational medicine data in the context

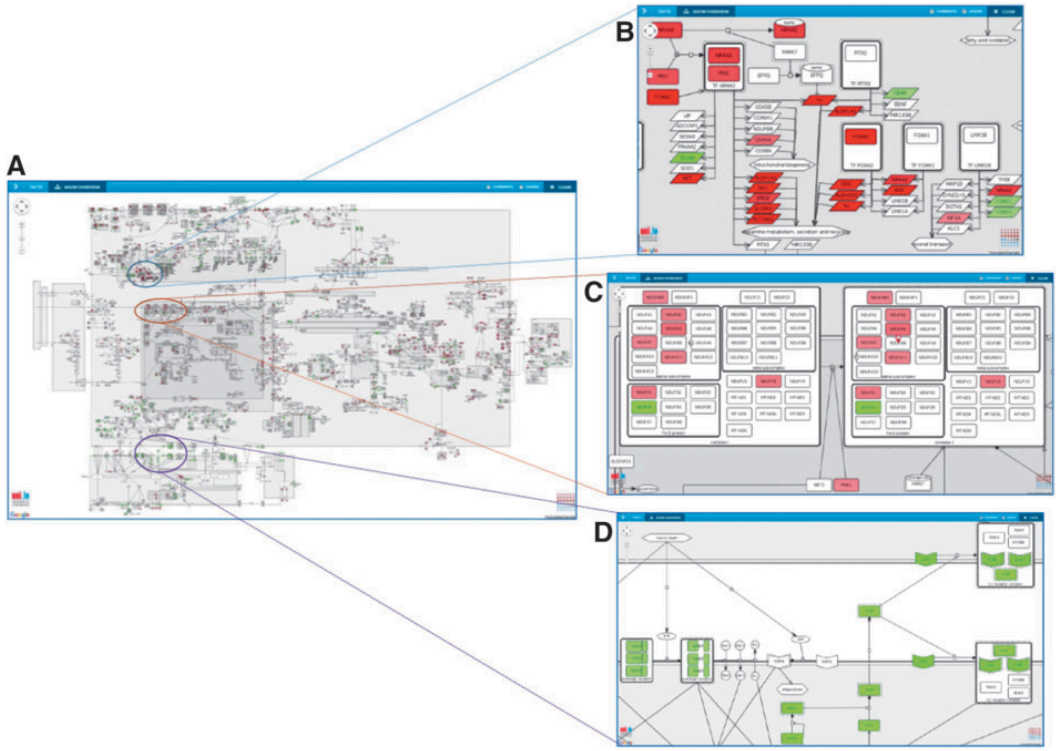


Figure 8.5: **Data visualization and analysis using PD map**. (A) Differential gene expression data comparing postmortem brain tissues from male PD patients versus controls are displayed on the PD map (green, upregulated; red, downregulated). Pathways and processes of conspicuous areas (colored circle) could be identified using the pathway and compartment layout view of the PD map. Detailed view on deregulated genes that encode for proteins involved in dopamine metabolism, secretion, and recycling (B), on mitochondrial electron transport chain, in particular elements of complex I (C), and on microglia activation (D). ©This figure is from my own publication (refer section 8.6) (Satagopam et al., 2016).

of human health. Three Web services - tranSMART, a Galaxy Server, and a MINERVA platform - are combined into one big data pipeline. Native visualization capabilities enable the biomedical experts to get a comprehensive overview and control over separate steps of the workflow. The capabilities of tranSMART enable a flexible filtering of multidimensional integrated data sets to create subsets suitable for downstream processing. A Galaxy Server offers visually aided construction of analytical pipelines, with the use of existing or custom components. A MINERVA platform supports the exploration of health and disease related mechanisms in a contextualized analytical visualization system. We demonstrate

the utility of our workflow by illustrating its subsequent steps using an existing data set, for which we propose a filtering scheme, an analytical pipeline, and a corresponding visualization of analytical results. The workflow is available as a sandbox environment, where readers can work with the described setup themselves. Overall, our work shows how visualization and interfacing of big data processing services facilitate exploration, analysis, and interpretation of translational medicine data.

8.5 Future development and directions

In our group we are currently working on different clinical and translational projects e.g., eTRIKS: European Translational Information & Knowledge Management Services (https://www.etriks.org), NCER-PD: National Centre of Excellence in Research on Parkinson's Disease (http://www.ncer-pd.lu). We would like to apply this TGM pipeline to these projects. We are planning to add more workflows to Galaxy. In the case of NCER-PD project, we are currently building an 'International Parkinson's Disease Variant Database' and will be connected to tranSMART and MINERVA platform. This new pipeline will facilitate the slicing and dicing of the clinical data (selection of sub-cohorts) in tranSMART instance and enrich the significant disease causing variants by connecting to the variant database and overlay these variants on PD map hosted in MINERVA platform for visual exploration of contextualized Parkinson's Disease knowledge.

Contributions: Myself, Wei Gu and Serge Eifes conceived and designed the project. Adriano Barbosa da Silva, Wei Gu and Serge Eifes curated the data and integrated into tranSMART. Myself and Serge Eifes implemented Galaxy workflows. Myself and Piotr Gawron developed the Galaxy-MINERVA interface. Stephan Gebel interpreted the experimental results. Myself, Reinhard Schneider and Rudi Balling supervised the project.

8.6 Publication

This research work has been published in Big Data journal (Satagopam et al., 2016) and the article is provided in the Publications section 9.3.

Chapter 9

Conclusions

In recent years we are witnessing breakthrough improvements of biomolecular knowledge and technologies. During this period many high-throughput technologies have been developed to investigate various aspects of cellular processes, such as sequence and structural variations of the genome, transcriptome, epigenome, proteome, metabolome, interactome (collectively -omics data), imaging, kinetic (mobile, sensor) and clinical data.

This technological advancement greatly helps scientists to study both complex diseases such as cancer, diabetes, neurodegenerative diseases like Parkinson's disease, Alzheimer's disease as well as normal physiological conditions at different levels: at cell, tissue, organ level and system as whole. There are several biological entities (e.g., genes, proteins, chemicals, metabolites, regulatory elements, non-coding RNAs) involved in these biological processes. These high-throughput experiments are producing an explosion of different types of data. There is a pressing need to integrate these heterogenous experimental datatypes, analyse and prioritise the bio-entities for either early detection of diseases (as disease biomarkers) or efficient patient stratification for personalised therapies. On the other side, there are a large amounts of published literature (>27millions scientific papers in PubMed) and various publicly available biological databases. We would like to take advantage of these vast amounts of public knowledge to annotate, validate and prioritise the bio-entries, markers, drug targets identified from above mentioned high throughput experiments. The main problem of public data resources is they are heterogenous in both content (e.g. genes, proteins,

regulatory elements, chemicals, metabolites, diseases, ontologies, reactions, interactions, pathways, literature etc.,) and format (e.g. flat-file, XML, relational database). These public data need to be parsed and integrated in order to use the knowledge seamlessly.

There is also a lot of valuable information available in the literature. For example, researchers are often interested in bio-entities associated with diseases, the phenotypic variation resulting from naturally occurring single nucleotide polymorphism or directed mutagenesis. Traditionally, this information must be manually collected by trawling literature repositories such as PubMed using author names, protein/gene/disease names, keywords or a mixture of all three. The vast amounts of literature makes such an approach inadequate.

In order to address the above mentioned needs, my thesis has focused on:

1. Data integration, knowledge management and analysis,
2. Text-mining of literature to extract knowledge like mutations,
3. Clinical and translational medicine data integration and visualization.

9.1 Data integration, knowledge management and analysis

9.1.1 bioCompendium

To address the data integration and knowledge management needs, I have developed bioCompendium, a knowledge base consisting of fine grained knowledge. It consists of more than 80 important public biological databases collected locally and integrated into the SRS system. In addition, it also integrated data related to model organisms (human, mouse, yeast) from Ensembl, BioMart, a text mining resource (AKS2) into bioCompendium data layer. All these databases are scanned for each gene from human, mouse and yeast and the relevant information is stored in a MySQL database (bioCompendium knowledge base). As scientists use different database identifiers (IDs) to represent their experiment

results, in order to compare the results from one IDs system to another, I have implemented an ID Conversion service. Researchers also perform experiments in different model organisms. Therefore, to achieve the cross-species comparison and enrichments, I am converting information from one genome to another in realtime by using orthology relationships. One more unique feature of bioCompendium is the handling of documents (PDF, MS-Word, Excel, PowerPoint, plain text) to extract gene lists. I have implemented an API to access the Reflect (O'Donoghue et al., 2010), a bio-entity tagging service. After converting any of the above mentioned documents to ascii text, that text is sent to the Reflect via an API and the tagged version of the text is received back. The bio-entities (genes/proteins) are extracted and processed like a gene set. The information in bioCompendium is Ensembl gene centric, all the information from methods - ID Conversion service, Cross species comparison based on orthology, Handling of documents, boils down to Ensembl genes and provides several bioinformatics analysis results.

bioCompendium web interface supports autocomplete based simple search and provides summary sheets for each gene/protein with rich annotations. It offers sequence similarity (homology) based as well as domain architecture similarity based clustering workflows and KEGG, PANTHER, Reactome, GeneOntology (GO) enrichments for given lists of genes/proteins. It provides transcription factor binding site profiling using 5kb upstream region staring from the first exon of each gene. bioCompendium integrates chemical information from several publicly available databases and literature and that is mapped to protein targets. In order to know if any of the protein or part of the protein sequence from user input is already patented, these sequences are searched in the patented sequences that are collected from European, US, Japan and Korean patent offices. This module presents further details of the patent hits and the records are linked back to the respective patent office.

bioCompendium is also equipped with 2D (Medusa, a java applet) and 3D (Arena3D, a java application) visualisation tools to visualise selected genes/proteins and their relationships to pathways, diseases, chemicals (drugs, ligands, metabolites) and GO terms. It also provides tissues expression profiling and highly relevant literature information by running customised queries to PubMed. The abstracts are tagged with Reflect service that provides useful popups for

genes, proteins, chemicals and wikipedia terms. Even though these two domains are exclusively developed for betaJUDO project, they will be soon made available for the public version of the bioCompendium. As bioCompendium handles the documents (PDF, MS-Word, Excel, PowerPoint, plain text), one can use this functionally to compare the experimental results with already published data by directly uploading gene dataset(s) and one or more PDFs of published literature. It will help the researcher by avoiding the painstaking step of collecting gene names manually from the literature. This functionality is also helpful to just extract the bio-entities from one or more scientific publications.

Apart from web based functionalities, bioCompendium provides RESTful API. It is very practical for programmers and bioinformaticians to access its functionality programatically. bioCompendium both front end - the web application and the backend - the database are designed in a very modular way. It is easy to add new knowledge domains to the existing schema and to develop respective web interface views to serve the new knowledge.

bioCompendium is a 'work horse' in our group to analyse datasets of genes or gene products obtained from various high throughput experiments from different projects. As this web application is open to the public it became a tool of choice not only for our institute but also for the researchers from all over the world as indicated in the web statistics of bioCompendium (refer section 2.4.1 from Chapter 2). bioCompendium provides two types of sessions, (a) temporary session; (b) permanent session through its session management component. Uploaded gene/protein datasets and analysis results are organised in temporary sessions and will be deleted at regular intervals with a cron job, but in-house and collaborative projects are kept longer using permanent sessions. Table 2.3 from Chapter 2 provides the list of the in house and collaborative projects, in which bioCompendium is used to analyse and annotate the experimental results. Some of these applications of bioCompendium are briefed below:

Integration with Garuda platform as Garuda gadget: 'Garuda - the way biology connects' is an ongoing effort to build an open source, community driven framework to integrate several bio-medical services (bioinformatics, computational and systems biology applications, algorithms, databases) thorough a com-

mon interface (Ghosh et al., 2011). As part of this initiative, I have integrated two services - bioCompendium analytics and SBML annotation service into Garuda using APIs from respective services. bioCompendium analytics plug-in within Garuda dashboard can analyse gene or protein datasets coming from any other tool. The SMBL annotation server, on the other hand, takes SMBL files as input, for example coming from a modeling tool of biochemical networks, CellDesigner. The annotation service parse the SMBL file and annotate each gene and protein with rich knowledge obtained as a result of integrating several publicly available biological databases.

betaJUDO analysis with bioCompendium: In this project, the palmitate treated and control samples were analyzed using LC-MS/MS mass spectrometry. The data has been analyzed and the significant proteins annotated and enriched with bioCompendium. Apart from the classical features of bioCompendium, I have added two new domains to the bioCompendium knowledge base exclusively for this project - a) tissue expression information from the Human Protein Atlas (HPA) and b) context specific literature that is very relevant for betaJUDO project. But eventually these new domains will also be made available in the public version of bioCompendium. This proteomics data has been integrated with lipidomics data and this combined analysis of isolated human islets exposed to palmitate reveals time-dependent changes in insulin secretion and lipid metabolism (Roomp et al., 2017).

A permanent session has been created in bioCompendium and is made available at URL `http://biocompendium.embl.de/betajudo` for betaJUDO consortium members.

Progeria: The global gene expression changes in fibroblasts from human subjects with HGPS were analyzed and found that 352 genes were significantly differentially expressed between fibroblasts from subjects with HGPS and controls. Protein interaction network analysis using MetaCore database (www.genego.com) and STRING interaction database identified Rb1 as the only one encoding a protein product, Rb, known to interact directly with A-type lamins (Mancini et al., 1994; Ozaki et al., 1994). The expression of Rb1 was downregulated in HGPS.

Microarray analyses results were validated by real-time RT-PCR. The differentially expressed genes in HGPS indicated that Rb is a key regulatory component affected by LMNA mutation and that it is at the center of a signaling network that is abnormally active in the disease. Thus, therefore, suggest that lamin A - Rb signaling network is a major defective regulatory axis. Lonafarnib, a protein farnesyltransferase inhibitor, treatment of fibroblasts with this drug reversed the gene expression defects. This study identifies Retinoblastoma protein (Rb) as a key factor in HGPS pathogenesis and suggests that its modulation could help in premature ageing and as well as physiological ageing.

Differentially expressed gene sets from progeria project and their analysis using bioCompendium are available at `http://biocompendium.embl.de/Progeria`.

TAMAHUD: The RNAi screen in a human embryonic kidney (HEK293) T-REx cell line overexpressing a full length mutated Huntingtin construct followed by validation in Drosophila, has yielded a number of potentially druggable targets which are suitable for HD. To gain further insight into the biological relevance of the data generated, TAMAHUD knowledgebase based on both public and experimental data has been developed.

The annotated TAMAHUD hits and their enrichments were organised in different logical sections as depicted in Figure 4.1. The gene expression datasets were organised in an expandable tableview along with number of significant DEGs and their corresponding p-value and fold-change cut offs. The user can select one or more DEG lists and send them to the bioCompendium service on the fly for further exploration and analysis as shown in Figure 4.2.

With the help of this centralized TAMAHUD database and the web server, we categorized the different sets of HD toxicity modulators according to their molecular function. Suppressors were enriched for certain classes of proteins such as GPCRs or transporters compared to the initial library, whereas the number of positive kinases in the screen was reduced, and no cytokines, growth factors or translational regulators were represented. We observed similar functional categorizations after selection from the cell and Drosophila screen. An Ingenuity Pathway Analysis of the hits obtained in the primary screen in cells shown in Figure 4.3 revealed that the majority of these proteins participate not only in general

processes such as GPCR- or cAMP-mediated signaling but also in canonical pathways related to neurodegeneration such as apoptosis, mitochondrial dysfunction, amyloid processing or protein ubiquitination. Notably, ten of these proteins have been previously related to HD signaling, including subunits of the succinate dehydrogenase complex and HTT-associated protein 1. Many of the genes validated in Drosophila are also involved in processes related to neurodegeneration but are enriched in mitochondrial metabolic pathways, especially those associated with fatty acid biosynthesis and metabolism.

TAMAHUD knowledgebase with both TAMAHUD experimental results as well as public high-dimentional Huntington's disease datasets has been developed as a result of this collaborative project. This is useful for many researcher working in the area of Huntington's disease and more broadly neurodegenerative area.

SMBL Map Annotation Service: This service has been developed in the context of MINERVA platform (Gawron et al., 2016) to annotate the hosted Parkinson's Disease (PD) map (Fujita et al., 2014) elements - genes, proteins, chemicals, drugs, metabolites. Here, I have applied data integration and knowledge management approaches to take advantage of vast amounts of publicly available biological databases (DBs). These DBs are integrated and indexed into Sequence Retrieval System (SRS) (Etzold and Verde, 1997). An ETL (Extract, Transform, Load) process has been developed and a comprehensive knowledgebase (SynonymDB) of bio-entities (e.g., genes, proteins, chemicals, drugs, metabolites), their synonyms, annotations and cross references to other databases for seven model organisms: yeast, worm, fly, zebrafish, rat, mouse and human is developed.

Two dedicated web-services have been developed and both of them are available under bioCompendium as sub-services. 1) SynonymDB web application, which takes list of bio-entities or their synonyms as input and return the annotated bio-entities and is available at `http://biocompendium.embl.de/synonym_db`; 2) Map annotation service, enriching the different molecular interaction networks and disease maps e.g., Parkinson's disease map by providing rich annotations to various species elements of the map and is available at `http://biocompendium.embl.de/map_annotator`. Both services are available as web

applications with RESTful APIs. As this resource consists of synonyms, it also serves to harmonise the bio-entities to the standard database identifiers. This part of the API call takes a list of synonyms as input and returns the primary identifiers/accession numbers of the database of user's choice.

Apart from the disease map annotation, systems biology tools including - CellDesigner, GARUDA, MINERVA platform - took the advantage of these biological data driven services to develop plugins (gadgets) using provided RESTful APIs to standardise, harmonise bio-entities and/or annotate them. In the future we would like to add more biological databases to this knowledgebase to cover more bio-entities.

9.1.2 Human-gpDB

Human-gpDB is a database employing visualization tools and data integration techniques to integrate GPCRs, G-proteins, effectors and their interactions. G-protein coupled receptors (GPCRs) are a major family of membrane receptors in eukaryotic cells. They play a crucial role in the communication of a cell with the environment. Ligands bind to GPCRs on the outside of the cell, activating them by causing a conformational change, and allowing them to bind to G-proteins. Through their interaction with G-proteins, several effector molecules are activated leading to many kinds of cellular and physiological responses. The great importance of GPCRs and their corresponding signal transduction pathways is indicated by the fact that they take part in many diverse disease processes and that a large part of efforts towards drug development today are focused on them. We have developed Human-gpDB, a database which currently holds information about 713 human GPCRs, 36 human G-proteins and 99 human effectors. The collection of information about the interactions between these molecules was done manually and the current version of Human-gpDB holds information for about 1663 interactions between GPCRs and G-proteins and 1618 interactions between G-proteins and effectors.

Human-gpDB, compared to the previous gpDB databases (Elefsinioti et al., 2004; Theodoropoulou et al., 2008) is now richer and contains new information concerning the classification of GPCRs (11 new subfamilies were added and all

existing subfamilies are classified based on the IUPHAR classification) and also contains interactions between all molecules. It is fully integrated with external data sources by bridging information that did not exist in the previous versions (e.g. drugs and chemicals). Human-gpDB database was built to provide a simple but yet a powerful tool for researchers in the life sciences field as it integrates a current, careful collection of human GPCRs, G-proteins, effectors and their interactions. Human-gpDB uses advanced visualization techniques to make the volume of data more informative and the advanced data integration techniques make Human-gpDB a unique tool, a reference guide in pharmaceutical research and especially in the areas of chemical and drug discovery for human diseases. In the future, the expansion of the current version of the database for other organisms starting from the ones that are evolutionarily closer to Humans is essential.

In all these projects that have been described so far - bioCompendium, betaJUDO, Garuda, Progeria, TAMAHUD, Map Annotation Service and Human-gpDB, I have applied data integration, knowledge management, analysis and visualization techniques to integrate both data from high-throughput experiments and rich publicly available biological databases. Vast amounts of this public knowledge, application of various analytical and visualization methods are helping in annotation and enrichment of experimental results and their prioritisation for further experimental validation.

Future development and directions: As there is lot of interest in bioCompendium resource from the scientific community, I would like to further continue developing the service. The transcription factors binding site profiling section need to be improved by providing the possibility for the user to select the genomic regions of interest. Also apart from the predictions based on PWMs, I would like to integrate ChIP-chip and ChIP-seq experimental data from ENCODE (ENCODE-Consortium, 2011) and other projects. There is lot of room to improve the visualisation part of the application, especially 2D Medusa applet, will be replaced by interactive javascript based visualisation to visualise the interactions between bio-entities.

More model organisms will be added to the resource, four new ones are in the

pipeline - rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), fruitfly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*). New knowledge domains will be added to the bioCompendium, especially noncoding RNAs that are playing very important role and we have limited knowledge about functionality of these ncRNAs and their interactions with other bio-entities. miRNAs, siRNAs, pseudogenes, circular and long ncRNAs are planned to be integrated into bioCompendium. Another important domain is mutations. We will systematically mine literature for mutations and their effect, both will be integrated into the resource.

Scientist working in plant science research would like to have similar resource for plant species. The bioCompendiumPlants will be developed in collaboration with plant science community.

9.2 Text-mining of literature to extract HIV mutations

Vast amounts of knowledge is buried in the unstructured literature without semantics and most of these full-text articles are not available for text-mining. Currently PubMed contains more than 27 millions publications, only a fraction (1.6 millions, 6%) of these publications are available for bulk retrieval and text mining. This was a major issue in mining full text articles.

Fortunately, recent pressure from government and scientific bodies and the rise of open access publishing has softened the stance of publishers and many are now receptive to waiving these restrictions. In our case, thankfully, the majority of the publishing companies and societies that we approached granted us permission to text-mine and index HIV mutation information contained in their literature. The list of publishers and journals that have given permission to the HIV mutation browser to access, data-mine and display articles can be found from the HIV mutation browser website (`http://hivmut.org/index.php?page=about`).

HIV is an important therapeutic target and has been the subject of a major research effort as evidenced by the large catalogue of HIV experimental literature. Currently there are over 275,000 HIV related articles in PubMed. We have the

permission to process approximately 45% of these full text articles published across 2,639 journals and these are used for text-mining to build the HIV mutation browser. These articles were mined for mutation information, and they were mapped to corresponding residues of HIV proteins. From this literature 8,061 distinct mutations were extracted. As each publication can consist of multiple mutations and each mutation can be mentioned in multiple publications, the 8,061 distinct mutations were from 52,281 mutation mentions in 5,875 articles.

The extracted mutations provide insight into the nature of the HIV research in the last decades. In our repository 2,990 of the 3,118 residues in the HIV proteome have one or more mutations. This indicates the HIV research has been broad in scope. On the other hand certain regions such as host interaction interfaces, catalytic sites of the protease and reverse transcriptase, are highly studied (refer Figure 7.3). HIV mutation database protein statistics are shown in the Table 7.2. It shows the distribution of the mutation data across the proteins of the HIV proteome. The number of publications, distinct mutations and distinct mutated positions per protein vary widely. The protein Pol which contains the 3 enzymatic chains, a protease, an integrase and a reverse transcriptase is by far the best studied protein.

The repository has been populated with text-mining results, and it is therefore the database contains some incorrectly assigned citations that are unavoidable. We have therefore provided a user feedback interface that allows users to curate the data by flag the quality of an entry either positive or negative.

The HIV Mutation Browser is one of the first resources to text-mine mutagenesis data from the HIV literature (Doughty et al., 2011; Krallinger et al., 2009; Laurila et al., 2010). It will complement the resource such as Stanford Drug Resistance database (Rhee et al., 2003), the UniProt knowledgebase (UniProt-Consortium, 2014) and the Los Alamos HIV Database (`http://www.hiv.lanl.gov`) which are manually annotated and curated. As these text-mining methods are generic, we expect that these methods or similar ones can be applied to other viral or cellular systems.

This resource will continue to evolve as HIV literature is produced continuously at a rate of approximately 1,500 articles per month and the database will be update on a monthly basis. Even though the resource contains the HIV related

journals, it is still incomplete, as we did not receive permission from all publishers to text-mine their HIV related articles. Journals from other publishers will be added when possible. Other point, not all mutations can be correctly identified and assigned by the text-mining methods. The main reason, many mutations are annotated in an article using non-standard patterns that are not widely used to describe directed mutations in mutagenesis experiments or polymorphisms. We will continue to improve the methods for text-mining and assignment. We encourage the community to utilise the feedback system for misannotated mutations in the resource and contact us about mutation data that should be in the resource that is not present yet. This community input will improve the quality of the annotated data and will pinpoint parts of the text-mining method that require improvement.

In summary, the HIV Mutation Browser is a valuable addition to the currently available HIV resources that will allow researchers to quickly and intuitively access data on mutagenesis and phenotypic variation. We expect the database to aid the process of experimental design and be a key resource for the HIV community.

The pipelines and software modules developed to build the HIV Mutation Browser are generic enough to apply for any other model organisms. We are in the process of building similar resource for HCV and human. Recently scientists from the in the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) contacted us to build a similar resource for Ebola virus and at this moment we are collecting the literature related to Ebola virus.

9.3 Clinical and translational medicine data integration and visualization

Medical field is going through a revolution that will change the practice of healthcare in virtually every way. Translational Medicine(TM) is a highly interdisciplinary domain that connects benchside, bedside and community. The main goal of TM is to promote enhancements in diagnosis, therapies and in general to improve the global healthcare system significantly (Cohrs et al., 2015).

Nowadays, the process of translation is supported by large amounts of heterogeneous data ranging from medical data to a whole range of -omics data. It is not only a great opportunity but also a great challenge, as translational medicine big data is difficult to integrate and analyze, and requires the involvement of biomedical experts for the data processing. The rise of translational and personal medicine have been made possible due to the advancement in many high-throughput technologies to study the cellular processes and molecular functions of organisms at different levels: cellular, tissue, organ and system as a whole. These technologies such as genome, transcriptome, proteome, lipidome, metabolome, epigenome, and microbiome collectively -omics of both single as well multi-cell samples, their integration and systems biology, have greatly advanced our understanding of human health and diseases (Canuel et al., 2015; Hawkins et al., 2010). However, the progress comes at a cost, translational research data integration requires large resources and efforts. In effect, important steps of the data life cycle in discovery: collection, curation, integration, analysis, and interpretation are challenges for biomedical research. Moreover, enabling biomedical experts to efficiently use big data processing pipelines and make sense of it is another challenge.

As translational medicine data become more and more rich and complex, their potential in informing both clinical and basic research grows (Regan and Payne, 2015). With constantly increasing presence of high-throughput molecular profiling, it becomes increasingly important to ensure that data interpretation capabilities follow generation of large-scale biomedical data sets (Costa, 2014; Mardis, 2010). Visualization can greatly support the processing of complex data sets on each of the steps of the data life cycle. This opportunity is actively explored in various domains of biomedical research, including clinical big data (West et al., 2015) or multiscale biomedical ontologies (de Bono et al., 2012).

Modern translational medicine approaches aim to combine clinical and molecular profiles of the patients to formulate informed hypothesis on the basis of stratified data (Tian et al., 2012). Integration of plethora of sources renders these data sets complex and difficult to process. Visualization of such integrated data sets can aid exploration and selection of key dimensions and subsets for downstream analysis. In turn, visually aided data analysis allows to comprehend

complicated workflows and aids interpretation of resulting data. Efficient data capturing, curation, harmonization, integration and analysis are challenging and pivotal for stratification of the patients, early detection of biomarkers and drug targets. We show here that visualization and interoperable workflows, combining multiple complex steps, can address at least parts of the challenge.

In this project, we presented an integrated workflow for exploring, analysis, and interpretation of translational medicine data in the context of human health. Three Web services, tranSMART, Galaxy, and MINERVA (TGM) are combined into one big data pipeline. Native visualization capabilities enable the biomedical experts to get a comprehensive overview and control over separate steps of the workflow. The capabilities of tranSMART enable a flexible filtering of multidimensional integrated data sets to create subsets suitable for downstream processing. A Galaxy Server offers visually aided construction of analytical pipelines, with the use of existing or custom components. A MINERVA platform supports the exploration of health and disease related mechanisms in a contextualized analytical visualization system. We demonstrate the capability of our workflow by illustrating its subsequent steps using an existing data set, for which we propose a filtering scheme, an analytical pipeline, and a corresponding visualization of analytical results. The workflow is available as a sandbox environment, where one can work with the described setup themselves. Overall, our work shows how visualization and interfacing of big data processing services facilitate exploration, analysis, and interpretation of translational medicine data.

Future development and directions: In our group we are currently working on different clinical and translational projects e.g., eTRIKS: European Translational Information and Knowledge Management Services (`https://www.etriks.org`), NCER-PD: National Centre of Excellence in Research on Parkinson's Disease (`http://www.ncer-pd.lu`). We would like to apply this TGM pipeline to these projects. We are planning to add more workflows to Galaxy. In the case of NCER-PD project, we are currently building an 'International Parkinson's Disease Variant Database' and will be connected to tranSMART and MINERVA platform. This new pipeline will facilitate the slicing and dicing of the clinical data (selection of sub-cohorts) in tranSMART instance and enrich the signifi-

cant disease causing variants by connecting to the variant database and overlay these variants on PD map hosted in MINERVA platform for visual exploration of contextualized Parkinson's Disease knowledge.

In this thesis I have demonstrated the power of connecting the dots, the metaphor for integrating the valuable knowledge deposited in various biological databases, literature and use of this compendium of knowledge to fill the gaps and to achieve data driven hypothesis generation and validation. This biological data driven approach has been helping several in-house as well as external projects to annotate, enrich and prioritise the experimental results that can be further validated both in the context of pathological (disease) as well as normal physiological states. To keep these services up-to-date is a challenge, as we call 'sisyphean' task. With the availability of service monitoring tools and time-based job schedulers, most of these update tasks are done automatically as we demonstrated in the case of HIV mutation browser as well as SRS system.

These advanced data curation, harmonization, integration and visual analytics are paving a path for stratification of patients, early detection of bio-markers, drug targets and drug discovery towards personalised medicine.

Appendix A

This appendix contains the supplementary information for Chapter 6.

Example annotation of PD species - SNCA: An example annotation of a gene SNCA (alpha-synuclein) in SBML is shown below.

```
<species metaid="s402" id="s402" name="SNCA" compartment="c1"
    initialAmount="0">
<notes>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<title/>
</head>
<body>
[updated by PD map annotation service]
HGNC_ID: 11138
Symbol: SNCA
Name: synuclein, alpha (non A4 component of amyloid precursor)
Description: RecName: Full=Alpha-synuclein; AltName: Full=Non-A beta
    component of AD amyloid; AltName: Full=Non-A4 component of amyloid
    precursor; Short=NACP;
Previous Symbols: PARK1, PARK4
Synonyms: NACP, PARK1
Chromosome: 4q21.3-q22
```

RefSeq_ID: NP_001139527.1, NP_000336.1, NP_001139526.1, NP_009292.1

EnrezGene_ID: 6622

Ensembl_ID: ENSP00000378442, ENST00000394986, ENSP00000378437,
ENSP00000338345, ENSG00000145335, ENST00000394991, ENST00000336904

UCSC_ID: uc010ikt.1, uc003hsp.1, uc003hso.1

Reactome_ID: REACT_76627

EMBL_ID: AAA16117.1, AAA98487.1, AAA98493.1, AAC02114.1, AAG30302.1,
AAG30303.1, AAH13293.1, AAI08276.1, AAL15443.1, AAY88735.1, AF163864,
AK290169, AY049786, BAA06625.1, BAF82858.1, BC013293, BC108275,
CAG33339.1, CH471057, CR457058, D31839, DQ088379, EAX06036.1, L08850,
L36674, L36675, U46897, U46898, U46899, U46901

IPI_ID: IPI00024107, IPI00218467, IPI00218468

UniGene_ID: Hs.21374

PDB_ID: 1XQ8, 2JN5, 2KKW, 2X6M, 3Q25, 3Q26, 3Q27, 3Q28, 3Q29

DIP_ID: DIP-35354N

MINT_ID: MINT-2515483

KEGG_ID: hsa:6622

GeneCards_ID: GC04M090646

PharmGKB_ID: PA35986

Pathway_Interaction_DB: alphasynuclein_pathway

GO_Function: GO:0000287[magnesium ion binding] [IDA:UniProtKB],
GO:0005509[calcium ion binding][IDA:UniProtKB],
GO:0008198[ferrous iron binding][IDA:UniProtKB],
GO:0008270[zinc ion binding][IDA:UniProtKB],
GO:0019894[kinesin binding][IPI:UniProtKB],
GO:0030544[Hsp70 protein binding][IPI:UniProtKB],
GO:0042393[histone binding][IDA:UniProtKB],
GO:0042802[identical protein binding][IPI:IntAct],
GO:0043014[alpha-tubulin binding][IPI:UniProtKB],
GO:0043027[cysteine-type endopeptidase inhibitor activity involved in

apoptotic process][IDA:UniProtKB],
GO:0045502[dynein binding][IPI:UniProtKB],
GO:0048156[tau protein binding][IDA:UniProtKB],
GO:0051219[phosphoprotein binding][IDA:BHF-UCL]
GO_Process: GO:0001921[positive regulation of receptor
recycling][IDA:UniProtKB],
GO:0006916[anti-apoptosis][IMP:UniProtKB],
GO:0006919[activation of caspase activity][IDA:BHF-UCL],
GO:0010040[response to iron(II) ion][IDA:UniProtKB],
GO:0010517[regulation of phospholipase activity][IDA:UniProtKB],
GO:0010642[negative regulation of platelet-derived growth factor receptor
signaling pathway][IDA:UniProtKB],
GO:0031115[negative regulation of microtubule polymerization][IDA:BHF-UCL],
GO:0031623[receptor internalization][IDA:UniProtKB],
GO:0032026[response to magnesium ion][IDA:UniProtKB],
GO:0032410[negative regulation of transporter activity][IDA:UniProtKB],
GO:0032496[response to lipopolysaccharide][IDA:UniProtKB.],
GO:0032769[negative regulation of monooxygenase activity][IDA:BHF-UCL],
GO:0033138[positive regulation of peptidyl-serine
phosphorylation][ISS:BHF-UCL],
GO:0034341[response to interferon-gamma][IDA:UniProtKB],
GO:0035067[negative regulation of histone acetylation][IDA:UniProtKB],
GO:0045807[positive regulation of endocytosis][IDA:UniProtKB],
GO:0045920[negative regulation of exocytosis][IMP:UniProtKB],
GO:0048488[synaptic vesicle endocytosis][ISS:UniProtKB],
GO:0051281[positive regulation of release of sequestered calcium ion into
cytosol][IDA:UniProtKB],
GO:0051585[negative regulation of dopamine uptake][IDA:UniProtKB],
GO:0051612[negative regulation of serotonin uptake][IDA:UniProtKB],
GO:0051622[negative regulation of norepinephrine uptake][IDA:UniProtKB],

GO:0060732[positive regulation of inositol phosphate biosynthetic process][IDA:UniProtKB],

GO:0070495[negative regulation of thrombin receptor signaling pathway][IDA:UniProtKB],

GO:0070555[response to interleukin-1][IDA:UniProtKB],

GO:0071902[positive regulation of protein serine/threonine kinase activity][IDA:BHF-UCL]

GO_Component: GO:0005634[nucleus][IDA:UniProtKB],

GO:0005829[cytosol][IDA:UniProtKB],

GO:0005886[plasma membrane][IDA:UniProtKB],

GO:0005938[cell cortex][IDA:UniProtKB],

GO:0015629[actin cytoskeleton][IDA:UniProtKB],

GO:0030054[cell junction][IEA:UniProtKB-KW],

GO:0030424[axon][IDA:UniProtKB],

GO:0030426[growth cone][IDA:UniProtKB],

GO:0043205[fibril][IDA:UniProtKB],

GO:0045202[synapse][IEA:UniProtKB-SubCell]

InterPro: IPR001058, IPR002460, Synuclein_alpha.

Pfam: PF01387, Synuclein

PANTHER: PTHR13820, Synuclein

HOVERGEN: HBG000481

[updated by PD map annotation service]

</body>

</html>

</notes>

<annotation>

<celldesigner:extension>

<celldesigner:positionToCompartment>inside</celldesigner:positionToCompartment>

<celldesigner:speciesIdentity>

<celldesigner:class>PROTEIN</celldesigner:class>

```
<celldesigner:proteinReference>pr1</celldesigner:proteinReference>
</celldesigner:speciesIdentity>
</celldesigner:extension>
</annotation>
</species>
```

Publications

The methodologies and results presented in Chapters 3, 4, 5, 7 and 8 have been published in the peer-reviewed journals and the articles are included below:

- **Satagopam VP**, Theodoropoulou MC, Stampolakis CK, Pavlopoulos GA, Papandreou NC, Bagos PG, Schneider R, Hamodrakas SJ. GPCRs, G-proteins, effectors and their interactions: human-gpDB, a database employing visualization tools and data integration techniques. *Database (Oxford).* 2010 Aug 5;2010:baq019. doi: 10.1093/database/baq019.

- **Satagopam V**, Gu W, Eifes S, Gawron P, Ostaszewski M, Gebel S, Barbosa-Silva A, Balling R, Schneider R. Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases. *Big Data.* 2016 Jun;4(2):97-108. doi: 10.1089/big.2015.0057.

- Davey NE*, **Satagopam VP***, Santiago-Mozos S*, Villacorta-Martin C, Bharat TA, Schneider R, Briggs JA. The HIV mutation browser: a resource for human immunodeficiency virus mutagenesis and polymorphism data. *PLoS Comput Biol.* 2014 Dec 4;10(12):e1003951. doi: 10.1371/journal.pcbi.1003951 (*Authors contributed equally).

- Marji J, O'Donoghue SI, McClintock D, **Satagopam VP**, Schneider R, Ratner D, Worman HJ, Gordon LB, Djabali K. Defective lamin A-Rb signaling in Hutchinson-Gilford Progeria Syndrome and reversal by farnesyltransferase inhibition. *PLoS One.* 2010 Jun 15;5(6):e11132. doi: 10.1371/journal.pone.0011132.

- Jimenez-Sanchez M, Lam W, Hannus M, Snnichsen B, Imarisio S, Fleming A, Tarditi A, Menzies F, Dami TE, Xu C, Gonzalez-Couto E, Lazzeroni G, Heitz F, Diamanti D, Massai L, **Satagopam VP**, Marconi G, Caramelli C, Nencini A, Andreini M, Sardone GL, Caradonna NP, Porcari V, Scali C, Schneider R, Pollio G, O'Kane CJ, Caricasole A, Rubinsztein DC. siRNA screen identifies QPCT as a druggable target for Huntington's disease. *Nat Chem Biol.* 2015 May;11(5):347-354. doi: 10.1038/nchembio.1790. Epub 2015 Apr 6.

During the duration of the work described here I have also co-authored the following articles. These are the applications of methods and tools developed in this work and are cited in this dissertation:

- Chowdhury A, **Satagopam VP**, Manukyan L, Artemenko KA, Fung YM, Schneider R, Bergquist J, Bergsten P. Signaling in insulin-secreting MIN6 pseudoislets and monolayer cells. *J Proteome Res.* 2013 Dec 6;12(12):5954-62.

- Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, Crespo I, Perumal TM, Jurkowski W, Antony PM, Diederich N, Buttini M, Kodama A, **Satagopam VP**, Eifes S, Del Sol A, Schneider R, Kitano H, Balling R. Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol Neurobiol.* 2014 Feb;49(1):88-102. doi: 10.1007/s12035-013-8489-4.

- Neumann B, Walter T, Hrich JK, Bulkescher J, Erfle H, Conrad C, Rogers P, Poser I, Held M, Liebel U, Cetin C, Sieckmann F, Pau G, Kabbe R, Wnsche A, **Satagopam V**, Schmitz MH, Chapuis C, Gerlich DW, Schneider R, Eils R, Huber W, Peters JM, Hyman AA, Durbin R, Pepperkok R, Ellenberg J. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature.* 2010 Apr 1;464(7289):721-7. doi: 10.1038/nature08869.

- Bali KK, Venkataramani V, **Satagopam VP**, Gupta P, Schneider R, Kuner R. Transcriptional mechanisms underlying sensitization of peripheral sen-

sory neurons by granulocyte-/granulocyte-macrophage colony stimulating factors. *Mol Pain.* 2013 Sep 25;9:48. doi: 10.1186/1744-8069-9-48.

- Bali KK, Selvaraj D, **Satagopam VP**, Lu J, Schneider R, Kuner R. Genome-wide identification and functional analyses of microRNA signatures associated with cancer pain. *EMBO Mol Med.* 2013 Nov;5(11):1740-58.

- Roomp K, Kristinsson H, Schvartz D, Ubhayasekera K, Sargsyan E, Manukyan L, Chowdhury A, Manell H, **Satagopam V**, Groebe K, Schneider R, Bergquist J, Sanchez JC, Bergsten P. Combined lipidomic and proteomic analysis of isolated human islets exposed to palmitate reveals time-dependent changes in insulin secretion and lipid metabolism. PLoS One. 2017 Apr 27;12(4):e0176391. doi: 10.1371/journal.pone.0176391. eCollection 2017.

OPEN ACCESS Freely available online

PLoS one

Defective Lamin A-Rb Signaling in Hutchinson-Gilford Progeria Syndrome and Reversal by Farnesyltransferase Inhibition

Jackleen Marji[1], Seán I. O'Donoghue[2], Dayle McClintock[1], Venkata P. Satagopam[2], Reinhard Schneider[2], Desiree Ratner[1], Howard J. Worman[3], Leslie B. Gordon[4], Karima Djabali[1,5]*

1 Department of Dermatology, College of Physicians and Surgeons, Columbia University, New York, New York, United States of America, **2** EMBL, Heidelberg, Germany, **3** Departments of Medicine and of Pathology and Cell Biology, College of Physicians and Surgeons, Columbia University, New York, New York, United States of America, **4** Department of Pediatrics, Warren Albert Medical School of Brown University, Providence, Rhode Island, United States of America, **5** Department of Dermatology, Technical University Munich, Munich, Germany

Abstract

Hutchinson-Gilford Progeria Syndrome (HGPS) is a rare premature aging disorder caused by a *de novo* heterozygous point mutation G608G (GGC>GGT) within exon 11 of *LMNA* gene encoding A-type nuclear lamins. This mutation elicits an internal deletion of 50 amino acids in the carboxyl-terminus of prelamin A. The truncated protein, progerin, retains a farnesylated cysteine at its carboxyl terminus, a modification involved in HGPS pathogenesis. Inhibition of protein farnesylation has been shown to improve abnormal nuclear morphology and phenotype in cellular and animal models of HGPS. We analyzed global gene expression changes in fibroblasts from human subjects with HGPS and found that a lamin A-Rb signaling network is a major defective regulatory axis. Treatment of fibroblasts with a protein farnesyltransferase inhibitor reversed the gene expression defects. Our study identifies Rb as a key factor in HGPS pathogenesis and suggests that its modulation could ameliorate premature aging and possibly complications of physiological aging.

Citation: Marji J, O'Donoghue SI, McClintock D, Satagopam VP, Schneider R, et al. (2010) Defective Lamin A-Rb Signaling in Hutchinson-Gilford Progeria Syndrome and Reversal by Farnesyltransferase Inhibition. PLoS ONE 5(6): e11132. doi:10.1371/journal.pone.0011132

Editor: Mikhail V. Blagosklonny, Roswell Park Cancer Institute, United States of America

Received February 12, 2010; **Accepted** May 20, 2010; **Published** June 15, 2010

Funding: This work was supported in part by the National Institutes of Health (R01AG025240) to H.J.W., and by the Irving Foundation and the National Institutes of Health (Grants K01AR048594 and R01AG025302) to K.D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kd206@columbia.edu

Introduction

Hutchinson–Gilford progeria syndrome (HGPS) is a rare, sporadic genetic disorder with phenotypic features of premature aging [1][2,3,4]. It is caused by *de novo* dominant mutations in *LMNA* [5,6,7]. *LMNA* encodes A-type nuclear lamins, with the predominant somatic cell isoforms lamin A and lamin C arising by alternative RNA splicing [8]. Lamins are intermediate filament proteins that polymerize to form the nuclear lamina, a meshwork associated with the inner nuclear membrane. HGPS is one of a spectrum of diverse diseases, sometimes referred to as "laminopathies," caused by mutations in *LMNA* [9].

Lamin A is synthesized as a precursor, prelamin A, which has a CaaX motif at its carboxyl terminus. The CaaX motif signals a series of catalytic reactions resulting in a carboxyl-terminal cysteine that is farnesylated and carboxymethylated [9]. Farnesylated, carboxymethylated prelamin A is normally cleaved near its carboxyl-terminus in a reaction catalyzed by ZMPSTE24 endoprotease, leading to removal of the farnesylated cysteine [9]. The *LMNA* G608G mutation responsible for the majority of cases of HGPS creates an abnormal splice donor site within exon 11, generating an mRNA that encodes a prelamin A with a 50 amino acid deletion at its carboxyl-terminal domain [5,6]. The ZMPSTE24 endoproteolytic site is deleted from progerin and hence retains a farnesylated and carboxymethylated cysteine at its carboxyl terminus [9]. Expression of progerin induces severe abnormalities in nuclear morphology, heterochromatin organization, mitosis, DNA replication and DNA repair [5,6,10,11,12,13,14,15]. Progerin toxicity is attributed at least in part to its farnesyl moiety, as chemical inhibitors of protein farnesyltransferase (FTIs) reverse abnormalities in nuclear morphology in progerin expressing cells [16,17,18,19,20]. In addition, FTIs and other chemical inhibitors of protein prenylation partially reverse progeria-like phenotypes in genetically modified mice that express progerin or lack ZMPSTE24, and therefore accumulate unprocessed, farnesylated prelamin A [21,22,23,24].

While several studies have clearly implicated farnesylated progerin in HGPS, the precise molecular mechanisms of how it induces HGPS pathology remain to be understood. Initial gene expression profiling of fibroblasts from human subjects with progeria syndromes and transfected cell models identified changes in sets of genes implicated in diverse pathways that have not always been consistent and have not been shown to be reversed by interventions such as treatment with FTIs [25,26,27,28]. Therefore, we carried out additional genome-wide expression studies in cells from children with HGPS to identify alterations in functional groups of genes that define defective signaling pathways and to determine if FTI treatment reverses these defects. Our results demonstrate a link between progerin

and the retinoblastoma protein (Rb) signaling pathway in HGPS.

Results

Lamin A-Rb signaling network is implicated in HGPS pathophysiology

To determine the mechanisms by which progerin exerts its pathological effect, we performed parallel microarray analyses of fibroblasts from subjects with HGPS and control individuals that were treated or untreated with the FTI lonafarnib for three days. We used RNA isolated from fibroblasts from five subjects with HGPS and five unaffected individuals to hybridize Affymetrix U133 plus 2.0 arrays. We identified 50,636 probe sets (Fig. 1A) and analyzed the data as described in Materials and Methods.

We first focused on the different gene expression profiles in fibroblasts from controls and subjects with HGPS that were not treated with FTI. We found that 352 genes were significantly differentially expressed between fibroblasts from subjects with HGPS and controls. Of those genes, 306 were downregulated and 46 upregulated in fibroblasts from subjects with HGPS. The assigned subcellular localizations indicated that at least 31.6% of the gene products were localized to the nucleus, 30.4% to the cytoplasm, 23.7% to the plasma membrane and 1.8% to the extracellular matrix, with the remainder unknown (Fig. 1B). Molecular function analyses (Fig. 1C) and physiological distributions (Fig. 1D) indicated that a significant number of genes were implicated in lipid metabolism, cell growth and differentiation, cell cycle, DNA replication and repair as well as cardiovascular system development.

Of the 352 genes differentially expressed in cells from subjects with HGPS, 280 genes had known interactions according to MetaCore database (www.genego.com). To build networks integrating these genes, we added *LMNA* because, although its levels of expression remained unchanged, mutations in this gene, which result in abnormal protein expression, are the cause of HGPS (Fig. 2). Of the genes with altered expression in HGPS, the MetaCore method identified *Rb1* as the only one encoding a protein product, Rb, known to interact directly with A-type lamins [29,30]. The expression of *Rb1* was downregulated in HGPS. Based on differential expression of other genes and known direct protein interactions and relationships, the MetaCore analysis identified additional downstream factors in a signaling network that were altered in cells from subjects with HGPS (Fig. 2). This signaling network started with lamin A/C (layer 1), which is altered by the *LMNA* mutation leading to progerin expression, and impacted Rb (layer 2). The most direct downstream targets of Rb1 were p107, NCOA2, SP1 and ATF2 (layer 3). We added E2F1 to this network, a direct Rb interacting partner although its expression level was not changed. These five factors then had downstream direct effects on 15 other transcription regulators (layer 4). Of these 15 genes, the expression levels of 13 were significantly downregulated in fibroblasts from subjects with HGPS with the exception of ANCO-1, which was upregulated. HNF4-alpha and c-Myc were added to the network even though their levels remained unchanged to allow the integration of the remaining downstream target genes. In the next downstream layer (layer 5), more nuclear factors were connected and all these genes had significantly lower levels of expression in fibroblasts from subjects with HGPS. From the 15 nuclear factors in layer 4, the network divided into several sub-networks, which included the JAK pathway implicated in the regulation of cell proliferation, differentiation and survival, a group of genes related to motility and cytoskeletal organization, and another group of genes implicated in DNA replication and repair (Fig. 2, and Figures S1, S2, S3, S4).

The differential expression of most of the identified genes in HGPS can potentially be explained by the hypothesis that an abnormal prelamin A variant causes Rb to differentially interact with or regulate downstream partners. Most of these genes have at least one direct connection to an upstream nuclear factor (Fig. 2, right panel). Importantly, the differentially expressed genes in HGPS indicated that Rb is a key regulatory component affected by *LMNA* mutation and that it is at the center of a signaling network that is abnormally active in the disease.

The lamin A-Rb signaling network is modulated by FTI treatment

To unravel the mechanism underlying reversal of the HGPS phenotype by blocking protein farnesylation, we determined the gene expression changes occurring in fibroblasts after FTI administration. Fibroblasts from normal subjects and subjects with HGPS were grown in medium supplemented daily with 1.5 μM of lonafarnib for three days. Trypan blue exclusion assays indicated that less than 2% of cells died whether or not they were treated with FTI (data not shown). We monitored the inhibition of protein farnesyltransferase by screening for unprocessed HDJ-2 and prelamin A. FTI-treated control cells exhibited an average of $64.0\% \pm 0.7$ (mean ± SD., $n = 3$, $P < 0.05$) non-farnesylated HDJ-2 and accumulation of prelamin A (Fig. 3A). FTI-treated fibroblasts from subjects with HGPS exhibited $62.3\% \pm 1.5$ ($n = 3$, $P < 0.05$) non-farnesylated HDJ-2 with a similar increase in prelamin A (Fig. 3A), indicating that protein farnesyltransferase activity was similarly inhibited in cells from control individuals and those with HGPS.

We first examined the effects of FTI treatment on normal fibroblasts by using Ingenuity Pathways Analysis (IPA) to compare gene expression profiles in fibroblasts from control individuals with or without FTI treatment. This analysis identified significant differences in the expression of 65 genes, with 47 downregulated and 18 upregulated in cells treated FTI. The majority of these genes were functionally assigned to cell cycle, DNA replication and repair and purine and pyrimidine metabolism; 63% of the encoded proteins were nuclear (Figure S5).

The 65 differentially expressed genes were assembled into a network based on known protein interactions in the MetaCore database. The changes in FTI-treated control cells can be explained by a direct effect of inhibition of protein farnesyltransferase (Fig. 3B). Inhibition of protein farnesyltransferase would subsequently affect three of its substrates, centromeric proteins F and E (CENP-F and CENP-E), the genes of which had altered expression in FTI-treated cells, and lamin A/C. Even though lamin A/C levels remained unchanged it was added to the network because FTIs induce the accumulation of unfarnesylated, unprocessed prelamin A. CENP-E interacts with TPX2 and CENP-F interacts directly with Rb. These proteins therefore were added to the network even though their levels remained unchanged in FTI-treated cells. Rb directly interacts with lamin A/C and with E2F1, repressing its transcriptional activity. Expression of all of the downstream targets of E2F1 was downregulated in the FTI-treated cells (Fig. 3B). These results indicate that FTI treatment affects a sequential cascade of events starting from inhibition of protein farnesyltranferase, which in turn modifies centromeric proteins and lamin A/C and possibly their interactions with Rb. Alterations in Rb function in turn influence the activity of the transcription factor E2F1, a key regulator of downstream genes with decreased expression in fibroblasts after FTI treatment.

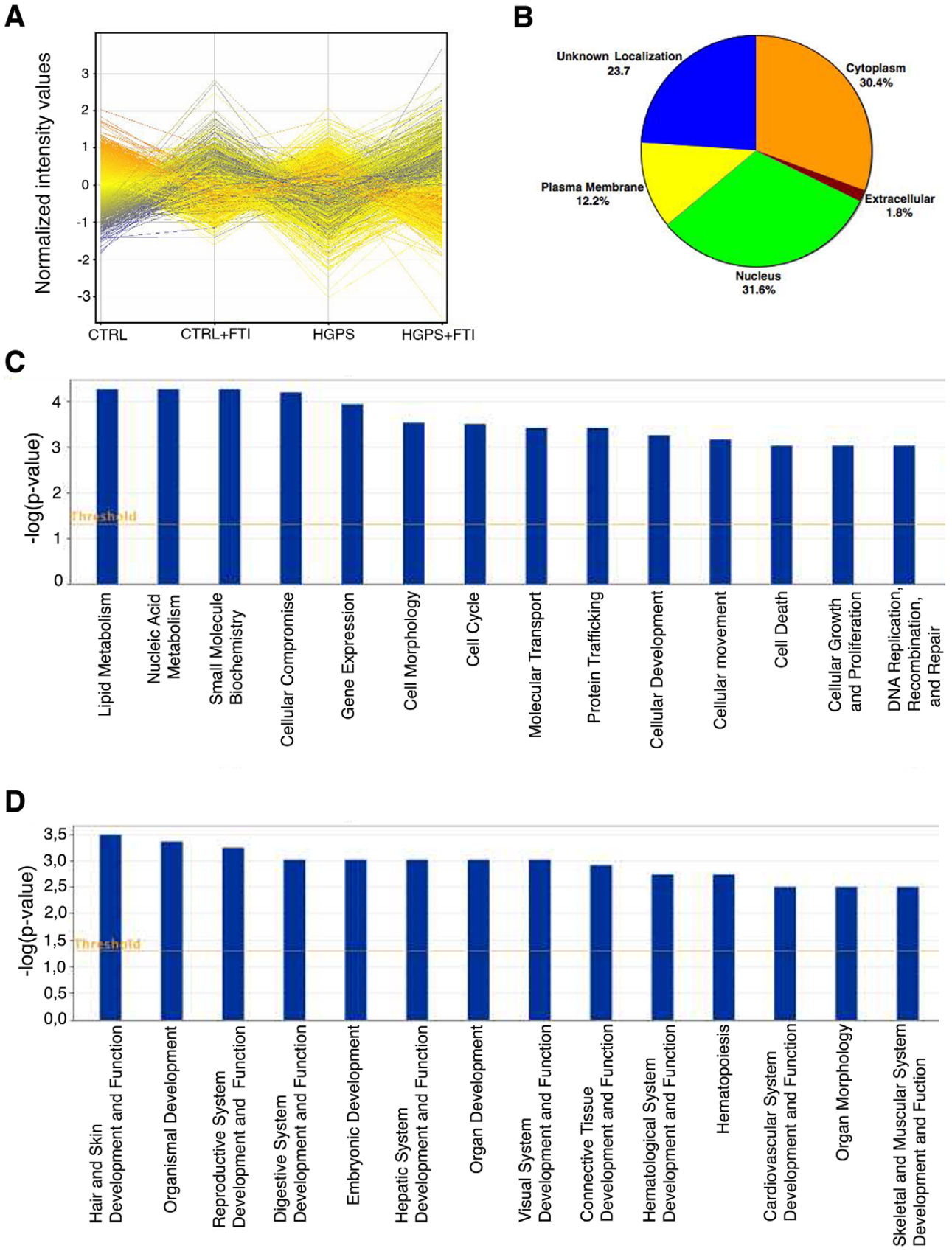

A
Normalized intensity values
3
2
1
0
-1
-2
-3
CTRL
CTRL+FTI
HGPS
HGPS+FTI
B
Unknown Localization
23.7
Cytoplasm
30.4%
Plasma Membrane
12.2%
Extracellular
1.8%
Nucleus
31.6%
C
-log(p-value)
4
3
2
1
0
Threshold
Lipid Metabolism
Nucleic Acid Metabolism
Small Molecule Biochemistry
Cellular Compromise
Gene Expression
Cell Morphology
Cell Cycle
Molecular Transport
Protein Trafficking
Cellular Development
Cellular movement
Cell Death
Cellular Growth and Proliferation
DNA Replication, Recombination, and Repair
D
-log(p-value)
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
Threshold
Hair and Skin Development and Function
Organismal Development
Reproductive System Development and Function
Digestive System Development and Function
Embryonic Development
Hepatic System Development and Function
Organ Development
Visual System Development and Function
Connective Tissue Development and Function
Hematological System Development and Function
Hematopoiesis
Cardiovascular System Development and Function
Organ Morphology
Skeletal and Muscular System Development and Fuction

Figure 1. Genome-wide expression profiling of HGPS and control fibroblast cultures. (A) Microarray plot profiles indicate changes in gene expression in control, HGPS, FTI-treated control and FTI-treated HGPS fibroblasts. Each continuous line corresponds to the normalized intensity value of an individual probe set. Line colors denote the intensity of the signal (red: strong and blue: low signal). Probes that satisfied a greater or less than two-fold cutoff and statistically significant difference of $p < 0.01$ are displayed. (B) Pie chart indicates the predicted subcellular localization of proteins encoded by the 352 genes differentially expressed in HGPS. The list of differentially expressed genes in HGPS versus control cells was analyzed using Ingenuity Pathway Analysis (IPA) and encoded proteins assigned a subcellular localization based on information contained in the Ingenuity Knowledge Base. (C) Genes differently expressed in HGPS (352 genes) were assigned to diverse cellular functions using the "Functional Analysis" tool of IPA software (www.ingenuity.com). Columns represent groups of genes associated with specific cellular functions (*x*-axis). The significant genes were compared to IPA database and ranked according to a p-values generated with Fisher extract test. P-values less than 0.05 indicates a statistically significant, non-random association between a set of significant genes and a set of all genes related to a given function in IPA database. The ratio (*y*-axis) represents the number of genes from the dataset that map to the pathway divided by the number of all known genes ascribed to the pathway. The yellow line represents the threshold of $p<0.05$. (D) As described above, the 352 genes were assigned to diverse physiological systems according to IPA.
doi:10.1371/journal.pone.0011132.g001

FTI restores a nearly normal gene expression profile in fibroblasts from subjects with HGPS

To determine the effects of a FTI on gene expression in cells from subjects with HGPS, we generated a Venn diagram to demonstrate overlapping alterations in expression in three datasets (Fig. 4A). The three datasets were: (1) genes differentially expressed in fibroblasts from controls subjects after FTI treatment compared to no treatment (65 genes); (2) genes differentially expressed in fibroblasts from subjects with HGPS compared to those from control subjects (352 genes); and (3) genes differentially expressed in fibroblasts from subjects with HGPS after FTI treatment compared to those from untreated control subjects (804 genes). Differential expression of 25 genes were specific to FTI-treated control fibroblasts (dataset 1); 40 genes were commonly differentially expressed in these cells as well as in cells from subjects with HGPS treated with FTI (dataset 1 compared to dataset 2). Only one gene, *SMC2*, a structural maintenance factor of chromosome 2 was common between the three datasets.

We focused on the gene expression profile resulting from FTI treatment of fibroblasts from subjects with HGPS (dataset 3), which indicated changes in the expression of 804 genes after FTI treatment. Of these 804 genes, 414 were upregulated and 390 were downregulated. This high number of genes with altered expression suggests that cells from subjects with HGPS were highly sensitive to FTI. For the gene products, 31.9% were predicted to be nuclear proteins (Fig. 4B), whereas in normal fibroblasts treated with FTI 63% of the differentially expressed genes were predicted to encode nuclear proteins (Figure S5). Of these 804 genes showing differential expression after FTI treatment of cells from subjects with HGPS, 64 were differentially expressed in untreated cells from subjects with HGPS compared to normal fibroblasts. Expression of the remaining 701 differentially expressed genes were uniquely altered in FTI-treated fibroblasts from subjects with HGPS and were functionally associated with gene expression, cell cycle, cell growth, DNA replication and repair, molecular transport and protein trafficking (Fig. 4C). They also were involved in several canonical pathways involved in PI3K/AKT signaling, protein ubiquitination and regulation of the actin cytoskeletal network (Fig. 4D).

Only 64 genes were differentially expressed in both the datasets comparing fibroblasts from control to HGPS (dataset 2) and comparing fibroblasts from control to HGPS treated with FTI (dataset 3). Hence, abnormal expression of 288 of the 352 genes in fibroblasts from subjects with HGPS was normalized after FTI-treatment.

In addition, an additional t-test analysis was done directly comparing fibroblasts from control individuals treated with FTI to fibroblasts from subjects with HGPS treated with FTI. This analysis indicated that only 1 gene *PIK3CB* (+2.166) was altered in FTI-treated HGPS cells (Fig 4E). *PIK3CB* is a phosphoinoitide-3-kinase that is functionally associated to apoptosis, survival, migration, proliferation, cell cycle and adhesion. The IPA analysis links *PIK3CB* with one network associated to lipid metabolism. Therefore, the difference in expression of only one gene indicated that 99% of the genes in FTI-treated cells from subjects with HGPS were expressed at the same levels as in FTI-treated normal fibroblasts.

Validation of the microarray analyses by real-time RT-PCR

Differential expression of selected genes found to be significantly different in the microarray analyses was confirmed by real-time RT-PCR. To compare fibroblasts from subjects with HGPS to unaffected controls, we selected genes identified in the microarray analysis that were upstream regulators in the signaling pathway. For normal fibroblasts with and without FTI treatment, we selected genes that were upstream regulators in the signaling network. For fibroblasts from subjects with HGPS with and without FTI treatment, we selected genes that showed the greatest fold changes in expression. We also measured the expression of PIK3CB in FTI-treated fibroblasts from subjects with HGPS and unaffected controls, as this was the only gene identified by microarray analyses to be differentially expressed between these experimental groups. In all cases, gene expression differences measured by real-time RT-PCR were consistent with those measured in the microarray analyses (Fig. 5A, B, C and Figure S6).

Potential relationship between gene expression alterations in HGPS and physiological aging

Our results implicate a defective lamin A-Rb signaling network as a pathogenic mechanism in HGPS, a disease with features of premature aging. We investigated whether alterations in this same signaling network might occur during physiological aging. We screened mRNA extracted from human skin biopsies from individuals 38 to 90 years of age. The skin samples were derived from an already established skin biopsy tissue bank (approved by the Columbia University Medical Center Institutional Review Board), which has been described previously [31]. Total mRNA extractions from human skin samples were performed as described previously [31]. Remarkably, lamin A/C transcripts were increased in skin from individuals 87 to 90 years old (Fig. 5D). In contrast, Rb mRNA was already decreased in skin from 70 to 72-year-old subjects (Fig. 5D), indicating that changes in *LMNA* and *RB1* expression occur during physiological aging. Furthermore, Western blot analyses of proteins extracted from skin showed that A-type lamins were increased in samples from middle aged and elderly individuals (Fig. 6A, B). Low levels of progerin were also detected by indirect immunofluorescence microscopy using a specific antibody in a restricted number of dermal fibroblasts in skin sections derived from elderly subjects (Fig. 6C). However, progerin expression was too low to detect by Western blot analyses of proteins extracted from skin as previously reported [31]. Rb, which is expressed at low levels in normal skin, was

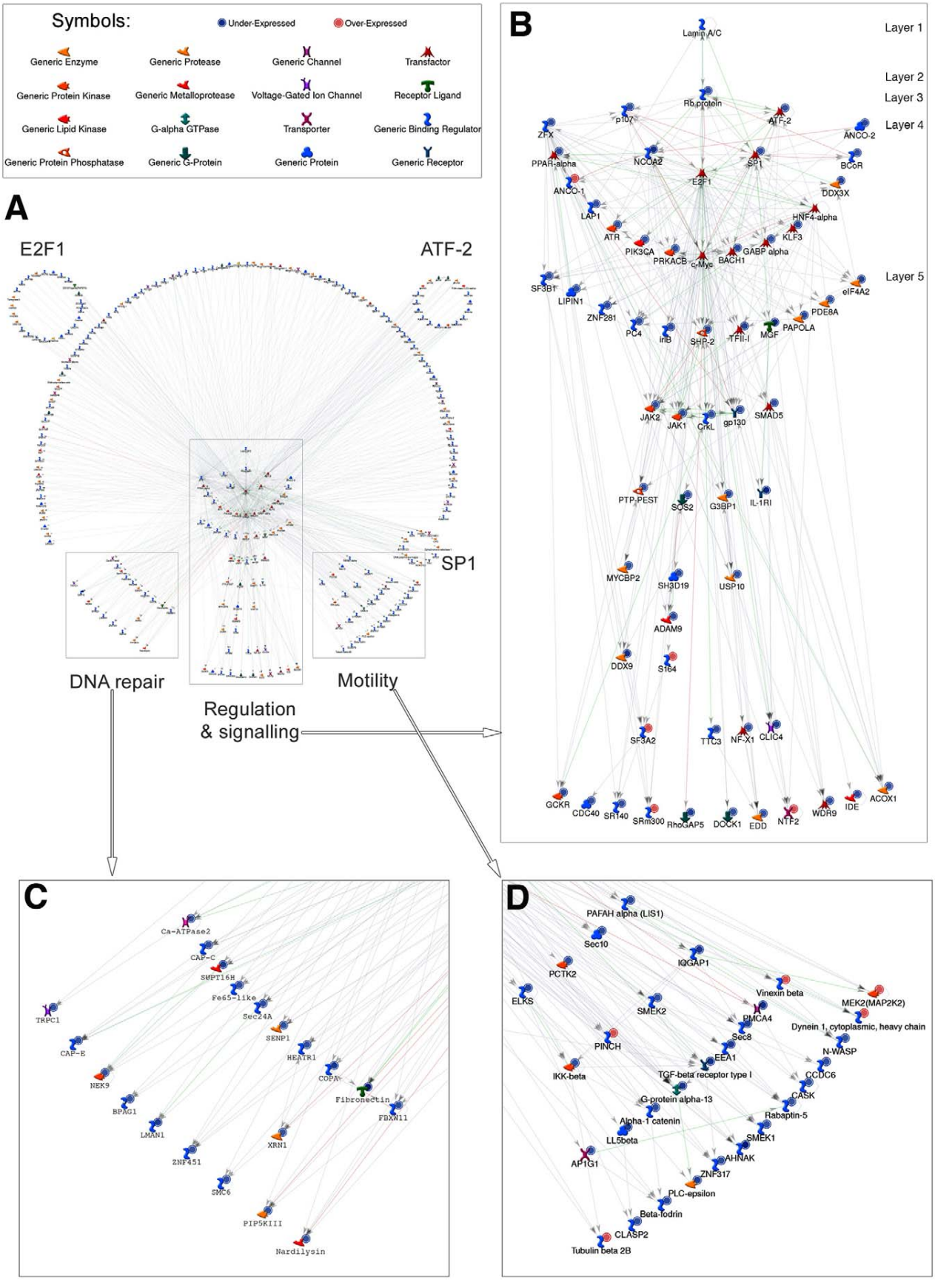

Symbols:
Under-Expressed
Over-Expressed
Generic Enzyme
Generic Protease
Generic Channel
Transfactor
Generic Protein Kinase
Generic Metalloprotease
Voltage-Gated Ion Channel
Receptor Ligand
Generic Lipid Kinase
G-alpha GTPase
Transporter
Generic Binding Regulator
Generic Protein Phosphatase
Generic G-Protein
Generic Protein
Generic Receptor
A
E2F1
ATF-2
SP1
DNA repair
Regulation & signalling
Motility
B
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Lamin A/C
Rb protein
p107
ATF-2
ZFX
ANCO-2
PPAR-alpha
NCOA2
SP1
BCoR
E2F1
ANCO-1
DDX3X
LAP1
HNF4-alpha
ATR
KLF3
PIK3CA
PRKACB
BACH1
GABP alpha
c-Myc
SF3B1
eIF4A2
LIPIN1
ZNF281
PDE8A
PC4
IntB
SHP-2
TFII-I
MGF
PAPOLA
JAK2
JAK1
CrkL
gp130
SMAD5
PTP-PEST
SOS2
G3BP1
IL-1RI
MYCBP2
SH3D19
USP10
ADAM9
DDX9
S164
SF3A2
TTC3
NF-X1
CLIC4
GCKR
CDC40
SR140
SRm300
RhoGAP5
DOCK1
EDD
NTF2
WDR9
IDE
ACOX1
C
Ca-ATPase2
CAP-C
SMPT16H
Fe65-like
Sec24A
TRPC1
SENP1
CAP-E
HEATR1
NEK9
COPA
Fibronectin
BPAG1
FBXW11
LMAN1
XRN1
ZNF451
SMC6
PIP5KIII
Nardilysin
D
PAFAH alpha (LIS1)
Sec10
IQGAP1
PCTK2
Vinexin beta
ELKS
SMEK2
MEK2(MAP2K2)
PMCA4
Dynein 1, cytoplasmic, heavy chain
Sec8
PINCH
N-WASP
EEA1
IKK-beta
TGF-beta receptor type I
CCDC6
G-protein alpha-13
CASK
Alpha-1 catenin
Rabaptin-5
LL5beta
SMEK1
AP1G1
AHNAK
ZNF317
PLC-epsilon
Beta-fodrin
CLASP2
Tubulin beta 2B

Figure 2. The lamin A-Rb network. The main network (center left) shows downstream interactions between lamin A/C and the 352 differentially expressed genes in HGPS fibroblasts. The network was built using MetaCore analyses, which finds known interactions between gene products. (A) The network divides into several distinct regions, the main region being the regulatory and signaling network. (B) The detailed view of the main region (top right) shows that the only gene differentially expressed in HGPS directly downstream of lamin A/C (layer 1) is that encoding Rb (layer 2). In turn, most of the immediate downstream partners of Rb included in this dataset are p107, NCOA2, SP1, and ATF-2 and E2F1 (layer 3). Of the remaining genes that occur downstream of layer 3, nearly half of the 280 genes interact with only one or more of these transfactors associated to the main network (center left); these gene products are placed above the centre. From layer 4 and 5, most of the genes can be connected at least to one entity according to GeneGO. In the HGPS dataset, 124 genes that had no known interactions in GeneGO are not shown. From the center regulatory and signaling network (zoom left, panel B)), several groups of genes segregate into six subnetworks, based on mutual interactions: a group denoted DNA repair (bottom left, panel (C)), motility (bottom right, panel (D)), and three circles of genes regulated by E2F1, ATF-2 and SP1, respectively. Symbols associating the genes with functions are indicated. Genes labeled with blue circles were downregulated and red circles were upregulated. Higher magnifications of panels A to D are provided in Figures S1, S2, S3 and S4.
doi:10.1371/journal.pone.0011132.g002

weakly detected using immunofluorescence microscopy in samples from young individuals and barely detected in samples from older subjects (Fig. 6A left panel). These findings indicate that lamin A and Rb levels change during physiological aging and suggest that defects in a lamin A-Rb signaling pathway might occur similarly to what we have identified in cells from subjects with HGPS.

Discussion

Inhibition of protein farnesyltransferase activity by FTIs has been shown to improve abnormalities in nuclear morphology in cells from subjects with HGPS and from animal models of the disease [15,16,17,18,19,20,32]. FTIs also improve the progeria-like phenotypes of mouse models of HGPS [22,23]. Presumably, blocking farnesylation of progerin, the truncated prelamin A expressed as a result of the G608G *LMNA* mutation which is responsible for most cases of HGPS, renders it less toxic. However, the molecular mechanisms by which progerin exerts its toxicity and whether FTI treatment reverses specific molecular defects induced by progerin remain largely unknown. We have now identified a defective lamin A-Rb signaling network in cells from subjects with HGPS and demonstrated that treatment of cells with an FTI reverses abnormalities in the expression of genes encoding proteins in this network.

We compared our gene expression profiles to those of previous studies that have examined gene expression in HGPS and found very little overlap (3.5% overlap of differentially expressed genes with Scaffidi and Mitseli [28] and less than 1% with Csoka *et al.* [27] (Table S1). Csoka *et al.*[27] used fibroblasts from only three subjects ages 8 to 14 years of age whereas we used cultured fibroblasts from five subjects with HGPS obtained at earlier ages (age 1 to 4 years). As the growth rate and proliferation potency of fibroblasts from subjects with HGPS decrease with cellular age *in vitro* and the donor's age, this could partially explain the differences between our results. Importantly, the average age of death in HGPS is 12–15 years, which suggests that the cellular phenotype may be severely altered in older children [4]. Scaffidi and Misteli [28] generated immortalized normal fibroblasts that overexpressed progerin and examined gene expression in only two lines, making a comparison difficult given the very different system and the low statistical power from using only two biological replicates. Separately, Fong et al. [33] used RT-PCR to examine expression of several genes selected from these previous studies in normal fibroblasts transfected with antisense oligonucleotides that increase *LMNA* alternative RNA splicing to generate progerin. We found that four of seven genes they tested were similarly differentially expressed in the fibroblasts we used from subjects with HGPS (Table S2). An earlier study by Ly *et al.* [25] used fibroblasts from three donors with HGPS of age 8 to 9 years and found 61 genes differently expressed in young subjects compared to old subjects and with HGPS subjects. Of these 61 genes, 49 were differentially expressed in HGPS versus young individuals [25]. Comparison of these 49 genes against the 352 genes differentially expressed in HGPS identified in our study indicated an overlap of 22 genes. An additional study from Park *et al* [26] compared gene expression in fibroblasts from only one subject with HGPS to that of cultured fibroblasts in replicative senescence; no statistical criteria were applied. Although HGPS is an extremely rare disease, to the best of our knowledge we have used the most biological replicates so far in any study of genome-wide gene expression in actual patient material. We also validated our microarray data using information from the literature, by screening databases and by performing RT-PCR for selected genes. We confirmed by RT-PCR that Rb mRNA levels were decreased in HGPS cells in comparison the control cells. No change in Rb gene expression level was reported in previous microarray analyses on HGPS cells or HGPS cell models [26,27,28].

Most importantly, rather than providing a descriptive list of genes with altered expression levels, our methods identified a novel signaling network starting with the lamin A-Rb interaction that is altered in HGPS.

The potential link between progerin and Rb in the orchestration of cell cycle and proliferation defects was previously suggested by studies demonstrating that lamin A and C interact with Rb [30,34]. Rb is a tumor suppressor and major cell cycle regulator that, in its hypophosphorylated state, binds to and inhibits the E2 factor (E2F) family of transcription factors required for cell cycle progression [36]. Upon hyperphosphorylation of Rb by cyclin/cyclin-dependent kinase complexes, E2F is released to initiate S phase. A role for lamins in this process is further suggested by the finding that hypophosphorylated Rb is tightly associated with lamin A/C-enriched nucleoskeletal preparations of early G1 cells [29]. In HGPS cells there is evidence for a significant reduction in hyperphosphorylated Rb in HGPS fibroblasts (Dechat et al. 2007). In our experiments, we found a decreased level of Rb expression at the mRNA and protein levels in HGPS (see Figure 2 and Figure 6).

Previously, a small fraction of lamin A/C was shown to colocalize with Rb and E2F1 in nuclear foci during G1-phase [37]. In HGPS cells, we could not detect lamin A foci. Whether Rb foci remain associated with E2F1 remains to be further investigated. In HGPS cells, our microarray analysis indicated a decrease in Rb levels without changes in E2F1 levels. We further investigated whether E2F1 levels might be altered by indirect immunofluorescence microscope and observed no changes in the nuclear intensity of the E2F1 signal between HGPS and normal fibroblast cells (data not shown). Previously, the localization of Rb and E2F1 in cells from $Lmna^{-/-}$ mice indicated that while Rb levels were reduced, a portion of Rb remained localized to nuclear foci but displayed reduced overlap with the E2F1 signal [35]. The expression levels of E2F1 *in* $Lmna^{-/-}$ cells in comparison to wild type cells has not yet been reported. However, the lack of association between Rb and E2F1 in nuclear foci in the absence of lamin A/C could indicate an alteration in Rb control of E2F1

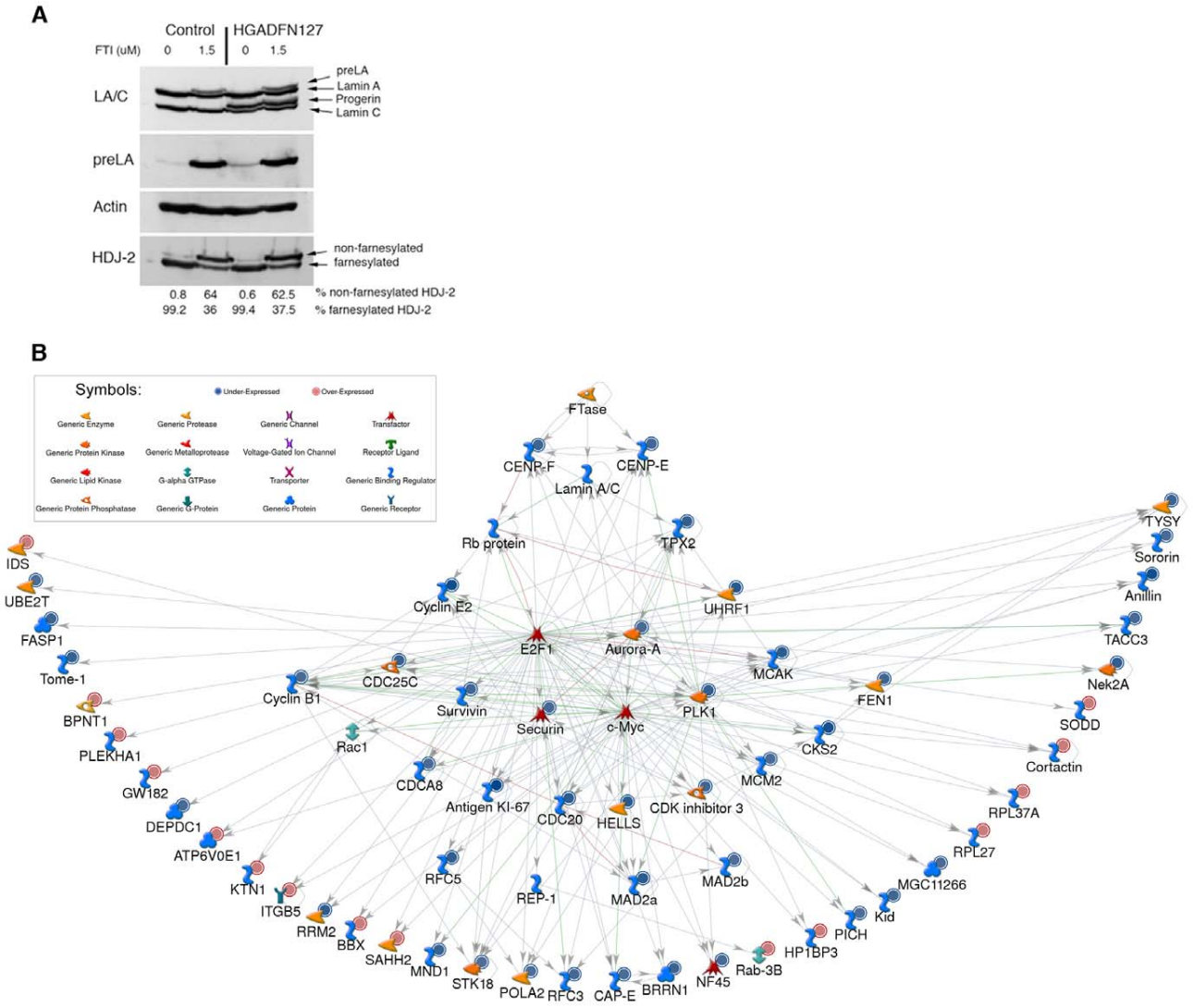


Figure 3. FTI inhibition of prelamin A and progerin farnesylation and FTI-induced gene expression changes in fibroblasts from normal subjects. (A) Western blot analysis of protein extracts of fibroblasts from individuals with HGPS and controls that were treated or untreated with FTI (1.5 μM lonafarnib daily for three days). Blots were probed with anti-lamin A/C (LA/C), anti-prelamin A (preLA), anti-actin, and anti HDJ-2 antibodies. Increase in levels of prelamin A and non-farnesylated HDJ-2 with FTI treatment are indicated. (B) 65 genes were differentially expressed in FTI-treated control cells. The signaling network was built upon functional association between protein farnesyltransferase (FTase), the enzymatic activity of which is inhibited by FTI, lamin A/C and Rb even though expression of all three transcripts remained unchanged with FTI treatment. Of note, lamin A is a known substrate for FTase. Downstream FTase interactions between all the 65 genes differentially expressed in FTI-treated control cells have been incorporated based on their interactions according to MetaCore analysis. The MetaCore analysis identified two transfactors, E2F1 and c-Myc, and one transfactor regulator, REP-1 that permit linking all those 65 genes into this single network. Symbols associating the genes functions are indicated. Genes labeled with blue circles were downregulated and by red circles were upregulated.
doi:10.1371/journal.pone.0011132.g003

activity [35]. Further studies will be needed to understand how expression levels of lamin A/C, Rb, and E2F1 might affect each other's expression levels, interactions and subnuclear localization. In addition to its role in cell cycle control, Rb also regulates cellular differentiation. In the absence of lamin A/C in muscle cells derived from *Lmna*$^{-/-}$ mice, reduced levels of Rb and other transcription factors regulating muscle cell differentiation was reported [38]. This study provide an additional link between lamin A/C-Rb and its potential role in muscle cell differentiation [38].

Immunohistochemical analysis of Rb distribution did not reveal any obvious alteration besides a decreased Rb signal in some nuclei of cells from subjects with HGPS (Figure S7). Rb was also detected in the most dysmorphic nuclei exhibiting a strong progerin staining, indicating the accumulation of high levels of progerin protein (Figure S7) as reported previously [15]. Based on these previous studies, decreased Rb expression or phosphorylation status in HGPS cells appears to be implicated in the deregulation of proliferation. Whether the interaction of A-type lamin with Rb is impaired in HGPS cells remains to be further addressed. However, our present study together with previously published observations support the potential implication of a defective lamin A/C-Rb signaling network in the HGPS cellular phenotype. In support of our finding, a similar potential link between an altered Lamin A/C-Rb signaling was previously reported in cells derived from Lmna$^{-/-}$ mice model [35]. In cells lacking lamin A/C the levels of Rb was decreased while its mRNA

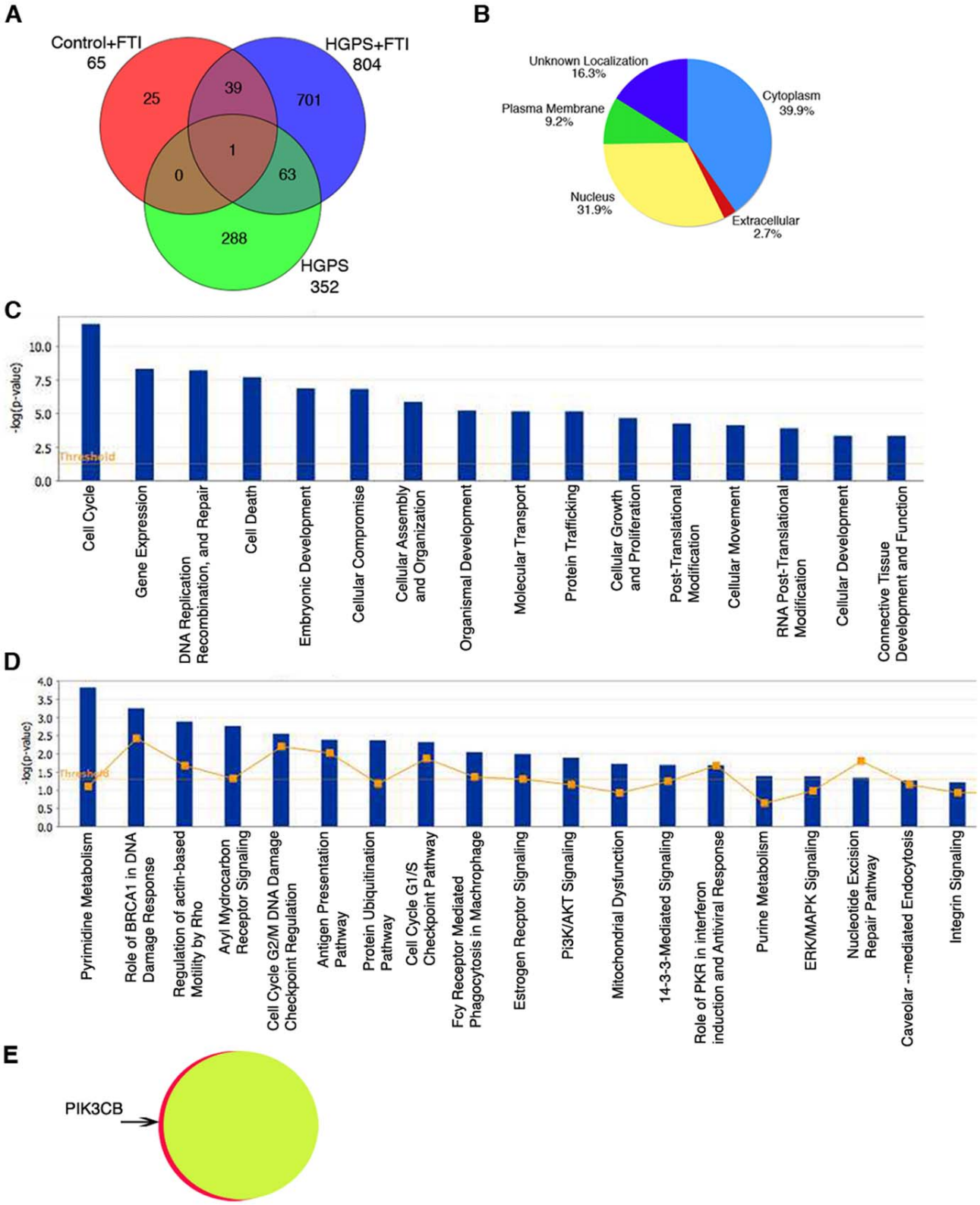


Figure 4. Genome-wide expression profiling of FTI-treated and untreated fibroblasts from subjects with HGPS. (A) Venn diagram comparison of microarray datasets of controls versus control-FTI, or HGPS and/or HGPS-FTI. 65 genes were differentially expressed in fibroblasts from control subjects after FTI treatment compared to no treatment (Control+FTI; red); 352 in fibroblasts from subjects with HGPS compared to those from control subjects (HGPS; green); and 804 in fibroblasts from subjects with HGPS after FTI treatment compared to untreated controls (HGPS+FTI; blue). Genes with altered expression common between these different datasets are shown as areas of overlap with different colors in the Venn diagram

with numbers of common genes indicated. (B) Pie chart indicates the subcellular localization of the encoded proteins of the differentially expressed genes between control and HGPS-FTI datasets according to information contained in the Ingenuity Knowledge Base. (C) Genes differentially expressed between controls versus HGPS-FTI datasets were assigned to diverse cellular functions according to IPA and (D) were associated to canonical pathways according to IPA. The significant genes were compared to IPA database and ranked according to p-values generated with Fisher extract test. P-values less than 0.005 indicate a statistically significant, non-random association between a set of significant genes and a set of all genes related to a given function in IPA database. The ratio (y-axis) represents the number of all known genes ascribed to the pathway. The Yellow line represents the threshold of $p<0.05$. (E) Comparison of gene expression alterations between FTI-treated control cells (yellow) with FTI-treated fibroblasts from subjects with HGPS (red). Using criteria of corrected p-value from unpaired t-test <0.01 and two-fold change in expression, only one gene, *PIK3CB* was upregulated in FTI treated cells from subjects with HGPS compared to cells from normal subjects treated with FTI. Therefore, 99% of the genes in FTI-treated cells from subjects with HGPS were expressed at the same levels as in FTI-treated normal fibroblast.
doi:10.1371/journal.pone.0011132.g004

level remained unchanged in comparison to wild type cells [35]. These findings suggested that the Rb was unstable in the absence of lamin A/C. In HGPS cells, we observed a decrease in both mRNA and protein levels of Rb. Our findings indicate that the lamin A/C-Rb signaling might be altered differently depending on the status of A-type lamins in cells (absence of A-type lamins expression versus expression of an abnormal lamin A variant, progerin). Further studies are required to understand the molecular mechanisms driven by alterations in A-type lamins on Rb signaling pathways.

FTI treatment of cells from subjects with HGPS leads to a significant reversal of the abnormal expression of genes encoding proteins in the lamin A-Rb signaling network. This is consistent with the improved phenotypes that result from blocking protein farnesylation in cellular and animal models of the disease [16,17,18,20,21,23]. Defective Rb activity appeared to be responsible for the repression of the large set of downstream transcription factors and regulators. To evaluate the impact of FTIs, we also investigated FTI-induced gene expression changes in fibroblasts from unaffected controls. Remarkably, Rb appeared again in this dataset as a key regulator of the downstream events. Not surprisingly, FTI-treated cells from subjects with HGPS demonstrated a near total reversal (99%) of abnormally expressed genes in comparison to FTI-treated control cells. Our finding suggests that a potential correction of the HGPS phenotype does not necessarily require a full inhibition of protein farnesyltransferase but rather points to the importance of normalizing Rb expression levels and function.

The results of our study further support the potential use of FTIs as therapeutic agents in HGPS. A phase II clinical trial treating children with HGPS with lonafarnib, the FTI we used in this study, was initiated in spring of 2007 [39]. Because of known toxicity, the dose of lonafarnib administrated to these children may generate tissue concentrations below those necessary to obtain the effects we observed *in vitro*. Therefore, in human subjects, other drugs in addition to an FTI may be necessary to inhibit protein prenylation to a similar extent as in cultured cells. One possibility is to add a statin and an aminobisphosphonate, which have been shown to inhibit prelamin A prenylation and improve the phenotype of mice deficient in Zmpste24 [24]. Furthermore, our findings suggest that screening candidate molecules against targets we identified in the lamin A-Rb signaling network could identify novel therapeutic approaches to treat HGPS. Proteins in this network could also serve as biomarkers for HGPS.

Because the lamin A-Rb interaction is a new signaling axis in the pathogenesis of HGPS, we investigated its potential role in physiological aging. We found that *LMNA* and *RB1* expression appeared to be inversely modulated in the elderly. These observations were similar to our findings in cells from subjects with HGPS. Although lamin A and lamin C levels were not altered, an abnormal prelamin A variant, progerin, is expressed and Rb expression is decreased. In addition to changes in the expression of A-type lamins, several studies suggest that abnormal prelamin A may accumulate in normal aging [14,31,40,41]. Our results, therefore, suggest that therapeutic approaches to re-establish a proper lamin A-Rb signaling network may be beneficial in preventing complications of physiological aging.

Materials and Methods

Cell culture and FTI treatment

Dermal fibroblasts from subjects with HGPS were obtained from the Progeria Research Foundation (www.progeriaresearch.org). The following fibroblasts were used: HGADFN003 (M, age 2), HGADFN127 (F, age 3), HGADFN155 (F, age 1), HGADFN164 (F, age 4) and HGADFN188 (F, age 2). Age-matched control dermal fibroblasts were obtained from Coriell Institute for Medical Research (Camden, NJ). The following cell lines were used: GM01652C (F, age 11), GM02036A (F, age 11), GM03349C (M, age 10), GM03348E (M, age 10), GM08398A (M, age 8). The Institutional Review Board at Columbia University Medical Center approved the use of human cells established from skin biopsies from patients with HGPS and unaffected individuals.

Cells were cultured in DMEM containing 15% fetal bovine serum, 1% glutamine and 1% penicillin/streptomycin. Cells were subcultured at 80% confluency to keep cultures in growth phase and collected at population doublings between 22 and 27.

Treatment with the FTI lonafarnib (Schering-Plough, Kennilworth, NJ) was initiated when the cells reached 45–50% confluency. Lonafarnib was added to the culture media to a concentration of 1.5 μM FTI daily for three days. Untreated fibroblasts were cultured in parallel with added vehicle (DMSO). Cellular toxicity was determined by Trypan blue exclusion. Preliminary studies were conducted with varying lonafarnib concentrations and 1.5 μM was selected, as higher concentrations caused increased cell death and lower concentrations less effectively blocked protein farnesylation (data not shown).

RNA Preparation

Total RNA was extracted from the cell pellets using the Rnase Mini kit (Qiagen, Valencia, CA). Spectrophotometric quantification of RNA was performed and the purity assessed by measurement of the 260 nm/280 nm absorbance ratio. Ratios between 1.9 to 2.1 were accepted. Additionally, samples were run on a 1.2% agarose gel to examine integrity of the RNA. All gels showed two discrete bands at the 4.5 kb and 1.9 kb, representing the 28S and 18S ribosomal subunits, respectively (data not shown).

Microarray processing

RNA samples were amplified once using the One Cycle Target Labeling and Control Reagents and labeled for hybridization using the Affymetrix reagents and protocols (Affymetrix, Santa Clara, CA). Prior to hybridization, biotin-labeled cRNA samples were examined on an Agilent 2100 Bioanalyzer (Agilent Technoligies, Palo Alto, CA) to ensure quality. The samples were hybridized to U133 Plus 2.0 Arrays (Affymetrix) at the Columbia University Microarray Core Facility.

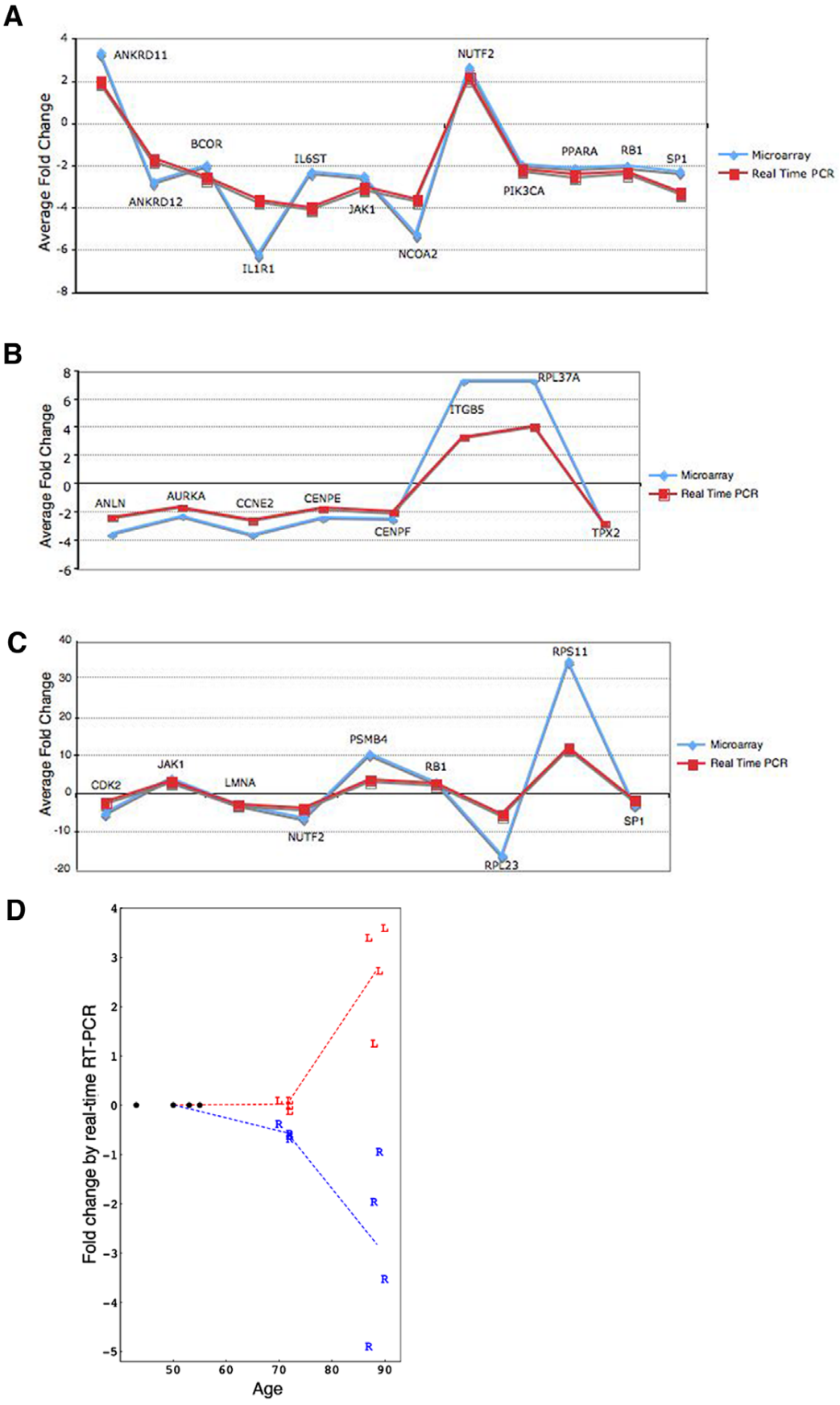

A
Average Fold Change
ANKRD11
ANKRD12
BCOR
IL1R1
IL6ST
JAK1
NCOA2
NUTF2
PIK3CA
PPARA
RB1
SP1
Microarray
Real Time PCR
B
Average Fold Change
ANLN
AURKA
CCNE2
CENPE
CENPF
ITGB5
RPL37A
TPX2
Microarray
Real Time PCR
C
Average Fold Change
CDK2
JAK1
LMNA
NUTF2
PSMB4
RB1
RPL23
RPS11
SP1
Microarray
Real Time PCR
D
Fold change by real-time RT-PCR
Age

Figure 5. Real-time RT-PCR validation of microarray results for selected genes. (A) Validation of a set of genes identified in microarray analysis comparing fibroblast from subjects with HGPS to controls. The mean value of expression for indicated genes measured using real time PCR (red; $p<0.05$), and microarray (blue; $p<0.01$) are shown. (B) Validation of a set of genes identified in microarray analysis comparing fibroblasts from control subjects with and without FTI treatment. (C) Validation of a set of genes identified in microarray analysis comparing fibroblasts from subjects with HGPS with and without FTI treatment. (D) Relative expression of *LMNA* (L, red) and *RB1* (R, blue) in human skin derived from 16 unaffected individuals of various age groups was analyzed by quantitative real-time RT-PCR. A relative quantification method was used and the data from the 38 to 55-year-old age group were calibrated to a relative quantity of 1. All values were normalized to GAPDH (Materials and Methods).
doi:10.1371/journal.pone.0011132.g005

The microarray data have been deposited in the European Bioinformatics Institute's ArrayExpress (www.ebi.ac.uk/arrayexpress) under accession number: E-MEXP-2597 along with detailed protocol notes.

Microarray data analysis

Data outputs were normalized and analyzed using GeneSpring GX 10.0 commercial software package (Agilent Technologies). Affymetrix data were uploaded into GeneSpring GX, normalized and quality control assessment was run to ensure the integrity of the samples. A t-test comparison was performed for each experimental group. The P-value cutoff was set to 0.01, a Benjamini-Hochberg correction was applied and the significant fold difference was considered two-fold above or below baseline. For interpretation of the results, MetaCore and GeneGo (www.Genego.com) and Ingenuity Pathways Analysis (Ingenuity Systems, www.ingenuity.com) software were used. Datasets containing gene identifiers and corresponding expression values were uploaded in the applications. An up or down change of two-fold or greater was set to identify genes whose expression was differentially regulated. These genes, or focus molecules, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base.

Network Analysis and Visualization

Gene sets were analyzed using MetaCore (www.genego.com). We built networks using as far as possible all and only genes belonging to the differentially regulated sets. The assembly of the networks exhibited in Fig. 2 and Fig. 3 were initially loaded with only the differentially regulated genes from the respective datasets. Then we added the known causal genes (encoding A-type lamins for Fig. 2 and protein farnesyltransferase for Fig. 3) and manually laid out the remaining genes in a semi-circular network based on the directness of the downstream interactions to the causal gene. In both cases, only a small number of genes had direct downstream interactions. For these and subsequent layers, we applied two rules. Rule 1: if a gene was downstream only to genes on the current layer or more inner layers, and if the gene had no genes downstream of it, it was moved to the outermost layer; otherwise, it was moved to the next layer. Rule 2: after the second layer, if the gene did not have downstream connections to genes in the current or previous layer, the gene was moved to a layer beyond the current layer. Following the initial manual layout, a number of differentially expressed genes were still not assigned to a downstream position from the causal gene. We used the MetaCore transfactor analysis tool to determine which transfactors could be implicated based on published data. This tool produces a list of transfactors belonging to the dataset analyzed. The list is ranked by an estimate of likelihood of being involved. Starting with the most likely, we added transfactors to the network, manually re-adjusting the network each time. We moved down the list until we had either exhausted the possible transfactors or finished the list of genes.

Real-time RT-PCR analysis

We synthesized cDNA using Omniscript Reverse Transcriptase (Qiagen) using total cellular RNA as template. RNA from fibroblasts from three subjects with HGPS and three controls, cultured with or without FTI, was used. Primers were designed using Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). The list of genes that were validated by RT-PCR and their corresponding primers are shown in Table S3.

Real-time RT-PCR reactions contained Power SYBR Green PCR mastermix (Applied Biosystems), 200 nM of each primer, and 0.1 µl of template in a 20-µl reaction volume. Amplification was carried out using the 7300 Real-Time PCR Detection System (Applied Biosystems) with an initial denaturation at 95°C for two minutes followed by 50 cycles at 95°C for 30 seconds and 62°C for 30 seconds. Three experiments were performed for each assay, in which the samples were run in triplicate. GAPDH was used as an endogenous control and quantification was performed using the relative quantification method where the real-time PCR signal of the experimental RNA was measured in relation to the signal of the control. The $2(\Delta\Delta C_T)$ method was used to calculate relative changes in gene expression [42].

Western blot analysis

Cells were extracted in Laemmli sample buffer (Bio-Rad) and heated for five minutes at 95°C. Approximatey 30 µg total protein extracts were loaded in parallel on a 7.5% polyacrylamide gel. An average of 7 mg of human skin tissues were extracted in Laemmli buffer using the bullet blender according to manufacturer instructions (Next Advance, Averill Park, NY). Approximately 40 µg skin protein extracts were loaded on the gel. After separation by electrophoresis, proteins were transferred to nitrocellulose membranes and incubated with blocking buffer as described previously [12]. Membranes were incubated sequentially with anti-prelamin A C20 antibodies (Santa Cruz Biotechnology, Santa Cruz, CA), anti-lamin A/C antibodies [43] (kindly provided by Dr. Nilabh Chaudhary), anti-actin antibodies (Sigma-Adrich, St. Louis, MO) and anti-HDJ-2 (Abcam, Cambridge, MA), washed and then incubated with a corresponding secondary antibody coupled to horseradish peroxidase (Jackson ImmunoResearch Laboratories, West Grove, PA). Other blots were similarly incubated sequentially with anti-Rb (BD Biosciences Pharmingen, San Diego, CA), anti- progerin [31], anti-lamin A (133A2, abcam), anti-laminA/C and anti-actin antibodies. Proteins were visualized using the enhanced chemiluminescence detection system (GE Healthcare, Piscataway, NJ). Signals obtained on the autoradiograms were analyzed by densitometry using Quantity One 1-D analysis software (Bio-Rad) on the scanned images.

Indirect immunofluorescence microscopy

Immunohistochemistry was performed on 6 µm frozen human skin sections fixed in methanol/acetone (1V/1V) at −20°C for 10 minutes and washed in phosphate-buffered saline, then blocked in phosphate-buffered saline containing 3% bovine serum albumin, 10% normal goat serum and 0.3% Triton X-100 for 30 minutes and 1 hour in the same buffer without Triton X-100. Cells and slides were incubated with anti-laminA/C and anti-Rb (BD Biosciences Pharmingen), or anti-progerin and anti-Rb or anti-lamin antibodies. The secondary antibodies were affinity purified

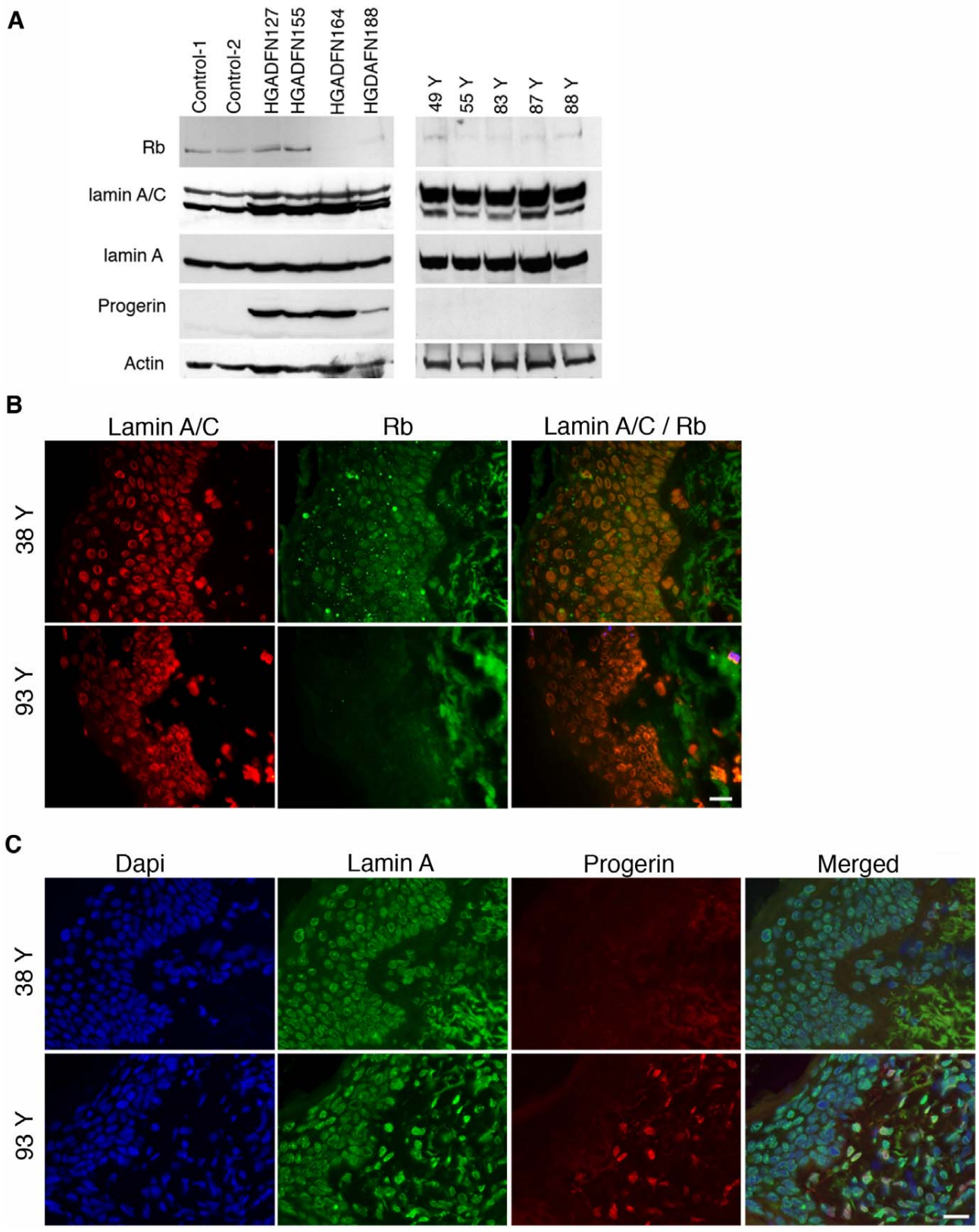


Figure 6. Detection of lamin A/C and Rb in cells from subjects with HGPS and in normal human skin biopsies. (A) Left panel: Western blot of proteins extracted from control fibroblasts (Control-1 [GM02036A], Control-2 [GM03349C]) and fibroblasts from subjects with HGPS (HGADFN127, HGADFN155, HGADFN164, HGADFN188) sequentially probed with anti-lamin A/C, anti- lamin A, anti-progerin (progerin) and anti-actin antibodies. Right panel: Western blot of protein extracts from skin of individuals at different ages as indicated (49 to 88 Year old, Y denote year). Blots

were sequentially probed with antibodies cited above. (B) In situ detection of lamin A/C and Rb in human skin sections. Skin sections from 93-year-old and 38-year-old subjects were probed with antibodies against lamin A/C (red) and Rb (green) and signal overlap is shown at right (Lamin A/C/Rb). (C) Skin sections as in panel (B) were probed with antibodies against lamin A (green) and progerin (red) and DNA was stained with 4′,6-diamidino-2-phenylindole (blue); the merged signals are indicated. Scale bar, 20 μM.
doi:10.1371/journal.pone.0011132.g006

Alexa Fluor 488 goat or donkey immunoglobulin G antibodies (Life Technologies, Carlsbad, CA) and Cy3-conjugated IgG antibodies (Jackson ImmunoResearch laboratories). All samples were also counterstained with 4′,6-diamidino-2-phenylindole (Sigma-Aldrich). Images were acquired on an Axioplan fluorescence microscope (Carl Zeiss, Thornwood, NY).

Supporting Information

Figure S1 High magnification of Fig. 2, panel A.
Found at: doi:10.1371/journal.pone.0011132.s001 (1.37 MB TIF)

Figure S2 High magnification of Fig. 2, panel B.
Found at: doi:10.1371/journal.pone.0011132.s002 (3.89 MB TIF)

Figure S3 High magnification Fig. 2, panel C.
Found at: doi:10.1371/journal.pone.0011132.s003 (1.40 MB TIF)

Figure S4 High magnification Fig. 2, panel D.
Found at: doi:10.1371/journal.pone.0011132.s004 (1.63 MB TIF)

Figure S5 Genome-wide expression profiling of FTI-treated and untreated control fibroblast cultures. (A) Genes differentially expressed in normal fibroblasts treated with FTI compared to untreated normal fibroblasts were assigned to diverse cellular functions according to IPA, and (B) were associated with canonical pathways according to IPA. (C) Pie chart indicates the subcellular localization of the protein products of the differentially expressed genes according to information contained in the Ingenuity Knowledge Base.
Found at: doi:10.1371/journal.pone.0011132.s005 (0.37 MB TIF)

Figure S6 Validation of the microarray analysis by real-time RT-PCR. (A) Validation of a set of genes identified in microarray analysis comparing fibroblasts from subjects with HGPS to control. The mean value of expression for indicated genes measured using real time RT-PCR are indicated (SD: standard deviation; $p<0.05$), and microarray fold change ($p<0.01$) are shown. (B) Validation of a set of genes identified in microarray analysis comparing fibroblasts from control subjects with and without FTI treatment. (C) Validation of a set of genes identified in microarray analysis comparing fibroblasts from subjects with HGPS treated with FTI to untreated fibroblasts from control subjects. (D) Validation of the only gene differentially expressed between FTI-treated fibroblasts from subjects with HGPS to FTI-treated fibroblasts from control subjects. Fold changes measured by real-time RT-PCR and microarray analyses are indicated.
Found at: doi:10.1371/journal.pone.0011132.s006 (0.71 MB TIF)

Figure S7 Distribution of progerin and Rb in cells from a subject with HGPS. Immunohistochemy was performed on fibroblasts from an unaffected control (GM03348) and a subject with HGPS (HGADFN003) at PPD 25 to 30. Cells were stained with anti-progerin antibody [31] (red) and anti-Rb monoclonal antibody (BD Biosciences Pharmingen) (green). Chromatin was stained with dapi. The triple merged signals are indicated. Scale bar, 20 μM.
Found at: doi:10.1371/journal.pone.0011132.s007 (1.31 MB TIF)

Table S1 Comparison of the differentially expressed genes in fibroblasts from subjects with HGPS from this study to the differentially expressed genes lists in studies by Scaffidi and Mitseli [28] and Csoka et al.[27] Control versus HGPS differentially expressed genes established after a statistical analysis using the t test with 1% significance and 2-fold cutoff were compared to the initial microarray analyses performed on three HGPS fibroblast strains (Coriell cell repositories) derived from patients at age 8 (AG11513), 9 (AG10750) and 14 years old (AG11498)[27]. This small overlap may be due to variation intrinsic to each cell, in addition to the fact that cells from subjects with HGPS exhibit increased variation in cellular phenotype with cellular age in vitro and with the donor's age. Our study was performed using five fibroblast cultures from five subjects with HGPS at age 2 to 4 years old, kindly provided by Progeria Research Foundation. Cells were collected at an early PPD (<25) when their growth rate remained similar to that of control fibroblast cultures. Phenotypic variations and different levels of progerin expression can contribute to the heterogeneity between fibroblasts from subjects with HGPS. We compared the control versus HGPS gene list with another study from Scaffidi and Mitseli that used a cellular model for HGPS by overexpressing progerin in normal immortalized fibroblasts [28] and found very little overlap.
Found at: doi:10.1371/journal.pone.0011132.s008 (0.07 MB TIF)

Table S2 Screening of a set of genes that were previously suggested to be perturbed in HGPS cells [28,33]. Using the same oligonucleotides for RT-PCR described by Fong et al 2009[2], we screened cultured fibroblasts from subjects with HGPS and from control individuals used in this study. The fold change between fibroblasts from subjects with HGPS and from control individuals in mRNA transcript for the corresponding genes are indicated with a $P<0.05$.
Found at: doi:10.1371/journal.pone.0011132.s009 (0.09 MB TIF)

Table S3 List of primers used for real time PCR.
Found at: doi:10.1371/journal.pone.0011132.s010 (21.31 MB TIF)

Acknowledgments

We would like to thank Dr. W. Robert Bishop (Schering-Plough) for providing the lonofarnib SCH66336. We thank the Progeria Research Foundation and the patient families for providing HGPS fibroblasts.

Author Contributions

Conceived and designed the experiments: KD. Performed the experiments: JM DEM. Analyzed the data: JM SIOD VPS RS HW LBG KD. Contributed reagents/materials/analysis tools: SIOD DEM VPS RS DR HW LBG. Wrote the paper: KD.

References

1. Hutchinson J (1886) Case of congenital absence of hair, with atrophic condition of the skin and its appendages, in a boy whose mother had been almost wholly bald from alopecia areata from the age of six. Lancet 1: 923.
2. Gilford H (1904) Ateleiosis and progeria: continuous youth and premature old age. British Medical Journal 2: 914–918.
3. DeBusk FL (1972) The Hutchinson-Guilford progeria syndrome. Journal of Pediatric 80: 697–724.
4. Merideth MA, Gordon LB, Clauss S, Sachdev V, Smith AC, et al. (2008) Phenotype and course of Hutchinson-Gilford progeria syndrome. New England Journal of Medecin 358: 592–604.

5. Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, et al. (2003) Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. Nature (London) 423: 293–298.
6. De Sandre-Giovannoli A, Bernard R, Cau P, Navarro C, Amiel J, et al. (2003) Lamin A truncation in Hutchinson-Gilford progeria. Science 300: 2055.
7. Cao H, Hegele RA (2003) LMNA is mutated in Hutchinson-Gilford progeria (MIM 176670) but not in Wiedemann-Rautenstrauch progeroid syndrome (MIM 264090). Journal of Human Genetics 48: 271–274.
8. Lin F, Worman HJ (1993) Structural organization of the human gene encoding nuclear lamin A and nuclear lamin C. Journal of Biological Chemistry 268: 16321–16326.
9. Worman HJ, Fong LG, Muchir A, Young SG (2009) Laminopathies and the long strange trip from basic cell biology to therapy. Journal of Clinical Investigation 119: 1825–1836.
10. Goldman RD, Shumaker DK, Erdos MR, Eriksson M, Goldman AE, et al. (2004) Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. Proceedings of the National Academy of Sciences 101: 8963–8968.
11. Lutz RJ, Trujillo MA, Denham KS, Wenger L, Sinensky M (1992) Nucleoplasmic localization of prelamin A: Implications for prenylation-dependent lamin A assembly into the nuclear lamina. Proceedings of the National Academy of Sciences 89: 3000–3004.
12. McClintock D, Gordon LD, Djabali K (2006) Hutchinson-Gilford progeria mutant lamin A primarily targets human vascular cells as detected by an anti-Lamin A G608G antibody. Proceedings of the National Academy of Sciences 103: 2154–2159.
13. Shumaker DK, Dechat T, Kohlmaier A, AAdam SA, Bozovsky MR, et al. (2006) Mutant lamin A leads to progressive alterations of epigenic control in premature aging. Proceedings of the National Academy of Sciences 103: 8703–8708.
14. Cao K, Capell B, Erdos MR, Djabali K, Collins FS (2007) A Lamin A protein isoform overexpressed in Hutchinson-Gilford progeria syndrome interferes with mitosis in progeria and normal cells. Proceedings of the National Academy of Sciences 104: 4949–4954.
15. Dechat T, Shimi T, Adam SA, Rusinol AE, Andres DA, et al. (2007) Alteration in mitosis and cell cycle progression caused by a mutant lamin A known to accelerate human aging. Proceedings of the National Academy of Sciences 104: 4955–4960.
16. Yang SH, Bergo MO, Toth JI, Qiao X, Hu Y, et al. (2005) Blocking protein farnesyltransferase improves nuclear blebbing in mouse fibroblasts with a targeted Hutchinson-Gilford progeria syndrome mutation. Proceedings of the National Academy of Sciences 102: 10291–10296.
17. Toth JI, Yang SH, Qiao X, Beigneux AP, Gelb MH, et al. (2005) Blocking protein farnesyltransferase improves nuclear shape in fibroblasts from humans with progeroid syndromes. Proceedings of the National Academy of Sciences 102: 12873–12878.
18. Capell BC, Erdos MR, Madigan JP, Fiordalisi JJ, Varga R, et al. (2005) Inhibiting farnesylation of progerin prevents the characteristic nuclear blebbing of Hutchinson-Gilford progeria syndrome. Proceedings of the National Academy of Sciences 102: 12879–12884.
19. Mallampalli MP, Huyer G, Bendale P, Gelb MH, Michaelis S (2005) Inhibiting farnesylation reverses the nuclear morphology defect in a HeLa cell model for Hutchinson-Gilford progeria syndrome. Proceedings of the National Academy of Sciences 102: 14416–14421.
20. Glynn MW, Glover TW (2005) Incomplete processing of mutant lamin A in Hutchinson-Gilford progeria leads to nuclear abnormalities, which are reversed by farnesyltransferase inhibition. Human Molecular Genetic 14: 2959–2969.
21. Fong LG, Frost D, Meta M, Qiao X, Yang SH, et al. (2006) A protein farnesyltransferase inhibitor ameliorates disease in a mouse model of progeria. Science 311: 1621–1623.
22. Yang SH, Meta M, Qiao X, Frost D, Bauch J, et al. (2006) A farnesyltransferase inhibitor improves disease phenotypes in mice with a Hutchinson-Gilford progeria syndrome mutation. Journal of Clinical Investigation 116: 2115–2121.
23. Capell BC, Olive M, Erdos MR, Cao K, Faddah DA, et al. (2008) A farnesyltransferase inhibitor prevents both the onset and late progression of cardiovascular disease in a progeria mouse model. Proceedings of the National Academy of Sciences 105: 15902–15907.
24. Varela I, Pereira S, Ugalde AP, Navarro CL, Suarez MF, et al. (2008) Combined treatment with statins and aminobisphosphonates extends longevity in a mouse model of human premature aging. Nature Medicine 14: 767–772.
25. Ly DH, Lockhart DJ, Lemer RA, Schultz PG (2000) Mitotic misregulation and human aging. Science 287: 2486–2492.
26. Park WY, Hwang CI, Kang MJ, Seo JY, Chung JH, et al. (2001) Gene profile of replicative senescence is different from progeria or elderly donor. Biochemical and Biophysical Research Communication 282: 934–939.
27. Csoka AB, English SB, Simkevich CP, Ginzinger DG, Butte AJ, et al. (2004) Genome-scale expression profiling of Hutchinson-Gilford progeria syndrome reveals widespread transcriptional misregulation leading to mesodermal/mesenchymal sefects and accelerated atherosclerosis. Aging Cell. pp 235–243.
28. Scaffidi P, Misteli T (2008) Lamin A-dependent misregulation of adult stem cells associated with accelerated ageing. Nature Cell Biology 10: 452–459.
29. Mancini MA, Shan B, Nickerson JA, Penman S, Lee W (1994) The retinoblastoma gene product is a cell cycle-dependent, nuclear matrix-associated protein. Proceedings of the National Academy of Sciences 91: 418–422.
30. Ozaki T, Saijo M, Murakami K, Enomoto H, Taya Y, et al. (1994) Complex formation between lamin A and retinoblastoma gene product: identification of the domain on lamin A required for its interaction. Oncogene 9: 2649–2653.
31. McClintock D, Ratner D, Lokuge M, Owens DM, Gordon LB, et al. (2007) The Mutant Form of Lamin A that Causes Hutchinson-Gilford Progeria Is a Biomarker of Cellular Aging in Human Skin. PLoS ONE 2: e1269.
32. Wang Y, Panteleyev AA, Owens DM, Djabali K, Stewart CL, et al. (2008) Epidermal Expression of the truncated prelamin A causing Hutchinson-Gilford progeria syndrome: effects on keratinocytes, hair and skin. Human Molecular Genetic 17: 2357–2369.
33. Fong LG, Vickers TA, Farber EA, Choi C, Yun UJ, et al. (2009) Activating the synthesis of progerin, the mutant prelamin A in Hutchinson-Gilford progeria syndrome, with antisense oligonucleotides. Human Molecular Genetic 18: 2462–2471.
34. Markiewicz E, Ledran M, Hutchinson CJ (2005) Remodelling of the nuclear lamina and nucleoskeleton is required for skeletal muscle differentiation in vitro. Journal of Cell Science 118: 409–420.
35. Johnson BR, Nitta RT, Frock RL, Mounkes L, Barbie DA, et al. (2004) A-type lamins regulate retinoblastoma protein function by promoting subnuclear localization and preventing proteasomal degradation. Proceedings of the National Academy of Sciences 101: 9677–9682.
36. Giacinti C, Giordano A (2006) RB and cell cycle progression. Oncogene 25:5220–5227. Oncogene 25: 5220–5227.
37. Kennedy BK, Barbie DA, Classon M, Dyson M, Harlow E (2000) Nuclear organization of DNA replication in primary mammalian cells. Genes and Development 14: 2855–2868.
38. Frock RL, Kudlow BA, Evans AM, Jameson SA, Hauschka SD, et al. (2007) Lamin A/C and emerin are critical for skeletal muscle satellite cell differentiation. Genes and Development 20: 486–500.
39. Gordon LB, Harling-Berg CJ, Rothman FG (2007) Highlights of the 2007 Progeria Research Foundation scientific workshop: progress in translational science. Journal of Gerontology A Biological Science and Medical Science 63: 777–787.
40. Scaffidi P, Misteli T (2006) Lamin A-dependent nuclear defects in human aging. Science 312: 1059–1063.
41. Rodriguez S, Coppedè F, Sagelius H, Eriksson M (2009) Increased expression of the Hutchinson-Gilford progeria syndrome truncated lamin A transcript during cell aging. European Journal of Human Genetic 28.
42. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402–408.
43. Chaudhary N, Courvalin JC (1993) Stepwise reassembly of the nuclear envelope at the end of mitosis. Journal of Cell Biology 122: 295–306.

ARTICLE
PUBLISHED ONLINE: 6 APRIL 2015 | DOI: 10.1038/NCHEMBIO.1790

siRNA screen identifies QPCT as a druggable target for Huntington's disease

Maria Jimenez-Sanchez[1], Wun Lam[1,2], Michael Hannus[3,9], Birte Sönnichsen[3,9], Sara Imarisio[1,2,9], Angeleen Fleming[1,4,9], Alessia Tarditi[5,8], Fiona Menzies[1], Teresa Ed Dami[1,4,8], Catherine Xu[1,4], Eduardo Gonzalez-Couto[5,8], Giulia Lazzeroni[5], Freddy Heitz[5,8], Daniela Diamanti[5], Luisa Massai[5], Venkata P Satagopam[6,7], Guido Marconi[5,8], Chiara Caramelli[5,8], Arianna Nencini[5], Matteo Andreini[5,8], Gian Luca Sardone[5], Nicola P Caradonna[5], Valentina Porcari[5], Carla Scali[5], Reinhard Schneider[6,7], Giuseppe Pollio[5], Cahir J O'Kane[2], Andrea Caricasole[5,8]* & David C Rubinsztein[1]*

Huntington's disease (HD) is a currently incurable neurodegenerative condition caused by an abnormally expanded polyglutamine tract in huntingtin (HTT). We identified new modifiers of mutant HTT toxicity by performing a large-scale 'druggable genome' siRNA screen in human cultured cells, followed by hit validation in *Drosophila*. We focused on glutaminyl cyclase (QPCT), which had one of the strongest effects on mutant HTT-induced toxicity and aggregation in the cell-based siRNA screen and also rescued these phenotypes in *Drosophila*. We found that QPCT inhibition induced the levels of the molecular chaperone αB-crystallin and reduced the aggregation of diverse proteins. We generated new QPCT inhibitors using *in silico* methods followed by *in vitro* screening, which rescued the HD-related phenotypes in cell, *Drosophila* and zebrafish HD models. Our data reveal a new HD druggable target affecting mutant HTT aggregation and provide proof of principle for a discovery pipeline from druggable genome screen to drug development.

HD is a fatal, currently incurable, late-onset neurodegenerative disorder. The disease signs include involuntary and repetitive choreic movements, psychological dysfunction and cognitive impairment, which result from progressive degeneration of cortical and striatal neurons[1,2].

HD is caused by the expansion of a CAG repeat tract in exon 1 of the gene encoding huntingtin (*HTT*), which results in an abnormally long polyglutamine stretch in the N terminus of the protein[3]. Although the mechanisms are not fully understood, it is believed that the disease arises from a toxic gain of function of the mutant protein[4,5]. A hallmark of HD is the presence of intracellular aggregates, which is also a characteristic of the other ten polyglutamine-expansion disorders as well as other neurodegenerative conditions such as Parkinson's or Alzheimer's disease[6]. The role of these aggregates in the disease is not clear, although an increasing importance of the oligomeric forms in toxicity is emerging[7,8], and reducing mutant HTT aggregation with strategies such as pharmacological upregulation of chaperone function has been pursued as a therapeutic strategy in HD[9]. Mutant HTT toxicity is believed to be accentuated or possibly induced after cleavage events, resulting in the formation of short N-terminal polyglutamine-containing fragments, which can also be produced by aberrant splicing[10]. Hence, exon 1 models have been frequently used for disease modeling.

Here, we combined two approaches to identify modifiers of mutant HTT toxicity by first performing a cell-based screen to identify genes that, when knocked down, could suppress mutant HTT-induced toxicity, using a library of 5,623 siRNAs selected according to the potential druggability of their targets with small molecules[11]. We performed this screen in two different HD models. Initially, we screened the effects of siRNAs in a mammalian cell line inducibly expressing HTT with an abnormal polyglutamine expansion. In a secondary analysis, we validated primary hits in a *Drosophila* model of HD.

One of the strongest suppressors of mutant HTT toxicity in both mammalian cells and *Drosophila* was an enzyme responsible for the modification of N-terminal residues of glutamine or glutamate into an N-terminal 5-oxoproline or pyroglutamate (pE) named QPCT. QPCT not only suppressed mutant HTT induced toxicity but also greatly reduced the number of aggregates. This effect is not HTT-specific as QPCT exerted a general effect on aggregation of different aggregate-prone proteins, including other proteins containing an expanded polyglutamine or polyalanine tract, which could be attributed to increased levels of the chaperone αB-crystallin upon QPCT inhibition. Furthermore, we designed small-molecule modulators of QPCT activity, which effectively suppressed mutant HTT aggregation and toxicity in cells, neurons, fly and zebrafish models of the disease.

[1]Department of Medical Genetics, University of Cambridge, Cambridge Institute for Medical Research, Cambridge, UK. [2]Department of Genetics, University of Cambridge, Cambridge, UK. [3]Cenix BioScience GmbH, Dresden, Germany. [4]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. [5]Siena Biotech, Siena, Italy. [6]Structural and Computational Biology, EMBL, Heidelberg, Germany. [7]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg. [8]Present addresses: TPV GmbH, München, Germany (A.T.); Department of Neuroscience, Psychology, Drug Research and Child Health, University of Florence, Italy (T.E.D.); PerkinElmer Informatics, Madrid, Spain (E.G.-C.); GenKyoTex S.A., Geneva, Switzerland (F.H.); Autifony S.r.l., Verona, Italy (G.M.); Novartis Vaccines and Diagnostics, Rosia, Italy (C.C.); Department of Neuroscience, IRBM, Pomezia, Rome, Italy (M.A. and A.C.). [9]These authors contributed equally to this work. *e-mail: a.caricasole@irbm.it or dcr1000@hermes.cam.ac.uk

RESULTS

Primary cell screen for suppressors of mutant HTT toxicity

We performed the primary screen using a stable human embryonic kidney (HEK293) T-REx cell line expressing full-length human HTT bearing 138 polyglutamines (Q138) under the control of a tetracycline-inducible promoter, which we refer to as HTT(Q138). We confirmed the expression of HTT(Q138) after inducing the cells with doxycycline using antibodies recognizing the N terminus of human HTT (**Supplementary Results**, **Supplementary Fig. 1a** and **Supplementary Note 1**) and quantitative RT-PCR using primers spanning different areas of the human *HTT* cDNA (**Supplementary Fig. 1b**). This cell line had reduced cell viability after expression of mutant HTT, which was reverted through treatment with a known reference compound (Y27632)[12] (**Supplementary Fig. 1c**), suggesting that this model could be used to identify potential modulators of mutant HTT cellular toxicity in a large-scale screen.

For our high-throughput screen, we used a strategy consisting of an iterative siRNA screen where positive genes were selected after three consecutive rounds to compensate for the variability of the assay. We eliminated nonpositive siRNAs and added new siRNAs targeting the selected genes in consecutive passes. We assessed rescue of cellular toxicity by each siRNA by fluorescence microscopy and automated image analysis using three independent readouts, (i) number of cell nuclei (#nuclei), (ii) apoptotic index and (iii) aberrant nuclei index, and we used rescue indices to express the effect of each individual siRNA for each parameter analyzed. In an initial screen, we tested three independent siRNAs for each of the 5,623 genes (a total of 16,869 siRNAs), from which we selected 670 primary genes (see **Supplementary Note 1** for screen assay and criteria selection). As shown in **Supplementary Figure 2a**, the three readouts were partially redundant, as more than 50% of the 1,000 top-scoring siRNAs of one rescue index also ranked among the top 1,000 siRNAs of at least one of the other rescue indices. In **Supplementary Figure 1b**, a representation of rescue indices obtained in pass 1 shows the relatively large variability of the assay, with the nontargeting negative control siRNAs, negQ and negF, showing a #nuclei rescue index of 14% and 3%, respectively, whereas using siRNA targeting HTT as a positive control rendered a mean #nuclei rescue index of 81%.

After three consecutive rounds of screening, we selected 257 genes and ranked these on the basis of all three rescue indices, using #nuclei rescue index as a primary criterion (**Supplementary Data Set 1**).

Secondary RNAi screening in a *Drosophila* model of HD

To validate the hits obtained in mammalian cells and to focus on targets with potential relevance *in vivo*, we performed a secondary screen in a *Drosophila* model that expressed a construct containing 48 polyglutamines (Q48) that causes eye degeneration when expressed using a *GMR-GAL4* driver[13]. For most genes selected, we studied two upstream activating sequence (UAS)-RNAi constructs from the Vienna *Drosophila* RNAi Center (VDRC): a *P* element (GD) and a phiC31 (KK) construct, the latter of which carries more GAL4-binding sites and should therefore express the RNAi more strongly[14]. Of the 257 mammalian genes previously selected, we detected 133 that had one or more gene orthologs in flies (**Supplementary Data Sets 1** and **2**). Of these 133 mammalian genes with fly orthologs, 74 *Drosophila* genes (corresponding to 66 mammalian genes) rescued the Q48-induced eye degeneration with at least one RNAi line, whereas the others showed no obvious effect (**Supplementary Fig. 3a**,**b** and **Supplementary Data Sets 1** and **2**). We crossed suppressor RNAi lines to transgenic flies that expressed EGFP, also driven by the same *GMR*-GAL4 driver. We used EGFP to test whether modifiers affected transgene protein synthesis as Q48 levels can be modified by aggregation or autophagic degradation, which do not affect EGFP levels. Two of these fly RNAi lines, targeting orthologs to human cathepsin F (CTSF) and to human ADAM8, ADAM11 and ADAM33 reduced EGFP levels on western blots (**Supplementary Data Set 2**), suggesting a general effect of these genes in protein expression, whereas suppression exerted by the other RNAi lines seemed to be polyglutamine specific.

Functional categorization of mutant HTT modifiers

To gain further insight into the biological relevance of the data generated, we categorized the different sets of HD toxicity modulators according to their molecular function. Suppressors were enriched for certain classes of proteins such as GPCRs or transporters compared to the initial library, whereas the number of positive kinases in the screen was reduced, and no cytokines, growth factors or translational regulators were represented. We observed similar functional categorizations after selection from the cell and *Drosophila* screen (**Supplementary Fig. 4a**). An Ingenuity Pathway Analysis of the hits obtained in the primary screen in cells (**Supplementary Table 1a**) revealed that the majority of these proteins participate not only in general processes such as GPCR- or cAMP-mediated signaling but also in canonical pathways related to neurodegeneration such as apoptosis, mitochondrial dysfunction, amyloid processing or protein ubiquitination. Notably, ten of these proteins have been previously related to HD signaling, including subunits of the succinate dehydrogenase complex and HTT-associated protein 1 (**Supplementary Table 1a**). Many of the genes validated in *Drosophila* (**Supplementary Fig. 4b** and **Supplementary Table 1b**) are also involved in processes related to neurodegeneration but are enriched in mitochondrial metabolic pathways, especially those associated with fatty acid biosynthesis and metabolism.

Validation of QPCT in *Drosophila*

We focused our attention on a gene that had one of the strongest and most consistent effects in rescuing mutant HTT-induced toxicity in the cell-based siRNA screen. The gene product has glutaminyl cyclase activity and is named QPCT. Two orthologs have been reported in fly[15], *Glutaminyl cyclase* (*QC*) and *iso Glutaminyl cyclase* (*isoQC*), which show about 39% amino acid identity; a third fly ortholog, *CG6168*, shows expression restricted to male accessory glands (http://www.flyatlas.org/) and is not considered further here. RNAi lines targeting either *QC* or *isoQC* partially rescued eye depigmentation and mediated a substantial decrease in the number of black spots in flies expressing Q48 (**Fig. 1a**,**b** and **Supplementary Fig. 5a**). Data are shown for GD- and KK-RNAi lines in the case of *QC*, but only a KK line was available for *isoQC*. These effects are most likely independent of transcription and translation of the Q48 as no changes in EGFP protein levels were seen when we crossed transgenic flies expressing EGFP driven by the same *GMR-GAL4* driver as Q48 with *QC* or *isoQC* RNAi lines (**Supplementary Fig. 5b**). Thus, QPCT represents an interesting candidate to study in HD.

To further evaluate the benefits of downregulating QPCT on HD, we took advantage of an additional *Drosophila* model of neurodegeneration, HD flies that express exon 1 of HTT with 120 polyglutamines, GMR-HTT.Q120, in eye photoreceptors[16]. *Drosophila* has a compound eye consisting of many ommatidia, each of which is composed of eight photoreceptors, seven of which can be visualized by light microscopy using the pseudopupil technique[17]. Neurodegeneration results in the loss of visible rhabdomeres of each photoreceptor and can be rescued or enhanced by genetic or chemical approaches[18]. Consistent with our data using the Q48 flies, the loss of visible photoreceptors in transgenic flies expressing GMR-HTT.Q120 was partially rescued when they were crossed with RNAi lines for either of the two QPCT fly orthologs, *QC* and *isoQC* (**Fig. 1c**). We observed no effect on the number of rhabdomeres in QPCT RNAi lines in the absence of GMR-HTT.Q120. The effects of QPCT knockdown on toxicity correlated with a reduction in HTT aggregation, which we assessed in flies expressing GFP-tagged expanded HTT exon 1, HTTex1-Q46-eGFP, in the eye[19] (**Fig. 1d**).

QPCT modulates HTT aggregation

To further validate QPCT, we first confirmed the protective effect of its knockdown against toxicity and aggregation in HEK293 cells expressing the exon 1 of *HTT* (from residue 8) with a 74-polyglutamine expansion fused at its N terminal to EGFP (EGFP-HTT(Q74))[20] (**Fig. 2a** and **Supplementary Fig. 6a**,**b**). The QPCT siRNAs used in these experiments as well as in the screen do not target QPCT-like, a paralogous protein that catalyzes a similar reaction and shares 51% sequence identity with QPCT (**Supplementary Fig. 6b**,**c**). We also validated the effect of QPCT knockdown on aggregation in HeLa cells (**Supplementary Fig. 6d**), which, like HEK293 cells, express QPCT[21]. We also confirmed a decrease in protein aggregation of a construct that expresses full-length HTT carrying 138 polyglutamines (similar to the one used in the initial screen; **Supplementary Fig. 6e**). QPCT siRNA did not have a general anti-apoptotic effect as it did not affect caspase 3 activity in response to staurosporine treatment (**Supplementary Fig. 6f**). Consistent with these data, QPCT shRNA reduced aggregation of EGFP-Q80 (80 glutamines fused to EGFP) in primary cortical neurons (**Fig. 2b** and **Supplementary Fig. 6g**). We could not assess the effect of QPCT knockdown on polyglutamine-mediated toxicity in these neurons, as the levels of cell death obtained in this assay were very low (**Fig. 2b**). While knocking down QPCT was protective, overexpression of QPCT in HeLa and HEK293 cells increased the numbers of apoptotic nuclei and also led to a large accumulation of HTT(Q74) aggregates (**Fig. 2c** and **Supplementary Fig. 7a**), whereas QPCT overexpression did not increase caspase activity upon staurosporine treatment (**Supplementary Fig. 7b**). The effects of QPCT were activity dependent, as the catalytically inactive E201Q mutant did not increase the percentage of cells with HTT(Q74) aggregates (**Fig. 2d** and **Supplementary Fig. 7c**,**d**).

We measured mRNA levels of QPCT in HD mice and found that its expression was reduced compared to that in their wild-type littermates, suggesting that QPCT expression may be downregulated as a compensatory mechanism (**Supplementary Fig. 8**) and that increased QPCT activity may not be a prerequisite for aggregation.

QPCT catalyzes the modification of N-terminal glutamines or glutamates into a pE residue. Although the presence of an extended polyglutamine tract makes HTT a potential substrate for QPCT, this enzyme only modifies N-terminal residues, suggesting that any modification on mutant HTT would require an N-terminal cleavage to reveal a glutamine at the N terminal that could be cyclated. The formation of a pE residue may then affect its stability and propensity to aggregate, a hypothesis that was previously suggested[22]. This cleavage model in either the polyglutamine tract or HTT exon 1 or GFP is unlikely, as QPCT modulated the aggregation of constructs consisting only of isolated polyglutamine expansions (Q57 and Q81) C-terminally fused to EGFP (**Fig. 2e**,**f**) or of HTT exon 1 with 74 glutamines fused to hemagglutinin (HA)[23] (**Supplementary Fig. 9a**,**b**), and QPCT siRNA also reduced the aggregation of an expansion of 37 alanines[24] (**Fig. 2f**). QPCT appeared to modulate the early stages of mutant HTT oligomerization as QPCT overexpression increased the amounts of Flag-tagged monomeric mutant HTT that were coimmunoprecipitated by GFP-tagged mutant HTT (**Fig. 2g**)[25]. As QPCT did not interact with

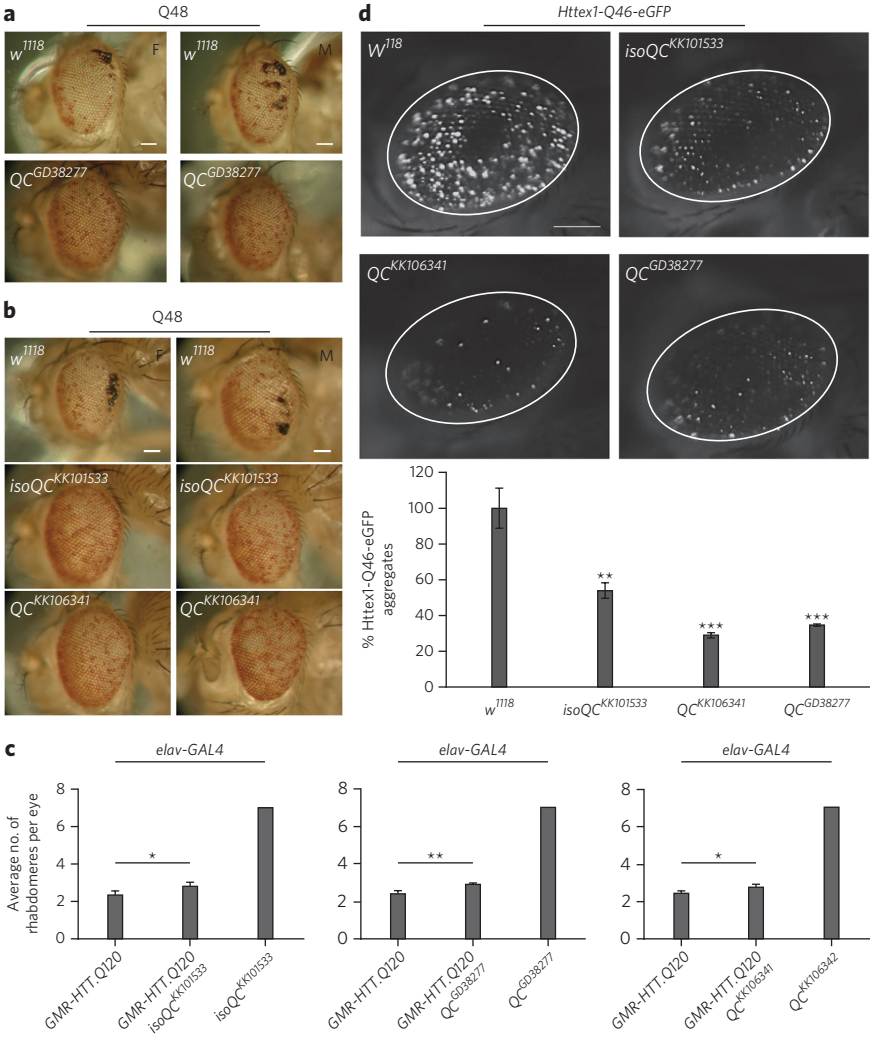


Figure 1 | Downregulation of QPCT in flies rescues HD toxicity. (**a**) The eye phenotype of flies that express Q48 crossed to w^{1118} (VDRC stock number 60,000) is rescued upon downregulation of *Drosophila QC* ($QC^{GD38277}$, VDRC GD-RNAi line 38277). Representative images of eye pigmentation rescue are shown. F, female; M, male. (**b**) Downregulation of QPCT fly orthologs *QC* and *isoQC* using KK-RNAi lines (lines $QC^{KK106341}$ and $isoQC^{KK101533}$) reduced the number of black necrotic-like spots on Q48 flies (see **Supplementary Fig. 5a** for quantification). Fisher's exact test was applied to statistically compare control and test genotypes. Females: $isoQC^{KK101533}P = 2.42 \times 10^{-14}$; $QC^{KK106341}P = 3.05 \times 10^{-12}$; males: $isoQC^{KK101533}P = 3.53 \times 10^{-8}$; $QC^{KK106341}P = 1.72 \times 10^{-9}$. (**c**) Loss of rhabdomeres due to expression of expanded HTT exon1 (*elav*-Gal4; GMR-HTT.Q120) in the eye was significantly rescued upon downregulation of QPCT fly orthologs *QC* or *isoQC* (GD- or KK-RNAi lines, as indicated). Graph shows the mean ± s.e.m. of the average number of rhabdomeres per eye from four independent experiments; one-tailed paired Student's *t*-test was used to test significance. (**d**) The number of aggregates in the eyes of flies expressing expanded HTTex1-Q46-eGFP using GMR-GAL4 was reduced by downregulating QPCT fly orthologs *QC* and *isoQC* (RNAi lines $isoQC^{KK101533}$, $QC^{KK106341}$ and $QC^{GD38277}$). Graph shows mean ± s.e.m. of the number of aggregates from four independent crosses for each genotype, with control levels set at 100%. One-tailed paired *t*-test was used for comparison between control and test genotypes ($n = 4$). In all panels, $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Scale bars, 200 µm.

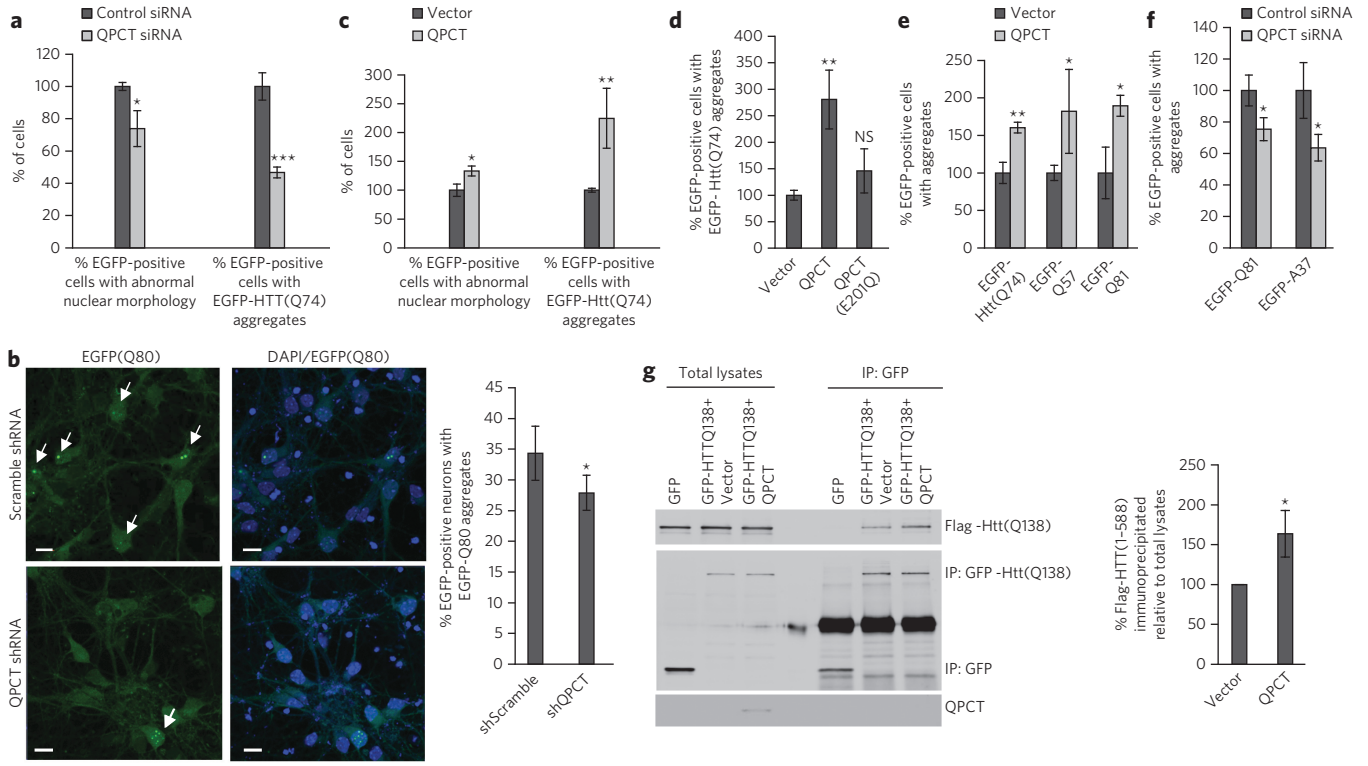


Figure 2 | QPCT modulates HTT toxicity and aggregation in mammalian cell lines and primary neurons. (**a**) The percentage of cells with apoptotic nuclei or HTT(Q74) aggregates is reduced in HEK293 cells transiently expressing EGFP-HTT(Q74) and treated with QPCT siRNA. Representative images are shown in **Supplementary Figure 6a**. (**b**) QPCT shRNA significantly reduced the number of aggregates in mouse primary cortical neurons expressing Q80-EGFP. Scale bars, 10 μm. The mean of three independent experiments, performed in triplicate, is represented in the graph. Significance was analyzed by two-tailed paired Student's *t*-test. (**c**,**d**) Overexpression of QPCT together with EGFP-HTT(Q74) in HeLa cells for 48 h increased the percentage of cells with apoptotic nuclear morphology and aggregates (**c**), but this effect is not observed with a catalytically inactive QPCT (QPCT(E201Q)-Flag) (**d**). (**e**) The percentage of HeLa cells expressing EGFP-HTT(Q74), EGFP-Q57 or EGFP-Q81 with aggregates is enhanced upon QPCT-Flag overexpression for 48 h. (**f**) QPCT siRNA reduces the percentage EGFP-Q81 or EGFP-A37 with aggregates in HEK293. (**g**) Overexpression of QPCT enhanced the amount of mutant HTT(1–548)-Flag coimmunoprecipitating with HTT(1–588)-GFP. Levels of Flag-HTT(1–588) co-immunoprecipitated relative to total lysates from five independent experiments are represented in the graph. Data were analyzed by two-tailed paired Student's *t*-test ($n = 5$ experiments). Full blot images are shown in **Supplementary Figure 17a**. In all panels, unless indicated, graphs show mean values with control conditions set to 100, and error bars represent s.d. from a triplicate experiment representative of at least three independent experiments. Statistical analyses were performed by two-tailed unpaired Student's *t*-test. $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$; NS, not significant.

HTT directly by immunoprecipitation (**Fig. 2g**), its effect on HTT oligomer formation is most likely mediated via intermediaries.

Design and characterization of compounds that inhibit QPCT

To target QPCT pharmacologically, we tested a previously described QPCT inhibitor[26] that did not rescue the HD phenotype in mammalian cells (**Supplementary Fig. 10a**,**b**). Although this compound has been effective in Alzheimer's disease models by reducing the formation of extracellular pE-Aβ, this effect may be due to extracellular QPCT inhibition[21,27]. Thus, we reasoned that the failure of this compound was most likely due to poor cell permeability. To generate new QPCT inhibitors, we used existing data on its structure and known inhibitors to generate three three-dimensional pharmacophore models, two ligand-based and one structure-based (using the human QPCT X-ray structure, Protein Data Bank (PDB) code 2AFW). We used these models, along with stringently applied central nervous system filters and a solubility model developed in house, to select 10,000 compounds from both commercially available screening compounds and the SienaBiotech compound library. We screened these molecules in a functional assay assessing the conversion of the H-Glu-AMC fluorogenic substrate into pyroGlu-AMC, as previously described[28], and selected hits associated with predicted robust binding for the hit-to-lead phase. The optimization strategy was based on physicochemical properties and approaches based on ensemble docking models. The ensemble docking methodology[29,30] was chosen to take into account the flexibility of the human QPCT catalytic site and was constructed using both X-ray structures and protein conformations coming from a 100-ns molecular dynamic study of the human QPCT 2AFW X-ray structure. The ensemble docking model was evolved during project development. Initially, only four X-ray structures were used (PDB codes 2AFW, 2AFX, 2AFZ and 2AFU[31]), and then a set of 16 protein conformations, selected by clustering of molecular dynamic simulations, was added to improve model accuracy. Recently, two more X-ray structures were added to the model (PDB code 3PBB[32] and 3SI0 (ref. 33)). All of the docking calculations were performed using CCDC Gold (versions 4 and 5)[34–36] along with an *ad hoc* developed program to rank and select the best-scoring ligand docking pose from the pool of QPCT conformations. Along with the biochemical readouts used during this optimization, we included a range of *in vitro* absorption, distribution, metabolism and elimination (ADME) assays, including solubility measurements, a central nervous system membrane permeability assay (PAMPA-BBB)[37] and stability

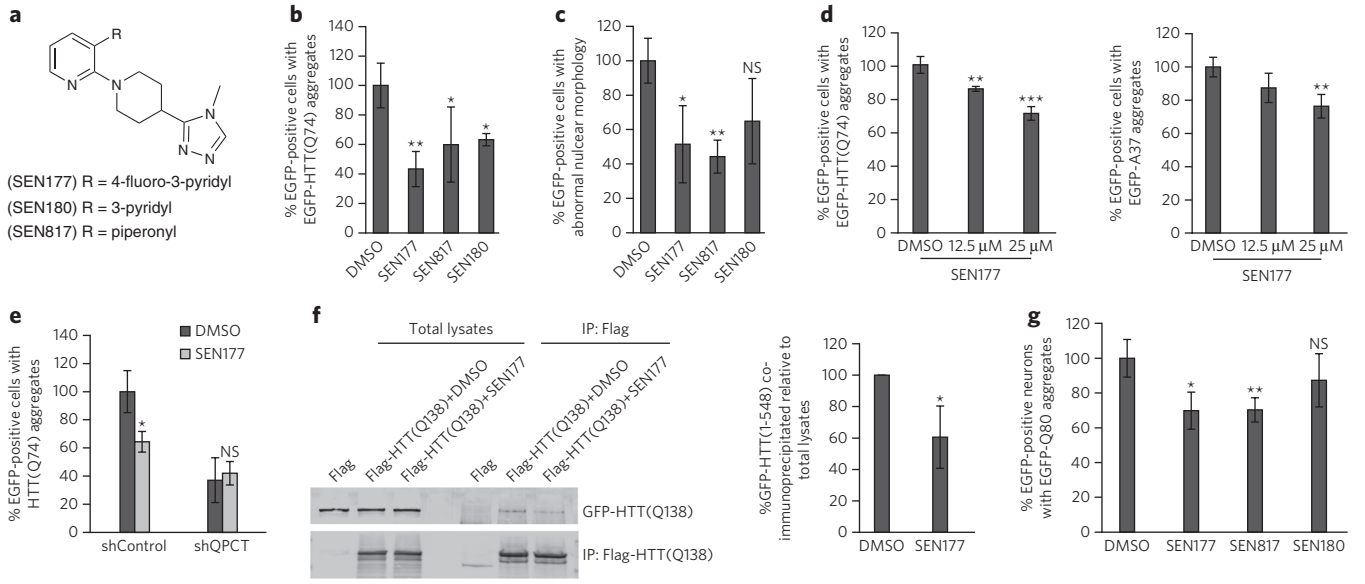


Figure 3 | Design of QPCT inhibitors that reduce mutant HTT aggregation. (**a**) Chemical structure of compounds designed to inhibit QPCT activity. The activity and *in vitro* ADME properties of the compounds are shown in **Supplementary Figure 11a**. (**b**,**c**) Treatment of HeLa cells expressing EGFP-HTT(Q74) with SEN177, SEN817 and SEN180 (50 μM) for 24 h reduced the percentage of cells with aggregates (**b**) and apoptotic nuclei (**c**). (**d**) SEN177 reduces the percentage of HEK293 cells with EGFP-HTT(Q74) or EGFP-A37 aggregates in a concentration-dependent manner. (**e**) SEN177 does not further reduce the percentage of EGFP-HTT(Q74) aggregates in QPCT shRNA-transfected cells. (**f**) SEN177 reduces the amount of HTT(1–588)-GFP coimmunoprecipitating with HTT(1–548)-Flag in HeLa cells (25 μM SEN177). The amount of GFP-HTT(1–548) immunoprecipitated relative to total lysates was quantified, and the average of five independent experiments is shown in the graph. Data were analyzed by two-tailed paired Student's *t*-test ($n = 5$ experiments). Full blot images are shown in **Supplementary Figure 17b**. (**g**) Primary neurons expressing EGFP-Q80 for 3 d were treated with 50 μM of the indicated compounds for a further 24 h. In all panels, unless indicated, graphs show mean values with control conditions set to 100, and error bars represent s.d. from experiments performed in triplicate, representative of at least three independent experiments. Statistical analyses were performed by two-tailed unpaired Student's *t*-test. $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$; NS, not significant.

measurements in the presence of human CYP3A4, a member of the cytochrome P450 mixed-function oxidase system and a key enzyme involved in the metabolism of xenobiotics in humans.

We selected a series of compounds on the basis of these properties and validated their effects on mutant HTT aggregation and toxicity in cells expressing HTT(Q74)GFP, which led to the selection of three of them, SEN177 (**1**), SEN817 (**2**) and SEN180 (**3**) (**Fig. 3a**, **Supplementary Note 2** and **Supplementary Fig. 11a**). Nontoxic concentrations of these compounds caused a dose-dependent reduction in the percentage of cells with aggregates, which correlated with suppression of mutant HTT-induced apoptosis (**Fig. 3b–d** and **Supplementary Fig. 11b**). As seen with genetic knockdown experiments, pharmacologic inhibition of QPCT using these compounds also reduced aggregation of polyalanines (**Fig. 3d**) and did not affect protein levels, as assessed by measuring GFP levels by western blotting (**Supplementary Fig. 11c**) or by metabolic labeling of wild-type HTT followed by detection of newly synthesized protein in the presence of SEN177 (**Supplementary Fig. 11d**). Notably, the effect of these compounds was blocked when QPCT expression was suppressed by shRNA, confirming that they protect by a mechanism that requires QPCT inhibition (**Fig. 3e** and **Supplementary Fig. 11e**,**f**). Thus, though these compounds also inhibited QPCT-like (**Supplementary Fig. 11a**), and though we cannot exclude the possibility that at least some of the effects observed may be mediated by this QPCT isoenzyme, their effects on aggregation were QPCT dependent, as the shRNA used did not target QPCT-like. Consistent with these data, SEN177 greatly reduced the early stages of mutant HTT oligomerization, as it decreased the amounts of GFP-tagged monomeric HTT that were coimmunoprecipitated by Flag-tagged HTT (**Fig. 3f**). The protective effect of these compounds was also confirmed in primary cortical neurons (**Fig. 3g**), with SEN177 and SEN817 significantly reducing the percentage of neurons with Q80 aggregates.

QPCT modulates the levels of αB-crystallin

The effects of QPCT inhibition on HTT aggregation appeared to be independent of effects on protein clearance pathways targeting mutant HTT (autophagy and the ubiquitin-proteasome system; **Supplementary Fig. 12**), changes in mRNA or protein levels (**Supplementary Fig. 13a**,**b**) or secretion of the enzyme into the medium (**Supplementary Fig. 13c**). QPCT is localized in the endoplasmic reticulum (ER) and secretory pathway, and its knockdown, overexpression or inhibition seemed to have inconsistent and rather modest effects on different readouts of the ER stress response, as measured by GRP78 (also known as BIP) levels or phosphorylation of eIF2α, which did not correlate with its effect on aggregation (**Supplementary Fig. 14**). Our data also suggested that cAMP response element binding protein (CREB) or extracellular signal-regulated kinase (ERK) signaling, recently reported to be activated upon QPCT inhibition[38] (**Supplementary Fig. 15a**,**b**), or JNK signaling (**Supplementary Fig. 15c**) were unlikely contributors to the effects we observed.

QPCT overexpression or knockdown did not modulate levels of HSP70, the main inducible stress response chaperone (**Supplementary Fig. 15d**). We performed transcriptional profiling to assess changes in alternative molecular chaperones induced by SEN177 in the presence of mutant HTT and observed upregulation of several small heat shock proteins (sHSPs; HSPB6 with 1.6-fold increase, HSPB3 with 1.5-fold increase HSPB7 with a 1.5-fold increase and, notably, αB-crystallin, with a >2.5-fold increase in transcript levels; **Supplementary Fig. 16a** and **Supplementary Data Set 3**). We confirmed this induction at the protein level as well as with other QPCT inhibitors (**Fig. 4a**). Genetic inhibition of QPCT markedly

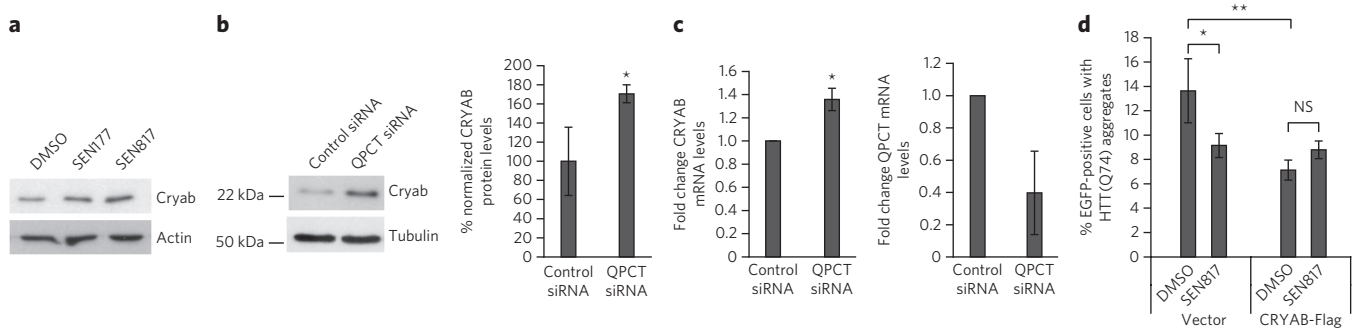


Figure 4 | QPCT inhibition induces alpha B-crystallin levels. (**a**) αB-crystallin (Cryab) protein levels were increased in cells transfected with HTT(Q74) GFP and treated with the indicated compounds at 25 μM for 24 h. Full blot images are shown in **Supplementary Figure 17c**. (**b**,**c**) Knockdown of QPCT for 24 h followed by transfection with HTT(Q74)GFP for another 24 h increased protein (**b**) and mRNA (**c**) levels of αB-crystallin. The fold change in mRNA of QPCT or αB-crystallin is represented in the graph, with error bars representing s.d. The mean of three independent experiments performed in triplicate was normalized to 1, and significance was calculated by one sample *t*-test. Full blot images are shown in **Supplementary Figure 17d**. (**d**) Overexpression of αB-crystallin (CRYAB-Flag) reduced the percentage of cells with HTT(Q74)GFP aggregates. SEN817 decreased aggregation when added at 25 μM for 24 h in control cells but not CRYAB-expressing cells. In all panels, unless indicated, graphs show mean values with control conditions set to 100 or 1, and error bars represent s.d. from an experiment performed in triplicate, representative of at least three independent experiments. Statistical analyses were performed by two-tailed unpaired Student's *t*-test. $^{**}P < 0.01$; $^{*}P < 0.05$; NS, not significant.

increased αB-crystallin protein and mRNA levels in the presence of HTT(Q74) (**Fig. 4b**,**c** and **Supplementary Fig. 16b**), whereas QPCT overexpression, which increased mutant HTT aggregation and toxicity (**Fig. 2c** and **Supplementary Fig. 7a**), reduced αB-crystallin levels (**Supplementary Fig. 16c**). QPCT also modestly modulated αB-crystallin levels in the absence of mutant HTT or in the presence of the nonpathogenic Q23 (**Supplementary Fig. 16b**,**c**).

As a sHSP, αB-crystallin acts as a molecular chaperone and is a suppressor of polyglutamine toxicity in cells and in *Drosophila*[39–41]. As expected, overexpression of αB-crystallin lowered the number of HTT(Q74) aggregates, whereas QPCT inhibitors failed to reduce aggregation further (**Fig. 4d** and **Supplementary Fig. 16d**), suggesting that this increase in αB-crystallin was a major contributor to the protection afforded by QPCT inhibition.

QPCT inhibition protects fly and zebrafish HD models

We tested QPCT inhibitors in flies expressing Httex1Q46 in the eye and found a reduction in the number of aggregates (**Fig. 5a**). The compound with a greatest effect, SEN177, was able to also rescue the number of visible rhabdomeres and prevent neurodegeneration (**Fig. 5b**).

A transgenic zebrafish expressing Htt exon 1 with 71Q fused to EGFP in the rod photoreceptors using the rhodopsin promoter has been established and validated as a model to study mutant HTT aggregation *in vivo*[42]. Zebrafish have two homologs with putative glutaminyl-peptide cyclotransferase activity, QPCT and QPCTLA, sharing 51% and 47% protein identity with QPCT and QPCT-like, respectively. To test the effect of pharmacologic inhibition of QPCT in this model, we first determined the maximum tolerated concentration for each of the three compounds tested in mammalian cells and subsequently treated HD larvae. SEN817 and SEN180 reduced the total number of EGFP aggregates in the retina (**Fig. 6a**), which correlated with a marked decrease in toxicity similar to the effect of the positive control, clonidine[42], as assessed by a rescue in the total area of eye photoreceptors (**Fig. 6b**).

Although the three compounds were protective, their effectiveness varied between these models, which might be due to the intrinsic properties of each system; SEN180 only mildly reduced aggregation in neurons, and the effect of SEN817 was not obvious in *Drosophila*. Although SEN177 had the highest *in vitro* activity and was able to efficiently reduce aggregates in mammalian cells, primary neurons and *Drosophila*, we found that this compound was tolerated at much higher concentrations than its analogs in zebrafish, and therefore the bioavailability in this model is much lower, which could explain the lack of effect in this system. All together, we have identified a number of small molecules that through QPCT inhibition have beneficial effects on the treatment of HD in a variety of *in vivo* models.

DISCUSSION

Our approach using a two-step screen, starting with an initial large-scale analysis in human cell models followed by validation in *Drosophila*, has yielded a number of potentially druggable targets which may be suitable for HD. A variety of high-throughput RNAi

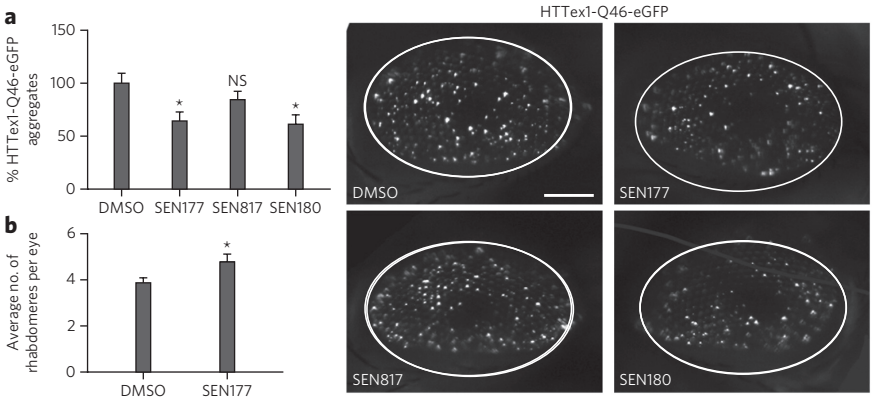


Figure 5 | Pharmacologic inhibition of QPCT in fly. (**a**) Flies that expressed HTTex1-Q46-eGFP in the eye have fewer aggregates after treatment with 50 μM of the indicated compounds. Data represent mean ± s.e.m. from four independent crosses for each compound. Statistical analyses were performed by one-tailed unpaired Student's *t*-test. Scale bar, 200 μm. $^{*}P < 0.05$; NS, not significant. (**b**) Flies expressing HTTex1-Q120 (GMR-HTT.Q120) show more rhabdomeres after treatment with SEN177 (50 μM). Graph represents the average number of rhabdomeres per eye ±s.e.m. from three independent experiments, with females and males counted separately and each experiment based on approximately 10 individuals per data point, scoring 15 ommatidia from each individual. Statistical analysis was performed using one-tailed paired Student's *t*-test. $^{*}P < 0.05$.

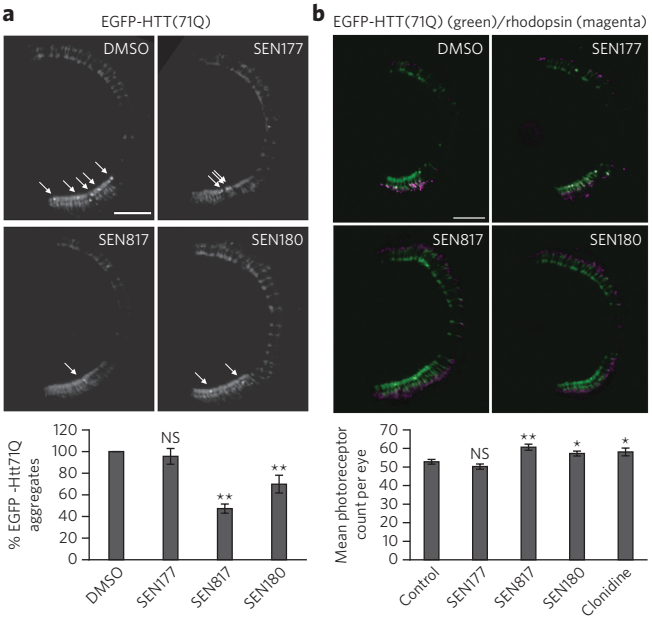


Figure 6 | Pharmacologic inhibition of QPCT in zebrafish. (**a**) Representative sections through the central retina of transgenic HD zebrafish at 7 d.p.f. treated with DMSO, SEN177 (1 mM), SEN817 (100 µM) or SEN180 (100 µM), showing aggregates (arrows) within the rod photoreceptors. Scale bar, 10 µm. Treatment with QPCT inhibitors resulted in reduction in aggregates (Student's *t*-test) for SEN187 and SEN810. (**b**) Representative sections through the central retina of transgenic HD zebrafish at 9 d.p.f. treated with DMSO, SEN177 (1 mM), SEN817 (100 µM) or SEN180 (100 µM). To demonstrate that loss of GFP corresponds to loss of photoreceptors, sections were stained with antirhodopsin (1D1) antibody (magenta). GFP labels the whole rod photoreceptor, whereas rhodopsin is present in the rod outer segment. Merged images show colocalization of GFP with the rhodopsin (magenta). Photoreceptor degeneration is ameliorated by SEN817 and SEN180. Scale bars, 10 µm. In all panels, $^{**}P < 0.01$; $^{*}P < 0.05$; NS, not significant.

screens have identified genetic suppressors of phenotypes mediated by mutant HTT N-terminal fragments in *Drosophila*, *Caenorhabditis elegans* and mammalian (mouse and human) cells[43–46]. In most cases, aggregation was the primary readout, often measured with C-terminal GFP fusions. Differences in the nature of the previous screens (species, cellular context, HTT fragment length, length of the polyglutamine expansion, primary readout, and differences in siRNA or shRNA sequences) complicates cross-screen comparisons. Also, virtually no screens in this area have examined their false negative rates owing to inefficient knockdown. Additionally, the screen presented here was biased toward the druggable component of the human genome, and a further selection was made in the course of triaging toward specific protein target classes. This most likely contributes to the relatively poor overlap of hits in the present and previous screens. A comparison with a screen performed in HEK293T cells to identify genetic suppressors of inducibly expressed mutant HTT exon 1 toxicity[45] revealed an overlap of only four genes (*CPA1*, *GRIN2A*, *NR3C2* and *USP21*) when considering the top 257 hits (**Supplementary Data Set 1**). However, matrix metalloproteases, identified in HEK293T cells as modulators of fragmentation and toxicity of N-terminal portions of mutant HTT[44] , were also identified in our data set together with PAK1, which we previously identified as a kinase promoting mutant HTT self-association and toxicity[25], thus validating the effectiveness of the screen.

Given the reproducible and clear rescue that QPCT inhibition exerts on mutant HTT toxicity in cells and in *Drosophila*, we focused on this target. A catalytically inactive QPCT was not able to increase the number of aggregates, suggesting that pE modifications modulate the levels of aggregates in HD models. Although one obvious mechanism would involve cleavage of the polyglutamine tract followed by cyclization of an N-terminal pE residue that may change properties such as stability or hydrophobicity, which would account for its change in aggregation[22], our data suggest that the effect of QPCT on HTT may be indirect. We found that the effect of aggregation modulation by QPCT was not restricted to mutant HTT as it affected aggregation of other aggregate-prone proteins, and we also found that QPCT influences the formation of mutant HTT oligomeric species. We observed an induction in several sHSPs, mostly αB-crystallin, suggesting that QPCT inhibition caused a stress response distinct from classical Hsp70 induction, which might be mediated by indirect substrates for pE modification. This molecular chaperone reduces aggregation of polyglutamine-containing proteins[39,41], α-synuclein[47,41]or amyloid-β peptide[48,49], underscoring QPCT inhibition as an effective target for misfolded protein disorders. As αB-crystallin is regulated at the transcriptional level, whereas QPCT resides in the secretory pathway, inhibition of QPCT may activate a signaling response that enhances αB-crystallin transcription. Our data suggest that this is most likely independent of an ER stress response or the involvement of ERK and CREB, which have been recently found phosphorylated upon QPCT inhibition[38], as well as other stress signaling pathways such as JNK. Further work will need to clarify the QPCT substrate mediating this effect. It is important to stress that the benefits of QPCT downregulation may not be restricted to αB-crystallin as an effector, as the upregulation of other related sHSPs may also contribute beneficially.

We identified and characterized a series of compounds that efficiently reduce mutant HTT aggregation in mammalian cell lines and also in primary mouse neurons, fly eye and in zebrafish. Although the levels of rescue obtained varied between compounds depending on the model used, this may be as a result of differences in absorption routes and bioavailability. Nevertheless, our data showed that pharmacologic inhibition of QPCT using this compound series can rescue HD phenotypes and provides proof of principle for QPCT as a potential therapeutic target for HD and possibly other related intracellular proteinopathies by modulating the formation of oligomeric forms, which have been proposed as the most toxic species in these diseases[7,8]. Clearly, further work, most likely including additional drug development, is required before we can consider whether this will be clinically relevant. Nevertheless, in a broader perspective, our data suggest that a discovery pipeline from druggable genome screen to drug development may be tractable for neurodegenerative diseases.

Received 14 August 2014; accepted 5 March 2015;
published online 6 April 2015

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Imarisio, S. *et al.* Huntington's disease: from pathology and genetics to potential therapies. *Biochem. J.* **412**, 191–209 (2008).
2. Zuccato, C., Valenza, M. & Cattaneo, E. Molecular mechanisms and potential therapeutical targets in Huntington's disease. *Physiol. Rev.* **90**, 905–981 (2010).
3. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971–983 (1993).
4. Mangiarini, L. *et al.* Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* **87**, 493–506 (1996).
5. Hodgson, J.G. *et al.* A YAC mouse model for Huntington's disease with full-length mutant huntingtin, cytoplasmic toxicity, and selective striatal neurodegeneration. *Neuron* **23**, 181–192 (1999).
6. Soto, C. & Estrada, L.D. Protein misfolding and neurodegeneration. *Arch. Neurol.* **65**, 184–189 (2008).

7. Takahashi, T. *et al.* Soluble polyglutamine oligomers formed prior to inclusion body formation are cytotoxic. *Hum. Mol. Genet.* **17**, 345–356 (2008).
8. Lajoie, P. & Snapp, E.L. Formation and toxicity of soluble polyglutamine oligomers in living cells. *PLoS ONE* **5**, e15245 (2010).
9. Hartl, F.U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332 (2011).
10. Sathasivam, K. *et al.* Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. USA* **110**, 2366–2370 (2013).
11. Bleicher, K.H., Böhm, H.-J., Müller, K. & Alanine, A.I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2**, 369–378 (2003).
12. Li, M., Huang, Y., Ma, A.A.K., Lin, E. & Diamond, M.I. Y-27632 improves rotarod performance and reduces huntingtin levels in R6/2 mice. *Neurobiol. Dis.* **36**, 413–420 (2009).
13. Marsh, J.L. *et al.* Expanded polyglutamine peptides alone are intrinsically cytotoxic and cause neurodegeneration in *Drosophila*. *Hum. Mol. Genet.* **9**, 13–25 (2000).
14. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448**, 151–156 (2007).
15. Schilling, S. *et al.* Isolation and characterization of glutaminyl cyclases from *Drosophila*: evidence for enzyme forms with different subcellular localization. *Biochemistry* **46**, 10921–10930 (2007).
16. Jackson, G.R. *et al.* Polyglutamine-expanded human huntingtin transgenes induce degeneration of *Drosophila* photoreceptor neurons. *Neuron* **21**, 633–642 (1998).
17. Franceschini, N. & Kirschfeld, K. Pseudopupil phenomena in the compound eye of *Drosophila*. *Kybernetik* **9**, 159–182 (1971).
18. Ravikumar, B. *et al.* Dynein mutations impair autophagic clearance of aggregate-prone proteins. *Nat. Genet.* **37**, 771–776 (2005).
19. Zhang, S., Binari, R., Zhou, R. & Perrimon, N. A genomewide RNA interference screen for modifiers of aggregates formation by mutant Huntingtin in *Drosophila*. *Genetics* **184**, 1165–1179 (2010).
20. Wyttenbach, A. *et al.* Effects of heat shock, heat shock protein 40 (HDJ-2), and proteasome inhibition on protein aggregation in cellular models of Huntington's disease. *Proc. Natl. Acad. Sci. USA* **97**, 2898–2903 (2000).
21. Cynis, H. *et al.* Inhibition of glutaminyl cyclase alters pyroglutamate formation in mammalian cells. *Biochim. Biophys. Acta* **1764**, 1618–1625 (2006).
22. Saido, T.C. Involvement of polyglutamine endolysis followed by pyroglutamate formation in the pathogenesis of triplet repeat/polyglutamine-expansion diseases. *Med. Hypotheses* **54**, 427–429 (2000).
23. Onodera, O. *et al.* Oligomerization of expanded-polyglutamine domain fluorescent fusion proteins in cultured mammalian cells. *Biochem. Biophys. Res. Commun.* **238**, 599–605 (1997).
24. Rankin, J., Wyttenbach, A. & Rubinsztein, D.C. Intracellular green fluorescent protein–polyalanine aggregates are associated with cell death. *Biochem. J.* **348**, 15–19 (2000).
25. Luo, S., Mizuta, H. & Rubinsztein, D.C. p21-activated kinase 1 promotes soluble mutant huntingtin self-interaction and enhances toxicity. *Hum. Mol. Genet.* **17**, 895–905 (2008).
26. Buchholz, M. *et al.* The first potent inhibitors for human glutaminyl cyclase: synthesis and structure-activity relationship. *J. Med. Chem.* **49**, 664–677 (2006).
27. Schilling, S. *et al.* Glutaminyl cyclase inhibition attenuates pyroglutamate Aβ and Alzheimer's disease–like pathology. *Nat. Med.* **14**, 1106–1111 (2008).
28. Schilling, S. *et al.* Continuous spectrometric assays for glutaminyl cyclase activity. *Anal. Biochem.* **303**, 49–56 (2002).
29. Yoon, S. & Welsh, W.J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comput. Sci.* **44**, 88–96 (2004).
30. Polgár, T. & Keserü, G.M. Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and β-secretase. *J. Chem. Inf. Model.* **46**, 1795–1805 (2006).
31. Huang, K.-F., Liu, Y.-L., Cheng, W.-J., Ko, T.-P. & Wang, A.H.-J. Crystal structures of human glutaminyl cyclase, an enzyme responsible for protein N-terminal pyroglutamate formation. *Proc. Natl. Acad. Sci. USA* **102**, 13117–13122 (2005).
32. Huang, K.-F. *et al.* Structures of human Golgi-resident glutaminyl cyclase and its complexes with inhibitors reveal a large loop movement upon inhibitor binding. *J. Biol. Chem.* **286**, 12439–12449 (2011).
33. Ruiz-Carrillo, D. *et al.* Structures of glycosylated mammalian glutaminyl cyclases reveal conformational variability near the active center. *Biochemistry* **50**, 6280–6288 (2011).
34. Jones, G., Willett, P. & Glen, R.C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53 (1995).
35. Jones, G., Willett, P., Glen, R.C., Leach, A.R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748 (1997).
36. Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W. & Taylor, R.D. Improved protein-ligand docking using GOLD. *Proteins* **52**, 609–623 (2003).
37. Palm, K., Luthman, K., Ros, J., Grasjo, J. & Artursson, P. Effect of molecular charge on intestinal epithelial drug transport: pH-dependent transport of cationic drugs. *J. Pharmacol. Exp. Ther.* **291**, 435–443 (1999).
38. Song, H. *et al.* Inhibition of glutaminyl cyclase ameliorates amyloid pathology in an animal model of Alzheimer's disease via the modulation of γ-secretase activity. *J. Alzheimers Dis.* **43**, 797–807 (2014).
39. Robertson, A.L. *et al.* Small heat-shock proteins interact with a flanking domain to suppress polyglutamine aggregation. *Proc. Natl. Acad. Sci. USA* **107**, 10424–10429 (2010).
40. Bilen, J. & Bonini, N.M. *Drosophila* as a model for human neurodegenerative disease. *Annu. Rev. Genet.* **39**, 153–171 (2005).
41. Tue, N.T., Shimaji, K., Tanaka, N. & Yamaguchi, M. Effect of αB-crystallin on protein aggregation in *Drosophila*. *J. Biomed. Biotechnol.* **2012**, 252049 (2012).
42. Williams, A. *et al.* Novel targets for Huntington's disease in an mTOR-independent autophagy pathway. *Nat. Chem. Biol.* **4**, 295–305 (2008).
43. Lejeune, F.-X. *et al.* Large-scale functional RNAi screen in *C. elegans* identifies genes that regulate the dysfunction of mutant polyglutamine neurons. *BMC Genomics* **13**, 91 (2012).
44. Miller, J.P. *et al.* Matrix metalloproteinases are modifiers of huntingtin proteolysis and toxicity in Huntington's disease. *Neuron* **67**, 199–212 (2010).
45. Miller, J.P. *et al.* A genome-scale RNA-interference screen identifies RRAS signaling as a pathologic feature of Huntington's disease. *PLoS Genet.* **8**, e1003042 (2012).
46. Yamanaka, T. *et al.* Large-scale RNA interference screening in mammalian cells identifies novel regulators of mutant huntingtin aggregation. *PLoS ONE* **9**, e93891 (2014).
47. Waudby, C.A. *et al.* The interaction of αB-crystallin with mature α-synuclein amyloid fibrils inhibits their elongation. *Biophys. J.* **98**, 843–851 (2010).
48. Raman, B. *et al.* αB-crystallin, a small heat-shock protein, prevents the amyloid fibril growth of an amyloid β-peptide and β2-microglobulin. *Biochem. J.* **392**, 573–581 (2005).
49. Hochberg, G.K.A. *et al.* The structured core domain of αB-crystallin can prevent amyloid fibrillation and associated toxicity. *Proc. Natl. Acad. Sci. USA* **111**, E1562–E1570 (2014).

Acknowledgments

We are grateful for funding from the UK Medical Research Council (COEN grant MR/J006904/1 to D.C.R. and C.J.O'K.), the Wellcome Trust (Principal Fellowship 095317/Z/11/Z), to D.C.R., and strategic award 100140), the National Institute of Health Research Biomedical Research Unit in Dementia at Addenbrooke's Hospital, the TAMAHUD project (European Community FP6 grant no. 03472 under the Thematic Call LSH-2005-2.1.3-8 "Early markers and new targets for neurodegenerative diseases") and the NEUROMICS project (European Community Seventh Framework Programme under grant agreement no. 2012-305121). We thank J.L. Marsh, N. Perrimon and the Vienna *Drosophila* RNAi Center for fly stocks; M. Renna and S. Luo for helpful comments; M. Lichtenberg for help with flow cytometry assays; F. Siddiqi and M. Garcia-Arencibia for help with primary cultures; W. Fecke for advice and assistance; and S. Gotta for help in high-resolution MS analysis of compounds.

Author contributions

M.J.-S. performed most post-screen cell biology experiments. W.L. and S.I. performed the *Drosophila* experiments. M.H. and B.S. performed the cell-based screen. A.F., T.E.D. and C.X performed the zebrafish experiments, and A.F. supervised these. A.T., E.G.-C., V.P.S. and R.S. performed the bioinformatics analyses. F.M. performed the chaperone transcription array experiments and nonradioactive pulse-chase. E.G.-C. and F.H. participated in experimental design of the screen. F.H., G.L., D.D., L.M. and G.P. generated and validated the stable cell lines for the screen. G.M., C.C. and A.N. synthesized and analyzed the compounds. M.A. performed the selection of compound for high-throughput screening and supported the hit-to-lead optimization by *in silico* drug design methodologies. V.P. optimized glutaminyl cyclase enzymatic assays for compound screening. G.L.S. and N.P.C. performed *in vitro* ADME experiments. C.S. provided support for experiments at Siena Biotech. C.O.K. supervised *Drosophila* experiments. G.P. also supervised molecular biology activities at Siena Biotech. A.C. supervised primary screen and chemical biology. D.C.R. supervised cell biology, *Drosophila* and zebrafish experiments. D.C.R and A.C. conceived the project and coordinated work between sites with assistance from G.P., M.J.-S. and D.C.R., and A.C drafted the manuscript, which was commented on by all authors.

Competing financial interests

The authors declare competing financial interests: details accompany the online version of the paper.

Additional information

Supplementary information and chemical compound information is available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to A.C. or D.C.R.

ONLINE METHODS

Assays for validation polyglutamine toxicity modifiers in *Drosophila*. Drosophila *fly stocks*. As a model of polyglutamine toxicity, flies that expressed a protein with 48 glutamines encoded by P{UAS-Q48.myc/flag}31 (ref. 13) in eyes under control of the GMR-Gal4 driver P{GAL4-ninaE.GMR}12 (ref. 50) (Q48) were used. Fly orthologs to the genes identified in the cell screen were selected by performing reciprocal BLASTP and cross checking with databases including HomoloGene (http://www.ncbi.nlm.nih.gov/homologene/), GeneCards (http://www.genecards.org/) and Ensembl (http://www.ensembl.org/index.html/). The RNAi lines corresponding to the identified genes were obtained from Vienna *Drosophila* RNAi Center (VDRC; http://stockcenter.vdrc.at/control/main/).

UAS-Q48.myc/flag was a generous gift from J.L. Marsh (University California, Irvine)[13], and UAS-Httex1-Q46-eGFP was a generous gift from N. Perrimon (Harvard Medical School)[19]. Fly lines that are not referenced here are documented in FlyBase (http://www.flybase.org/). All fly crosses and experiments were performed at 25 °C.

Drosophila RNAi screen. Five virgins of genotype *w*; GMR-GAL4; UAS-Q48.myc/flag (Q48) were crossed to males carrying each UAS-RNAi (GD- and KK-RNAi collections, VDRC, http://stockcenter.vdrc.at/control/main/). Genetic background was controlled by crossing *w*; GMR-GAL4; UAS-Q48.myc/flag females to w^{1118} males that share the same genetic background (VDRC stock number 60000 for the GD-RNAi lines and 60100 for the KK-RNAi lines). For *Glutaminyl cyclase* (CG32412), the GD-RNAi line 38277 and the KK-RNAi line 106341 were used. For *iso Glutaminyl cyclase* (CG5976), the KK-RNAi line 101533 was used. For GD-RNAi lines, degeneration was determined by scoring the eye depigmentation in the progeny of the above crosses 4 d after eclosion, assessing modification of polyglutamine loss of pigmentation and black necrotic-like spots. As the background in KK-RNAi lines leads to dark eye pigmentation (http://stockcenter.vdrc.at/control/protocols), toxicity was assessed by scoring the presence or absence of black necrotic-like spots in the eyes of 10-d-old flies. Fisher's exact test was performed to compare the numbers of necrotic spot–containing flies in the KK-RNAi crosses with controls using an arbitrary $P < 0.005$ as a statistical cutoff for significance. Eyes were imaged using a Nikon CoolPix 990 digital camera attached to a dissecting microscope.

EGFP expression levels assessed in Drosophila RNAi lines. Western blot analysis was performed using progeny of crosses between virgins of the genotype *w*; GMR-GAL4; UAS-EGFP and males of each VDRC-RNAi line used or background control (VDRC stock number 60100). Fly heads were homogenized in Laemmli sample buffer. Rabbit polyclonal anti-GFP at 1:1,000 (AbCam, Ab6556) and monoclonal anti-β-tubulin at 1:10,000 (Developmental Studies Hybridoma Bank) were used. Blots were scanned using Odyssey Fc Imaging System (LI-COR Biosciences). This validation was initially performed once on each suppressor, and subsequently RNAi lines showing an apparent reduction in EGFP levels were retested using the progeny of three independent crosses. Statistical analysis was performed by Student's two-tailed paired *t*-test between the RNAi lines and the control line.

Pseudopupil assay. Analysis was performed as previously described[17]. Virgins of genotype *elav*-GAL4^{C155}; {GMR-HD.Q120}4.62/TM3 (*elav*-Gal4; GMR-HTT.Q120)[16] were crossed with males carrying the RNAi construct for *Glutaminyl cyclase* (lines $QC^{GD38277}$ or $QC^{KK106341}$) or *iso Glutaminyl cyclase* (line $isoQC^{KK101533}$) and compared to a background control line.

To evaluate the effect of QPCT inhibitors, virgins of genotype *yw*; {GMR-HD.Q120}2.4 (GMR-HTT.Q120) were allowed to mate with w^{1118} control males for 48 h on standard cornmeal food and then transferred on fly food containing the compounds.

The number of rhabdomeres per ommatidium was scored in progeny of the above crosses at 3 d (GMR-Q120) or 4 d (*elav*-Gal4; *GMR*-HTT.Q120) post eclosion. Statistical analysis was performed using one-tailed *t*-test on data from three or four independent experiments, each based on approximately ten individuals for each genotype, scoring 15 ommatidia per eye. When compounds were tested, the analysis was done on females and males of each treatment separately.

Aggregate counting. Virgins of genotype *w*; GMR-GAL4; UAS-Httex1-Q46-eGFP[19] were crossed with males of QPCT UAS-RNAi lines or from the background KK-RNAi control line as all the background controls show similar aggregate scoring. Eye pictures of 18-d-old progeny were taken using a Leica MZ16F microscope connected to a Leica DFC340FX digital camera. For each genotype, GFP punctae indicating aggregate formation was counted using an ImageJ 'Cell Counter' plugin in the eyes of 20 males, a pool of 5 males from four independent crosses. For compound testing, virgins of genotype *w*; GMR-GAL4; UAS-Httex1-Q46-eGFP were crossed with w^{1118} control males, and females of the progeny were scored 15 d post eclosion. The experiment was repeated at least three times, and for each experiment at least four female eyes were scored. An unpaired one-tailed *t*-test was used to determine statistical significance for single comparisons between two groups using GraphPad Prism.

Compound treatment. Flies were reared on food (Instant Fly Food, Philip Harris, Ashby de la Zouch, UK) containing either QPCT inhibitor (50 μM) dissolved in DMSO or DMSO alone. The progeny were flipped every 2 d on fresh food containing the specific inhibitor or DMSO.

Bioinformatics analysis. Ingenuity Pathways Analysis (Ingenuity Systems; http://www.ingenuity.com/) was used to analyze the distribution of siRNAs tested among the different protein classes as well as to determine the canonical pathways associated to the confirmed primary actives.

Assays for validation of polyglutamine toxicity and aggregation modifiers in human cell lines. *Cell culture.* HEK293 (human embryonic kidney) and HeLa (human cervical carcinoma) cells and Atg5-deficient ($Atg5^{-/-}$) mouse embryonic fibroblasts (MEFs) (gift from N. Mizushima (University of Tokyo)) were grown in Dulbecco's modified eagle medium supplemented with 10% FBS, 100 U/ml penicillin-streptomycin and 2 mM L-glutamine at 37 °C in 5% CO_2. UbG76V-GFP–expressing stable HeLa cell line (a kind gift from N.P. Dantuma (Karolinska Institute) was maintained in medium containing 0.5 mg/ml G418.

Isolation and culture of mouse primary cortical neurons. Primary cortical neurons were isolated from C57BL/6 mice (Jackson Laboratories) embryos at E16.5. Briefly, brains were harvested and placed in ice-cold PBS-glucose, where the meninges were removed and the cerebral cortices were dissected. After mechanical dissociation using sterile micropipette tips, dissociated neurons were resuspended in PBS-glucose and collected by centrifugation. Viable cells were seeded on polyornithine-coated 12-multiwell plates. Cells were cultured in Neurobasal medium supplemented with 2 mM glutamine, 200 mM B27 supplement and 1% penicillin-streptomycin at 37 °C in a humidified incubator with 5% CO_2. One-half of the culture medium was changed every 2 d until treatment. After 5 d of culturing *in vitro*, differentiated cortical neurons were infected with lentiviral particles bearing EGFP-Q80 and scramble or QPCT-directed shRNAs. Compounds were added 3 d after EGFP-Q80 viral infection and left for another 24 h. When EGFP-Q80 was expressed together with shRNA, 5–6 d were needed before cultures were fixed in a 2% PFA-7.5% glucose solution.

DNA constructs. Human QPCT (NM_012413) plasmid was purchased from Origene (pCMV6-XL5-QPCT). A C-terminal Flag-tagged QPCT construct was generated by PCR amplification of QPCT cDNA from pCMV6-XL5-QPCT using primers overhanging HindIII and BglII sites and insertion into the pCMV5-Flag in HindIII and BamHI restriction sites, using standard restriction enzyme digestion and ligation procedures. QPCT(E201Q)-Flag was generated using QuikChange II Agilent Site-Directed mutagenesis kit with the following primers: forward, 5′-CTTCTTTGATGGTCAAGAGGCTTTTCTTCACTGG-3′, and reverse, 5′-CCAGTGAAGAAAA GCCTCTTGACCATCAAAGAAG-3′. pcDNA or pCMV5-Flag empty vectors were used as mock controls for pCMV6-XL5-QPCT or QPCT-Flag, respectively.

Constructs expressing the first exon of the *Htt* gene carrying 74 polyglutamines expressed from pEGFP-C1 (Clontech) (EGFP-HTTQ74) or pHM6 (Roche Diagnostics) (HA-HTTQ74) or with only 23 polyglutamines (EGFP-HTTQ23) were described previously[51]. pEGFP-N1-Q57 and pEGFP-N1-Q81 (ref. 23) and pEGFP-C1-A37 (ref. 24) have been previously described. Mutant HTT(1–588)-Flag was provided by M.R. Hayden (University of British Columbia), and mutant HTT(1–548)GFP was generated by S. Luo[25]. 3×Flag-CRYAB construct has been previously described[52]. The pGL3-BIP/GRP78-luciferase construct was kindly provided by M. Renna (University of Cambridge)[53].

Reagents. Chemical compounds used in cell culture were the autophagy inhibitors bafilomycin A1 (400nM, DMSO; 4 h; Millipore) and 3MA (10 mM, 16 h; SIGMA), staurosporine (3 μM) and the proteasome inhibitor MG132 (10 μM). PBD150 was synthesized as described in ref. 26.

Transfection. Cells were transfected in six-well plates with 0.5–1.5 μg of DNA and 5 μl of Lipofectamine (Invitrogen) or TransIT-2020 (Mirus) per well for 4 h in Optimem (GIBCO-BRL) and then incubated in full medium for 48 h. Gene knockdown experiments were performed using ON-TARGETplus SMARTpool siRNA (Dharmacon) for human QPCT, consisting of four siRNAs with the following sequences, which do not target the QPCT-like sequence: CUAUGGGUCUCGACACUUA, GUACCGGUCUUUCUCAAAU, CCUUAAAGACUGUUUCAGA and GGAACUUGCUCGUGCCUUA. For siRNA treatment, either a single transfection protocol using 50 nM siRNA for 48 h or a double transfection protocol which consisted of a first 50-nM siRNA transfection followed by a second 50-nM siRNA transfection after 48 h was used.

Western blotting. Cells were washed once in PBS and harvested on lysis buffer (20 mM Tris-HCl, pH 6.8, 137 nM NaCl, 1 mM EGTA, 1% Triton X-100, 10% glycerol, 1× Roche cOmplete mini protease inhibitor). Equal loading was obtained by protein concentration determination using a Bio-Rad assay followed by resuspension and boiling in Laemmli buffer. Samples were subjected to 12% SDS-PAGE and transferred to a PVDF membrane (Immobilon-P, GE Healthcare). Blots were proved with the following primary antibodies: anti-LC3 (1:2,000; Novus Biologicals, NB100-2220), anti-Hsp70 (1:1,000; Enzo SPA810), anti-CRYAB (1:1,000; Cell Signaling 8851), anti-actin (1:2,000; Sigma, A2066), anti-α-tubulin (1:4,000; T9026, Sigma), anti-Flag epitope (1:2,000; SIGMA, F7425), anti-GFP (1:1,000; Clontech, Living Colors, polyclonal), eIF2α (1:1,000, Abcam 5369) and phospho-S51-eIF2α_(1:1,000, Abcam 32157), GRP78 (1:1,000, Abcam 21685), anti-phospho-ERK (1:1,000, Cell Signaling, 9101), anti-ERK (1:1,000, Cell Signaling, 9102), anti-phospho-CREB (S133) (1:1,000, Cell Signaling 9191), anti-CREB 86B10 (1:1,000, Cell Signaling, 9104), anti-phospho-JNK (1:1,000, Cell Signaling, 9255), anti-JNK (1:1,000, Cell Signaling, 9252). The appropriate anti-mouse or anti-rabbit secondary antibodies were used and visualized using an ECL detection kit (Amersham) or LI-COR Biosciences infrared imager (Odyssey).

Caspase 3/7 activity assay. Cells were seeded in a 96-well plate 24 h before the assay, and 1 μM staurosporine or DMSO was added for the last 8 h. Caspase 3/7 activity was measured by using a luminogenic caspase 3/7 substrate (Caspase 3/7-Glo Assay, Promega) following manufacturer's protocols in a Glomax luminometer (Promega). Protein concentration was determined in each cell lysate, and caspase 3/7 activity was normalized to protein levels.

Coimmunoprecipitation assays. Assays were performed as previously described[25], where HTT(1–588)Flag(Q138) and HTT(1–548)GFP(Q138) were expressed in HeLa cells together with QPCT plasmid for 48 h or treated with 25 μM SEN177 for 24 h. Cells were lysed in buffer B containing 10 mM Tris, pH 7.4, 150 mM NaCl, 1 mM EDTA pH8, 1% Triton and 1× Roche complete mini protease inhibitor for 20 min on ice, followed by centrifugation at 13,000 r.p.m. for 10 min. Five hundred micrograms total protein were incubated with primary anti-Flag M2 (Sigma) or anti-GFP (Clontech, Living Colors, polyclonal) at 5 μg/ml overnight at 4 °C. Protein G Dynabeads (Life Technologies) were added and incubated for further 2 h. Beads were washed three times with buffer B and eluted using 0.1 M glycine, pH 2.5, followed by boiling in Laemmli buffer. Samples were subjected to western blotting and visualized using LICOR. A fraction of the total lysates was run simultaneously.

Reverse-transcriptase PCR analysis. Total RNA was isolated from cell pellets using Trizol Reagent (Invitrogen) and treated with DNase I, and cDNA synthesis was performed by SuperScript III First-Strand Synthesis System (Invitrogen). Standard conditions were used for cDNA amplification, and PCR products were analyzed by agarose gel electrophoresis and ethidium bromide staining or quantified with real-time PCR. For real-time PCR analysis, the reaction mixture containing cDNA template, primers and SYBR Green PCR Master Mix (Invitrogen) was run in a 7900 Fast Real-time PCR System (Applied Biosystems, Carlsbad, CA). Fold changes in mRNA levels were determined using a standard curve and after normalization to internal control β-actin RNA levels. Primer sequences used in this study are as follows: QPCT, 5′-CATGGCATGGATTTATTGG-3′ and 5′-GACGGTATCAGATCAAAC-3′; QPCT-like, 5′-CAGCGTCTCTGGAGCACTTA-3′ and 5′-GCCTCCA GGAACTTTCTGACT-3; GFP 5′-ACGTAAACGGCCACAAGTTC-3′ and 5′-T TCAGGGTCAGCTTGCCGTA-3′; actin, 5′-AGAAAATCTGGCCCACACC-3′ and 5′-GGGGTGTTGAAGGTCTCAAA-3′; CRYAB, 5′-TCTTGAGCTCA GTGAGTACTGG-3′ and 5′-AGCTCACCAGCAGTTCATGG-3′; and mouse QPCT, 5′-CGACTTGAGCCAATTGCTGA-3′ and 5′-CTTCC GGGTTAAGAGTGCTG-3′.

mRNA isolation from mouse brain. All mouse experiments were performed under appropriate UK Home Office licenses and following institutional procedures. We analyzed samples from N171 mutant HD mice and wild-type littermate controls at 20 weeks. mRNA was extracted from brains homogenized in Trizol (Invitrogen) using an Ultra Turrax homogenizer.

Lentivirus infection. shRNA containing pLKO.1 vectors targeting both mouse and human QPCT (TRCN032432) were obtained from the RNAi Consortium (TRC), and scramble shRNA vector was generated in D. Sabatini's laboratory (Whitehead Institute; available from Addgene, plasmid 1864). Lentiviral plasmids to express Q80-GFP were kindly provided by J. Uney (University of Bristol)[54]. Lentiviral particles were produced and transduced following the RNAi Consortium protocols.

Cell toxicity and aggregation assays. Cells were fixed for 7 min in 4% paraformaldehyde (PFA). For EGFP-tagged constructs, slides were mounted in Citifluor (Citifluor, Ltd.) containing 4′,6-diamidino-2-phenylindole (DAPI; 3 μg/ml; Sigma) and visualized using an Eclipse E600 fluorescence microscope (plan-apo 60×/1.4 oil immersion lens) (Nikon). For detection of HA-tagged constructs, immunofluorescence with an anti-HA (Covance laboratories 1:500) and anti-mouse Alexa 488 secondary antibody (Invitrogen, 1:1,000) was performed followed by mounting in Citifluor-DAPI. We assessed the percentage of transfected cells (EGFP- or HA-positive cells) with at least one aggregate per cell. Apoptotic cell death was determined by assessing the nuclear morphology (nuclei fragmented or condensed) in transfected cells. Slides were blinded, and at least 200 transfected cells per slide were scored; each individual experiment was performed in triplicate.

Detection of nascent protein synthesis. Protein synthesis was assessed by metabolic incorporation of AHA (l-azidohomoalanine) into cells transfected with EGFP-HTT(Q23). Briefly, 12 h after HeLa cells transfection, medium was washed and replaced with l-methionine/l-cysteine–free medium and treated with DMSO or SEN177 (50 μM) for 1 h before addition of AHA (l-azidohomoalanine) to the medium and collection of cells every 2 h. Labeled protein was detected by western blotting after performing Click-IT protein detection assay (Life Technologies) using biotin, following manufacturer protocols.

Luciferase reporter assay. Cells were transfected with 1 μg of GRP78-luciferase (firefly) reporter construct and 50 ng of *Renilla* luciferase (pRL-TK) as an internal transfection efficiency control. Cells were collected in Passive lysis buffer, and luciferase activity was measured using the Dual-luciferase Reporter Assay System (Promega) following the manufacturer's protocol in a Glomax Luminometer (Promega). GRP78-luciferase relative activity was calculated relative to the *Renilla* luciferase transfection efficiency control activity for each sample; experiments were performed in triplicate.

Statistical analysis. Quantification of immunoblots was performed by densitometric analysis using the ImageJ software or the LI-COR Biosciences infrared imager software and normalized to loading control (actin or tubulin, as indicated in each figure legend). The *P* values were determined by two-tailed Student's *t*-test.

Aggregates were counted in at least 200 cells per slide (with the observer blinded to their identity), and the percentage was calculated relative to control conditions. *P* values were determined by unpaired two-tailed Student's *t*-test. All experiments were done at least three times in triplicate, and a representative blot or graph from a triplicate experiment is shown unless indicated.

Heat shock proteins and chaperone PCR array. The Human Heat Shock Proteins and Chaperones RT2 Profiler PCR Array (SABiosciences, Frederick, MD) was used to study the expression profile of 84 heat shock proteins according to the manufacturer's procedure. Briefly, total RNA was extracted from cells transfected with HTT(Q74)GFP treated with DMSO or 25 μM of SEN177 inhibitor for 24 h using Trizol (Invitrogen) and was further purified using RNeasy mini kit with on-column DNAse digest (Qiagen). cDNA was then synthesized using an RT2 First Strand kit (SABiosciences), and real-time PCR was performed using a 7900HT fast real-time PCR system (Applied Biosciences). Data were analyzed with RT2 profiler PCR array data analysis software version 3.5.

Assays for validation of polyglutamine aggregation modifiers in zebrafish. *Maintenance of zebrafish stocks and collection of embryos*. All zebrafish husbandry and experiments were performed in accordance with UK legislation under a license granted by the Home Office and with local ethical approval. Zebrafish were reared under standard conditions[55] on a 14 h light/10 h dark

doi:10.1038/nchembio.1790

cycle. Embryos were collected from natural spawnings, staged according to the established criteria[56] and reared in embryo medium (5 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM Mg_2SO_4 and 5 mM HEPES).

Determination of the maximum-tolerated concentration of compounds in larval zebrafish. Compound exposure experiments were performed on wild-type larvae (TL strain) 2–3 d post-fertilization (d.p.f.). Concentration response assays were performed over log intervals, from 100 nM to 1 mM, to determine the maximum nontoxic concentration (MTC) for subsequent aggregate analysis assays (n = 10 larvae per concentration). Compound exposure experiments were performed in the dark at 28.5 °C.

Measuring aggregate number and rhodopsin protein levels in transgenic HD zebrafish. Aggregate counting and analysis of rod photoreceptor degeneration (photoreceptor number) was performed using heterozygous larvae from Tg (rho:EGFP-HTT71Q)^{cu5} zebrafish[42] (hereafter referred to as transgenic HD zebrafish). Embryos from outcrossed transgenic HD zebrafish were raised in 0.2 mM 1-phenyl-2-thiourea (PTU) from 1–3 d.p.f. to inhibit pigment formation, screened for transgene expression using EGFP fluorescence and then washed twice in the embryo medium to remove PTU. From 3–9 d.p.f., transgenic HD zebrafish larvae were dark-reared in embryo medium alone or in embryo medium containing either DMSO, 1 mM SEN177, 100 µM SEN180 or 100 µM SEN817. Embryo medium and compounds were replenished daily. Larvae were anaesthetized by immersion in 0.2 mg/ml 3-amino benzoic acid ethyl ester (MS222), then fixed for aggregate counting at 7 d.p.f. or for photoreceptor analysis at 9 d.p.f. Anaesthetized larvae were fixed using 4% paraformaldehyde (PFA) in PBS at 4 °C. Larvae were washed briefly in PBS, allowed to equilibrate in 30% sucrose in PBS then embedded in OCT medium (Tissue-Tek) and frozen on dry ice for subsequent cryosectioning. Sections were cut at 10-µm thickness using a cryostat (Bright Instruments). For aggregate counting, sections were mounted in 50% glycerol in PBS, and the total number of GFP-positive aggregates were counted over 100 µm of the central retina on either side of the optic nerve head, and mean values were calculated (n = 5 fish (10 eyes)) for each treatment group. For quantification of photoreceptor number, the GFP-positive area of the central retina was quantified using image thresholding and area analysis in ImageJ ($n \geq 5$ fish (10 eyes) for each treatment group). To demonstrate that loss of GFP corresponds to loss of photoreceptors, sections were stained with anti-rhodopsin (1D1) antibody (a kind gift from P. Linser, University of Florida)[57] and mounted using VectaShield hard set mounting medium (Vector Laboratories). Sections were viewed and representative images were acquired using a GX Optical LED fluorescent microscope, GXCAM3.3 digital camera and GX Capture software.

50. Freeman, M. Reiterative use of the EGF receptor triggers differentiation of all cell types in the *Drosophila* eye. *Cell* **87**, 651–660 (1996).
51. Narain, Y., Wyttenbach, A., Rankin, J., Furlong, R.A. & Rubinsztein, D.C. A molecular investigation of true dominance in Huntington's disease. *J. Med. Genet.* **36**, 739–746 (1999).
52. D'Agostino, M. *et al.* The cytosolic chaperone α-crystallin B rescues folding and compartmentalization of misfolded multispan transmembrane proteins. *J. Cell Sci.* **126**, 4160–4172 (2013).
53. Renna, M., Caporaso, M.G., Bonatti, S., Kaufman, R.J. & Remondelli, P. Regulation of *ERGIC-53* gene transcription in response to endoplasmic reticulum stress. *J. Biol. Chem.* **282**, 22499–22512 (2007).
54. Howarth, J.L. *et al.* Hsp40 molecules that target to the ubiquitin-proteasome system decrease inclusion formation in models of polyglutamine disease. *Mol. Ther.* **15**, 1100–1105 (2007).
55. Westerfield, M. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio rerio)* (University of Oregon Press, 1995).
56. Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B. & Schilling, T.F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310 (1995).
57. Hyatt, G.A., Schmitt, E.A., Fadool, J.M. & Dowling, J.E. Retinoic acid alters photoreceptor development *in vivo*. *Proc. Natl. Acad. Sci. USA* **93**, 13298–13303 (1996).

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2010, Article ID baq019, doi:10.1093/database/baq019

Original article

GPCRs, G-proteins, effectors and their interactions: human-gpDB, a database employing visualization tools and data integration techniques

Venkata P. Satagopam[1], Margarita C. Theodoropoulou[2], Christos K. Stampolakis[2], Georgios A. Pavlopoulos[1,3], Nikolaos C. Papandreou[2], Pantelis G. Bagos[3], Reinhard Schneider[1] and Stavros J. Hamodrakas[2,*]

[1]Structural and Computational Biology Unit, EMBL, Meyerhofstrasse 1, Heidelberg D69117, Germany, [2]Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01 and [3]Department of Computer Science and Biomedical Informatics, University of Central Greece, Lamia 35 100, Greece

***Corresponding author:** Tel: +30 210 727 4931; Fax: +30 210 727 4254. Email: shamodr@biol.uoa.gr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Submitted 1 June 2010; Revised 21 July 2010; Accepted 22 July 2010

G-protein coupled receptors (GPCRs) are a major family of membrane receptors in eukaryotic cells. They play a crucial role in the communication of a cell with the environment. Ligands bind to GPCRs on the outside of the cell, activating them by causing a conformational change, and allowing them to bind to G-proteins. Through their interaction with G-proteins, several effector molecules are activated leading to many kinds of cellular and physiological responses. The great importance of GPCRs and their corresponding signal transduction pathways is indicated by the fact that they take part in many diverse disease processes and that a large part of efforts towards drug development today is focused on them. We present Human-gpDB, a database which currently holds information about 713 human GPCRs, 36 human G-proteins and 99 human effectors. The collection of information about the interactions between these molecules was done manually and the current version of Human-gpDB holds information for about 1663 connections between GPCRs and G-proteins and 1618 connections between G-proteins and effectors. Major advantages of Human-gpDB are the integration of several external data sources and the support of advanced visualization techniques. Human-gpDB is a simple, yet a powerful tool for researchers in the life sciences field as it integrates an up-to-date, carefully curated collection of human GPCRs, G-proteins, effectors and their interactions. The database may be a reference guide for medical and pharmaceutical research, especially in the areas of understanding human diseases and chemical and drug discovery.

Database URLs: http://schneider.embl.de/human_gpdb; http://bioinformatics.biol.uoa.gr/human_gpdb/

Background

Signal transduction refers to these cellular processes by which stimuli, either physical or chemical, induce specific cellular responses, through chosen molecular mechanisms. The specificity of a cellular response to a signal depends on the receptor expressed on the target cell.

G-protein coupled receptors (GPCRs) are a very important superfamily of cell membrane receptors in eukaryotic cells. They may interact with both the environment outside and inside the cell and they play a crucial role in receiving stimuli signals from the environment. In response they induce certain cellular responses. GPCRs have a characteristic

Database, Vol. 2010, Article ID baq019, doi:10.1093/database/baq019

structure comprised of seven transmembrane-spanning α-helices, an extracellular N terminus, an intracellular C terminus and three interhelical loops on each side of the membrane (1). Several classification systems have been used for this superfamily categorization. The most frequently system used (2,3) classifies GPCRs in six classes, based on their sequence homology and their functional similarity. These are: Class A or 1 Rhodopsin-like, Class B or 2 Secretin receptor family, Class C or three Metabotropic glutamate/pheromone, Class D or four Fungal mating pheromone receptors, Class E or five Cyclic AMP receptors and Class F or six Frizzled/Smoothened like, first presented by (4). GPCRs that are not yet characterized or classified are called orphan GPCRs. Furthermore, a number of putative classes of some newly discovered GPCRs exist, whose nomenclature has not yet been accepted by the scientific community (5,6). Ligands bind to GPCRs on the outside of the cell, activating the GPCRs by causing a conformational change, and allowing them to bind to G-proteins (7).

G-proteins form heterotrimers composed of Gα, Gβ and Gγ subunits, which possess a binding site for a GTP or a GDP molecule. They are characterized by their α-subunits, which are further grouped into the Gαs, Gαi/o, Gαq and Gα12 families (8). The stimulation of GPCRs leads to the activation of G-proteins, which dissociate into their Gα and Gβγ subunits. The subunits then activate several effector molecules that lead to many kinds of cellular and physiological responses (1). Effectors form a diverse group of proteins through their interaction with G-proteins that act either as secondary messengers, or lead directly to a cellular and physiological response. Many proteins such as tubulins, adenylate cyclases, ion channels and others act as effectors (5). GPCRs, G-proteins, effectors and their interactions compose one of the main mechanisms for signal transduction and activation or deactivation of pathways within the cell. A large part of efforts towards drug development today is focused on finding chemicals that affect the ability of ligands to bind to GPCRs (9) either to inhibit or accelerate certain cellular processes. GPCRs play a crucial role in a wide range of human diseases.

Human-gpDB was developed as a tool for integrating together human GPCRs, G-proteins and effectors. It does not only present how they interact with each other but it also reveals information about the pathways they are involved in. Human-gpDB was built as a useful tool for drug research and as a platform that reveals new patterns for therapeutic paths.

Construction and content

Data integration

Our initial step was to collect sequence information individually about human GPCRs, G-proteins and effectors from the UniProt/Swissprot database (10). The entries were acquired using suitable scripts written in Perl in order to parse the DE (description), the GN (gene) or the DR (database cross reference) field of a respective database entry. The data sets were then checked manually in order to eliminate duplicates. Our main goal was to include database unique entries from UniProt/SwissProt. Perl scripts were used for data manipulation.

For each of the three sets, the next step was to isolate and keep these proteins that have at least one connection with another protein, a GPCR with a G-protein and a G-protein with an effector and vice versa. For the extraction of information concerning the connections between human GPCRs, G-proteins and effectors, an extensive literature search was performed attempting to detect terms that co-occur in the same abstract and are biologically related. None of the available text mining engines was used in order to avoid false negative results and to increase the reliability of the results that are presented. Currently the database holds 1663 connections between GPCRs and G-proteins and 1618 connections between G-proteins and effectors. In addition, PubMed reference articles that provide the literature support for each recorded connection are included.

Efficient ways were used to show how G-proteins and effectors might be categorized into families, subfamilies and types. G-proteins' classification is the most commonly used and is based on their α subunits sequence homology, while effectors' classification is based on their function. Many different classifications exist regarding GPCRs; the classification used here was according to the IUPHAR classification (2). All classifications were done manually.

UniProt (10) identifiers were used as starting points to integrate the Human-gpDB with various external data sources. The systems that were used to help us with this integration were ENSEMBL (11), BioMart (12) and SRS (13). For each of the proteins, information about the name, the sequence, the description, the family and the subfamily it belongs to, together with the full record coming from the Dasty2 DAS client (14) was collected. Furthermore, a collection of a vast variety of linked identifiers was obtained, to enrich the information for each protein. Thus, information comes from various databases like for example Uniprot (10), RefSeq Proteins (15), Entrez Proteins (16) and Ensembl Proteins databases (11). Information about the gene location and its properties are provided from Ensembl (11), EMBL (17), EntrezGene (18), RefSeq DNA (15) and UniGene databases. Domain links are provided for Smart (19), InterPro (20) and Pfam (21,22) databases. Structures are linked to PDB (23), HSSP (24) and PSSH (25) databases. Information about diseases comes from the OMIM (26) database and information about protein function from the Gene Ontology (27) database. Chemical information is provided by HMDB (28) and

DrugBank (29) and pathway information comes from KEGG (30), Panther (31) and Reactome (32) databases. Drugs related to three categories of molecules of Human-gpDB (GPCRs, G-proteins, effectors) were collected from DrugBank (29), Madator (33) and AKS2 (34). All available interactions between drugs and the three categories of molecules of the database are presented to the user using visualization tools (see 'Visualization' section for more details). All of the aforementioned information was collected for each protein using ENSEMBL, BioMart and SRS and the results were stored in a MySQL database.

Going one step further, Human-gpDB is not only linked to other sources but it also comes with some analysis features to make the data integration part more useful. Therefore, Human-gpDB comes with domain architecture analysis, protein–protein/protein–chemical interactions and pathway enrichment that will be explained in detail in the 'Utility and Discussion' section below.

Implementation details

All the results are delivered to the user through a web application. The database was implemented in mySQL whereas for the graphical user interface (GUI), the HTML language was used. The dynamic parts of the interface, as for example the auto-complete forms and the advanced search capabilities, are supported by Javascript. The communication between the GUI and the database to extract information from Human-gpDB was achieved with the use of CGI scripts and all the calculations were performed using Perl. The entire application is set up behind an apache web server. The pop-up window that provides links to external or internal data sources was implemented with the help of the Overlib library. For the visualization, the Arena3D standalone Java application (35) was used to support the projections of large-scale networks whereas for smaller networks the Medusa Java applet (36) visualization tool was used. Figure 1 shows an overview of Human-gpDB web application.

Visualization

The visualization module in Human-gpDB was designed in such a way, in order to give maximum flexibility to the user to visualize the interactions between different knowledge domains (GPCRs, G-proteins, effectors, drugs related to GPCRs, drugs related to G-proteins and drugs related to effectors) at different levels (depths) by taking advantage of the hierarchical categories of GPCRs, G-proteins and effectors.

Medusa, a 2D visualization tool (36), was used to graphically visualize interaction partner bioentities such as proteins or drugs as they come from the DrugBank, MATADOR and AKS2 for each of the GPCR, G-protein and effector proteins. The newer version of Medusa tool that supports the current version of Human-gpDB is now more interactive and many layout algorithms are implemented that make the networks much more informative and the extraction of the biological knowledge easier. Like Arena3D, Medusa comes with a set of layout algorithms that are able to minimize the crossovers between the nodes and make the network visually simpler. Medusa is currently provided as a java applet.

To show either the whole network consisting of drugs, GPCRs, G-proteins, effector proteins and their interactions or some large scale sub-networks, the Arena3D standalone visualization tool (35) was used. According to Arena3D, drugs, GPCRs, G-protein and effectors were separated onto four different layers following a multi-layer graph concept, a stack of 2D networks. Arena3D among others currently comes with a clustering layout algorithm that is able to visualize very efficiently predefined distinct clusters. The separation of the clusters is done by placing the nodes that belong to the same cluster together either in 2D or 3D groups. The rich color scheme helps the user to immediately recognize which node belongs to which cluster since nodes that belong to the same cluster are colored similarly. Taking advantage of this functionality, GPCRs, G-proteins and effectors were clustered individually onto their different layers into families and subfamilies according to their properties as mentioned in the previous section, whereas drugs were not clustered. Arena3D is highly interactive and gives the opportunity to the user to isolate either individual,or a set of, connections that reaches his interests for a more focused research. Users can visually highlight and observe patterns that can be easier processed by the user. Such an example could be the answer to a question like whether a protein targets a specific protein family or not. In this version of Human-gpDB, pre-generated input files for the interaction networks are available for download and can be used as input files for the Arena3D application, since it is currently available as a standalone application only and not as a web based tool.

To visually show the domain structure, a static HTML view was implemented. For the pathway visualization, the KEGG schemas are given as they come from the KEGG database. The parts of the pathways that the proteins of the Human-gpDB are involved in are then highlighted to allow the users to easier mark and distinguish the signal transition paths that each protein gets involved in. This functionality is also a strong point of Human-gpDB since researchers can immediately see the influence of a specific protein on a pathway.

Utility and discussion

Browse section

In this section, information about the three distinct categories according to the protein type is presented. These are

Database, Vol. 2010, Article ID baq019, doi:10.1093/database/baq019

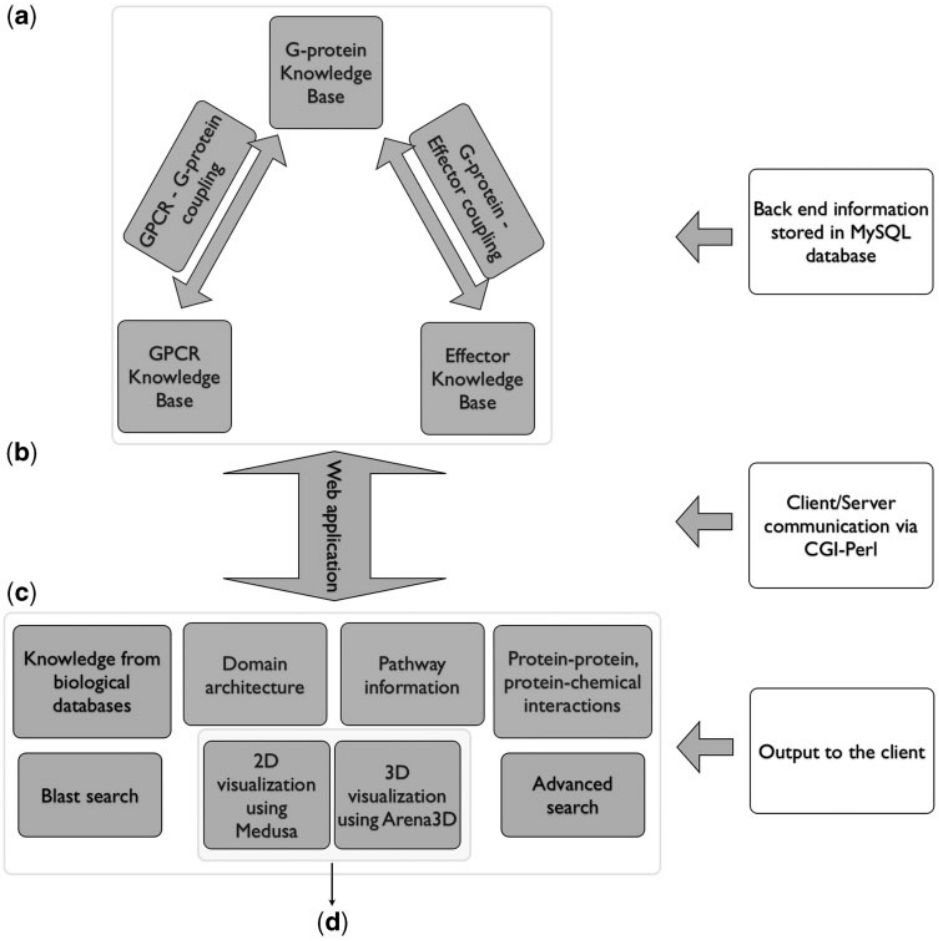


Figure 1. Over view of Human-gpDB web application. (**a**) Backend of the application consisting of manually collected information regarding GPCRs, G-proteins, effectors and their interactions as well as a wide range of publicly available information for each of these proteins stored in a MySQL database ('Data Integration' section for more details). (**b**) A CGI-Perl script handles the communication between the client and the server and (**c**) provides a wide range of information as output to the user (see 'Utility' section for more details). The system can be coupled with (**d**) a 2D visualization tool (Medusa) and a 3D visualization tool (Arena3D), which allows the easy visualization of the relationships between GPCRs, G-Proteins, effectors and the drugs (see 'Visualization' sections for more details).

the 'GPCRs', 'G-proteins' and 'effectors' respectively. Each category of proteins is further sub-divided into classes, families, sub-families hierarchically. One can easily navigate the results by choosing the respective category and then by following the '+' (expand) and the '−' (collapse) button to navigate through the different levels of the hierarchy. Once the user fully expands these categories and goes to the lowest levels that contain the proteins, s/he can see some information about these proteins like the name, the description, the gene names, hyper links to UniProt and links to the relevant entry page, a summary sheet and visualization module (node level).

Entry page. The entry page of each protein contains information about the class, the family, the subfamily that each protein belongs to according to the IUPHAR (2) classification, the sequence and a variety of cross references to various available databases like the UniProt, EMBL, InterPro, PRODOM, GENEID, MIM, PRINTS and PFAM. In the entry page, links about proteins that interact with a selected protein in the form of a sorted list and link to node level visualization module. Redirections to the literature (PubMed) that reveal the evidence of the recorded interaction are also given, so that users can observe the biological relevance of each connection and see where the interaction comes from.

Summary sheet. It provides information about each individual protein coming from a vast variety of publicly available databases. The collected information is further

split into different sections; these are labeled as genes, proteins, domains, structural features, diseases, gene ontology (GO), pathways, chemicals and orthology. The information of each section can be further explored by giving to researchers the opportunity to go deeper into the volume of knowledge in a more efficient way. All the information is presented under the same web page, which not only makes the exploration easier but it also reduces dramatically the loss of time that someone needs to find the information by browsing and querying the available databases. For example, if one checks the information in the section 'Proteins', he can see that it comes from four resources; these are 'Ensembl Proteins', 'Entrez Proteins', 'RefSeq Proteins' and UniPort. The information in these four resources related to the protein of interest is either complementary to each other or it provides additional evidence. Furthermore, Dasty2 DAS client links are given, to present many of the protein features. Simultaneously, a link to visualization module is provided to visualize the interactions between GPCRs, G-proteins and effectors graphically. In contrast to the entry page, the summary sheet provides to the user all the relevant information about how each protein is linked to other bioentities like genes, structures, diseases, pathways, domains, chemicals, etc. The links that are provided were explained in a previous paragraph of this section.

Visualization module. Human-gpDB provides a very flexible visualization module, where a user can select knowledge domains of interest, level or depth and also the tool in order to visualize the particular network of user's choice. The user may choose to include all types of molecules (GPCRs, G-proteins, effectors, drugs related to GPCRs, drugs related to G-proteins and drugs related to effectors) in the network or some of them. Also the user can decide the level/depth (Class, Family, Subfamily, Type and Node) in which the visualization will be made. This feature is available for GPCRs, G-proteins, effectors, but not for drugs. Two visualization tools are offered as options: Medusa (36), for 2D representations, and Arena 3D (35), for 3D visualization. Drugs are not classified and thus they appear as individual nodes connected to the respective type of molecule (GPCRs, G-proteins and effectors). The main reason for giving the possibility to visualize the interactions at these levels is because otherwise the information gets overcrowded and it becomes difficult to clearly see the information. Arena3D software was chosen to overcome the problem of the 2D space limitations for visualizing larger scale networks. These networks consist of hundreds of nodes and hundreds of connections. The main feature of Arena3D is that it utilizes 3D space to project the data. Like Medusa, it also comes with efficient algorithms to minimize the crossovers between the connections so that the network becomes more informative. The Arena3D tool was used to visualize the interactions between Drugs, GPCRs, G-proteins and effectors for any selected category and any protein level. Simultaneously, the four different molecule categories were separated onto four different 2D layers by following a multi-layered graph representation. These are the Drug, the GPCR, the G-protein and the effector layers. Arena3D does not only visualize the nodes and the edges of the network but it can also very efficiently visualize precalculated clusters, which in the case of Human-gpDB, represent the subfamilies of the proteins (see 'Visualization' section for more details).

Pop-up window. While a user navigates through the data in the browse section, he/she may further answer questions that refer to a set of proteins that belong to a specific level of the hierarchy and not only to individual ones. A researcher for example, might want to see some information about the whole B or C class of the GPCRs or the 'Gamma-aminobutyric' acid receptor subfamily. This way, a user may explore the biological knowledge related to each category and sub category by following the hyperlinks provided by a pop-up window after clicking on highlighted names. The information that a particular user may retrieve about a set of proteins is explained below.

Knowledge from biological databases. Biological information from different databases for all of the members of the selected category and sub category are displayed in a table view. A variety of sorting choices is provided so that researchers can sort the target proteins according to their names, their description, the family, the class or the subfamily that they belong to if any.

Domain architecture. A static comparative domain architecture view in order to detect patterns and investigate if the selected category possesses specific structural features is provided. Domain information of each protein was collected from the SMART database and a HTML based visualization tool was developed, to display this knowledge. In this view, each protein is hyperlinked to the SMART and ENSEMBL databases. Mouse actions over a specific domain allow the user to interact with the GUI interactively and get further information about the selected domain.

Pathway information. Each protein was mapped to KEGG pathways in order to find which pathways are enriched for the selected category. The results are provided in a tabular view accompanied by a KEGG pathway identifier, a pathway name and a list of proteins involved in that pathway. Each pathway is hyperlinked to a red flag, which ultimately displays the pathway, with the proteins from the selected category highlighted in red color.

Database, Vol. 2010, Article ID baq019, doi:10.1093/database/baq019

Protein–protein and protein–chemical interactions. Information about protein–protein and protein–chemical interactions are provided by the STITCH database (37). The collection of this information was done through the available API. The generated networks do not only include proteins stored in Human-gpDB database but also proteins that are recorded in the STITCH database. These proteins do not necessarily belong to one of the three categories of proteins that Human-gpDB holds. The STITCH database (37) goes one step further by providing interacting proteins that were found experimentally besides proteins that co-occur in the literature. Together with the variety of information that STITCH database holds, this is the main reason why this tool was selected to provide to the user the relevant protein–protein and protein–chemical interaction networks.

Visualization. This feature of the pop-up window redirects the user to the Visualization module referring to the particular level of the tree [see Visualization module and 'Visualization' section (main manuscript) for more details].

Blast search section

This feature of Human-gpDB database gives the opportunity to the users to search for homologies by providing one or more protein sequences in Fasta format. Wu-BLAST (38) was used to align a given set of sequences against the selected protein categories of Human-gpDB like for example against the GPCR category. The user may provide the advanced blast options in order to narrow down the search results. The results are then grouped according to their category and sorted by significance. Each result that is found is then hyperlinked to an entry page, summary sheet and node level visualization module. Here, the user can see the alignments as well.

Advanced search section

This feature gives users the option to search the given fields in the database. The user can enter any word in one or more of the available boxes under the name: Gene/Protein, Class, Family, Subfamily, Type, Description and Function. Expressions in separate search fields are combined with the AND operator, so every entry of the result set will satisfy the expressions of all the search fields the user has chosen. The user has the option to choose whether the query will be performed against the GPCRs, the G-proteins or the effectors included in the database.

Results

The database currently holds information about 713 human GPCRs, 36 human G-proteins and 99 human effectors. The collection of information about the interactions between these molecules was done manually and the current status of Human-gpDB reveals information about 1663 connections between GPCRs and G-proteins and 1618 connections between G-proteins and effectors. GPCRs are categorized in four classes. Table 1 shows the number of families and subfamilies in each GPCR class, while Table 2 shows the distribution of GPCRs' subfamilies based on the number of Gα families with which they interact. G-proteins are categorized in Gα, Gβ and Gγ groups. Gα consists of four respective families, as described initially in the 'Backround' section. From the 36 human G-proteins, 17 are characterized as Gα, 7 as Gβ and 12 as Gγ. Effectors are categorized in 20 families, 29 subfamilies and 63 types based on their biological function (Theodoropoulou, M.C.,

Table 1. Number of families and subfamilies in each GPCR class

GPCRs' Class	No. of GPCR Families	No. of GPCR Subfamilies
Class A	55	640 (422 subfamilies of olfactory receptor)
Class B	6	16
Class C	4	41 (29 subfamilies of taste receptors)
Frizzled/Smoothened	2	11
Total	67	708

Class A is the largest and consists of 55 families and 640 subfamilies (422 subfamilies of olfactory receptors). Class B consists of 6 families and 16 subfamilies. Class C consists of 4 families and 41 subfamilies (29 subfamilies of taste receptors). Frizzled/Smoothened class consists of 2 families and 11 subfamilies.

Table 2. Distribution of GPCR subfamilies based on the number of Gα families with which they interact

Couples with	No. of GPCRs' Subfamilies
1 Gα family	623
2 Gα families	48
3 Gα families	15
All 4 Gα families	1
Unknown coupling	21
Total	708

One subfamily of GPCRs, the TSHR family, couples with members of all four Gα families. Most of the GPCR subfamilies couple with members of one Gα family (623 out of the 708 subfamilies of GPCRs). Fifteen GPCR subfamilies couple with members from 3 Gα families, whereas 48 couple with members from 2 Gα families. Twenty-one GPCR subfamilies do not have known coupling.

Bagos, P.G. and Hamodrakas, S.J., manuscript in preparation). The two most highly populated effectors' families are Ion Channels and Tubulins.

Visualization of the interactions between GPCRs, G-proteins, effectors and drugs together with the rich data integration part is one of the main features of Human-gpDB. As described in the 'Visualization' section Medusa application was used for 2D representation of the networks of interactions. Arena3D was chosen for 3D and more efficient representation of either the whole network of interactions or dense subparts of it. In order for the user to evaluate the visualization tools offered, five chosen different examples of visualization are given in Figures 2–6. Medusa, which is a Java applet, offers the user a first glance of the respective network. However Medusa still has disadvantages compared to Arena3D mainly due to the fact that the visualization it offers is in 2D so the space might be a limiting factor for larger or dense networks.

Case study of human Prostanoid TP receptor's network of interactions

In order to demonstrate the utility of Human-gpDB, a case study of human Prostanoid TP receptor's network of interactions (Figure 5) is presented. Human Prostanoid TP receptor may couple with all four subfamilies (G_q, G_{11}, G_{14} and $G_{15/16}$) of the $G_{q/11}$ family of G-proteins. These subfamilies of G-proteins interact with 21 different types of effectors belonging to 8 different families (Tubulins, PI3/PI4 kinases, Phospholipases C, Ser/Thr protein kinases, TPR repeat proteins, Tyr protein kinases, Guanine nucleotide exchange factors and Ezrin-radixin-moesin-binding phosphoproteins). For this specific receptor 23 different drugs exist. Using information from STITCH (37) about protein–protein and protein–chemical interactions, information regarding mostly known natural or synthetic ligands is presented, complementing the functional role of the receptor. Prostanoid TP receptor is related with a bleeding disorder in cases of defection of the receptor [information retrieved from 188 070 entry of OMIM (26) database]. Regarding the receptor's participation in particular KEGG (30) pathways, the receptor is involved in a calcium-signaling pathway (hsa04020). Using the known coupling preferences of the receptor, a researcher may relate the receptor with other specific KEGG pathways (for example gap junction), in which $G_{q/11}$ G-proteins are known to participate. After launching BLAST against Human-gpDB using human Prostanoid TP receptor as query sequence, Prostanoid FR receptor is the most similar entry. Prostanoid FR receptor interacts with the same subfamilies of G-proteins, shares four mutual ligands and agonists [comparison of the networks presented by STITCH (37)] and two mutual drugs [comparison of the networks presented by Medusa (36)] with Prostanoid TP receptor, and, both receptors

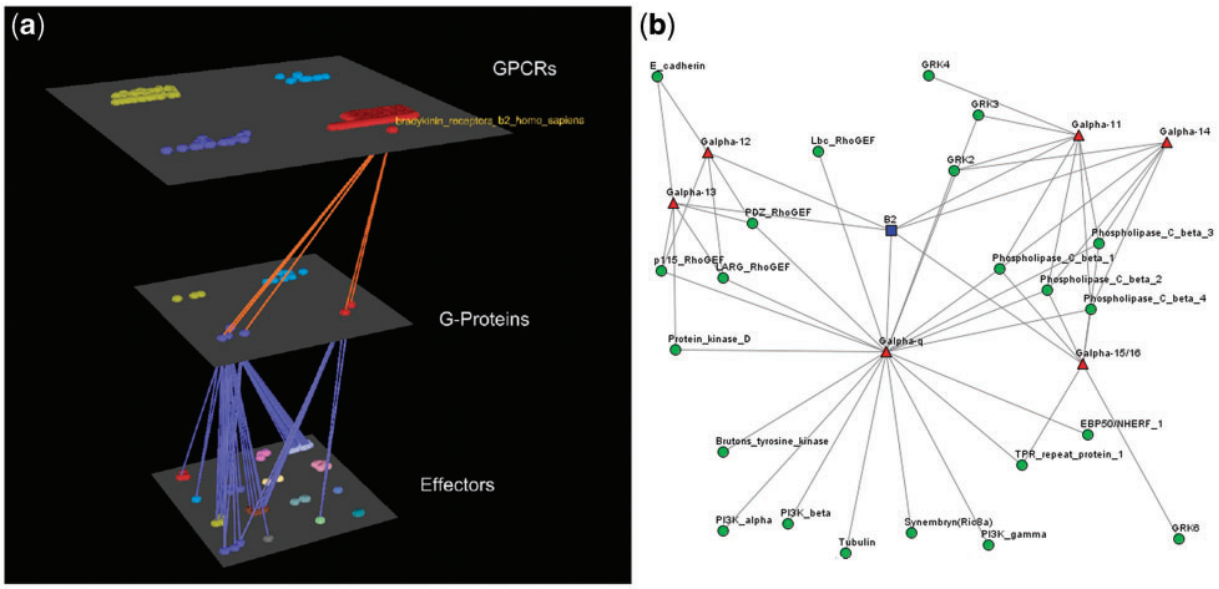


Figure 2. Visualization of human Bradykinin B2 receptor's interactions (**a**) Arena3D Visualization: human Bradykinin B2 receptor targets six different subfamilies of Gα G-Proteins belonging to Gα-q/11 and Gα-12/13 families. The G-Proteins are connected to 22 different types of effectors belonging to nine families. (**b**) Medusa 2D Visualization: human Bradykinin B2 receptor targets Gα G-Proteins belonging to six distinct Gα subfamilies (Gα-q, Gα-11, Gα-14, Gα-15/16, Gα-12 and Gα-13). These Gα G-Proteins interact with 22 types of effectors.

Database, Vol. 2010, Article ID baq019, doi:10.1093/database/baq019

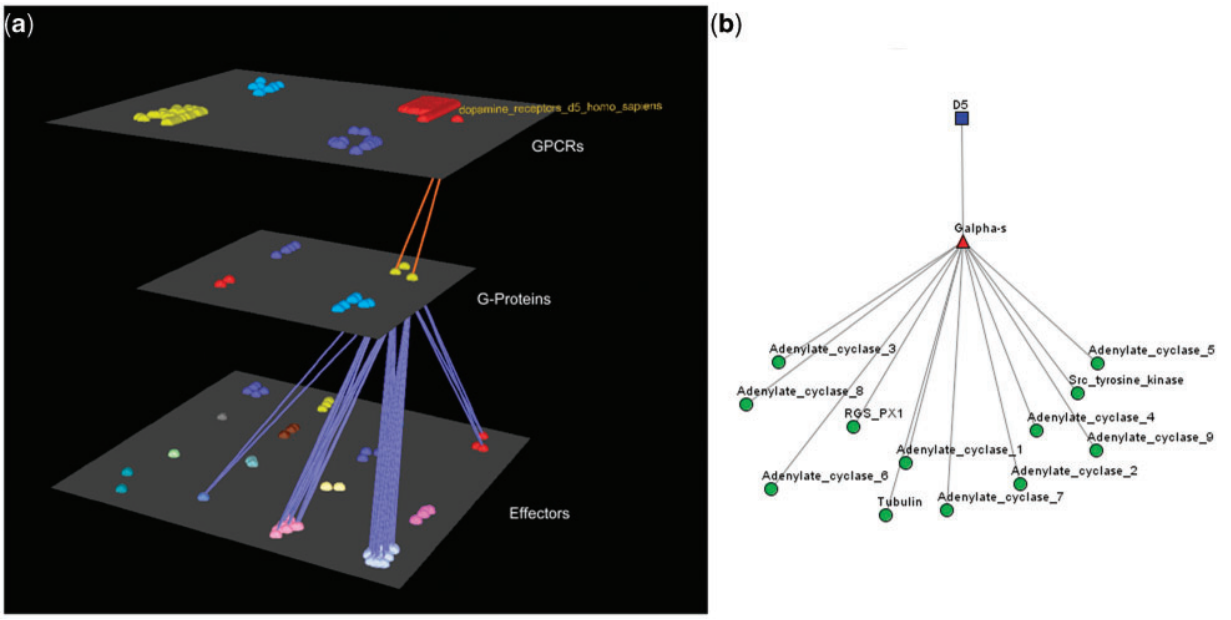


Figure 3. Visualization of human Dopamine D5 receptor's interactions. (**a**) Arena3D Visualization: Human Dopamine D5 receptor targets two Gα-s G-Proteins. The G-Proteins are connected to 12 types of effectors belonging to four specific families. These are: Adenylate cyclases, Regulators of G-Protein signaling, Tyr protein kinases and tubulins. (**b**) Medusa 2D Visualization: human dopamine D5 receptor targets Gα-s G-Proteins, which interact with the following four families of effectors: Adenylate cyclases, Regulators of G-Protein signaling, Tyr protein kinases and tubulins.

participate in the calcium-signaling pathway. In general, the two receptors seem to share similar functions. Based on that, a researcher may assume that chemicals that interact with one of the receptors may also interact with the other, leading to similar results (if a chemical is an agonist for one receptor, probably it will also be for the other one too). Therefore drugs known to affect the function of Prostanoid TP receptor (for this receptor more drugs are known) may also affect Prostanoid FR receptor too. Prostanoid FR receptor is not yet related with any disease (according to OMIM [26]), nevertheless there are 12 drugs related with this receptor. One of these drugs is Latanoprost, which is used for controlling the progression of glaucoma or ocular hypertension by reducing intraocular pressure and is a prostaglandin analogue. Based on the known usage of this drug and also the fact that Prostanoid FR receptor participates in the calcium-signaling pathway, there are indications that this specific receptor may be related with hypertension. As already shown, the combination of different information retrieved from Human-gpDB may help the researchers to design specific experiments by which they will clarify the pathways in which the receptors participate, propose a mechanism for the specific disease in the case of Prostanoid TP receptor, propose a relation between Prostanoid FR receptor and hypertension and/or comprehend the side effects of drugs.

Conclusions

Human-gpDB compared to the previous gpDB databases (39,40) is now richer and focuses only on human GPCRs, G-proteins and effectors. Human-gpDB is not simply a gpDB subset, since it contains more recent data (last update of gpDB was done in March 2008, whereas all data of Human-gpDB were retrieved until December 2010), but it also contains new information concerning the classification of GPCRs (11 new subfamilies were added and all existing subfamilies are classified based on the IUPHAR classification) and also contains interactions between all molecules. It is fully integrated with external data sources by bridging information that did not exist in the previous versions (e.g. drugs and chemicals) and it now comes with a new user-friendly environment supported by advanced visualization techniques. The interface makes the navigation friendlier, the exploration of information more efficient and the extraction of new knowledge easier. Human-gpDB database was built to provide a simple but yet a powerful tool for researchers in the life sciences field as it integrates a current, careful collection of human GPCRs, G-proteins, effectors and their interactions. Human-gpDB uses advanced visualization techniques to make the volume of data more informative and the advanced data integration techniques make Human-gpDB a unique tool, a reference

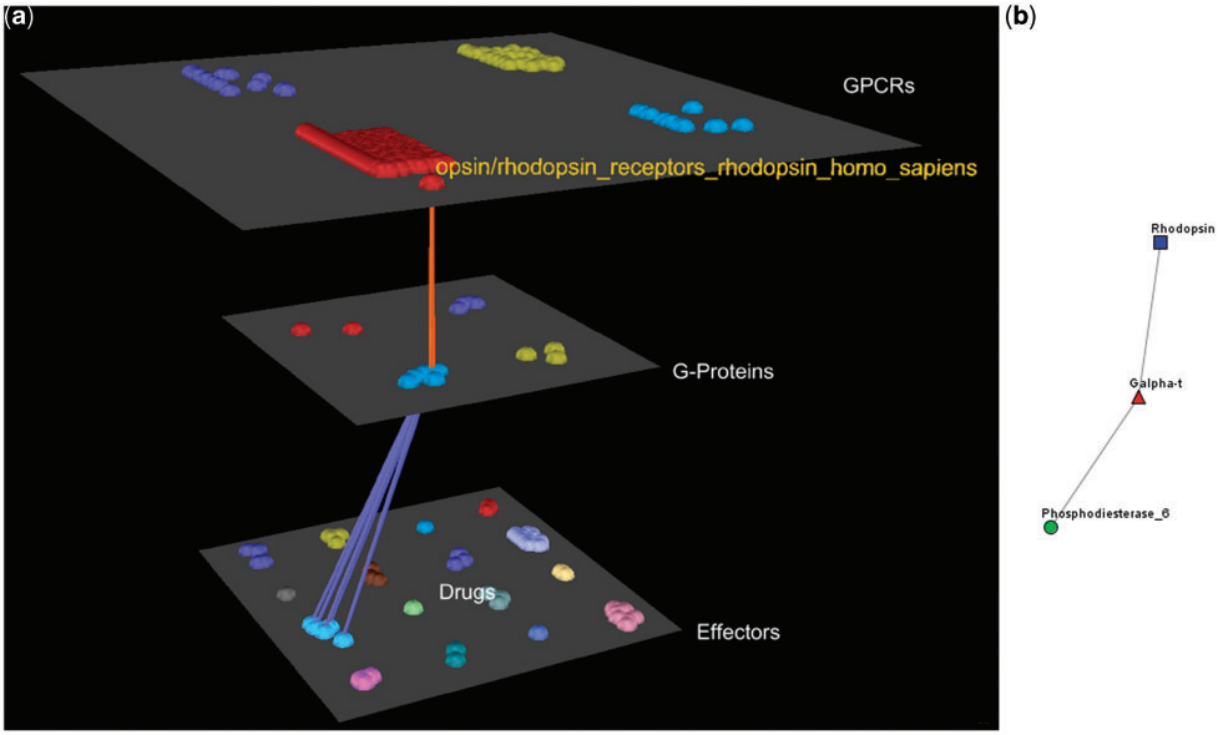


Figure 4. Visualization of human Rhodopsin receptors' interactions. (**a**) Arena3D Visualization: Human Rhodopsin receptor targets three Gα G-Proteins that belong to the Gα-t subfamily. The three G-Proteins interact with five effectors belonging to the Rhodopsin-sensitive cGMP-specific PDEases subfamily and more specifically to Phosphodiesterase 6 type of effectors. (**b**) Medusa 2D Visualization: the Rhodopsin subfamily of the Opsin/Rhodopsin family of the Class A of the GPCRs interacts with the Gα-t subfamily of the G-Proteins which interact with the Rhodopsin-sensitive cGMP-specific PDEases effectors' subfamily.

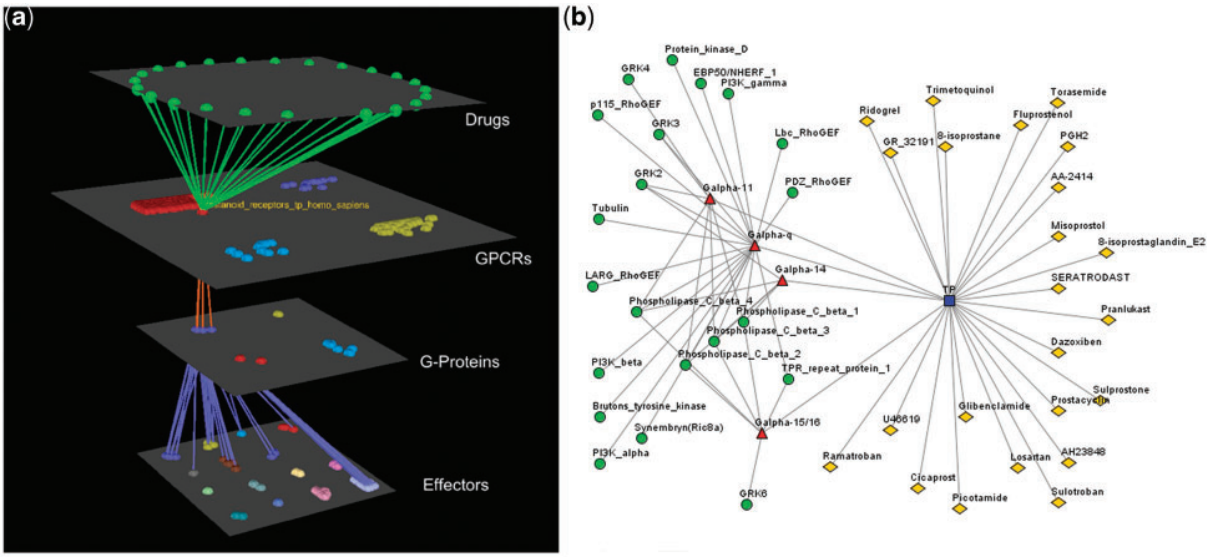


Figure 5. Visualization of human Prostanoid TP receptor's interactions (drugs included). (**a**) Arena3D Visualization: Human Prostanoid TP receptor protein of Class A GPCR family targets four Gα G-Proteins that belong to $G_{q/11}$ family. The G-Proteins are connected to effectors proteins belonging to eight specific families. For this specific receptor 23 different drugs exist. (**b**) Medusa 2D Visualization: Human Prostanoid TP receptor protein targets Gα G-Proteins that belong to $G_{q/11}$ family. These G-Proteins interact with 11 different subfamilies of effectors. For this specific receptor 23 different drugs exist.

Database, Vol. 2010, Article ID baq019, doi:10.1093/database/baq019

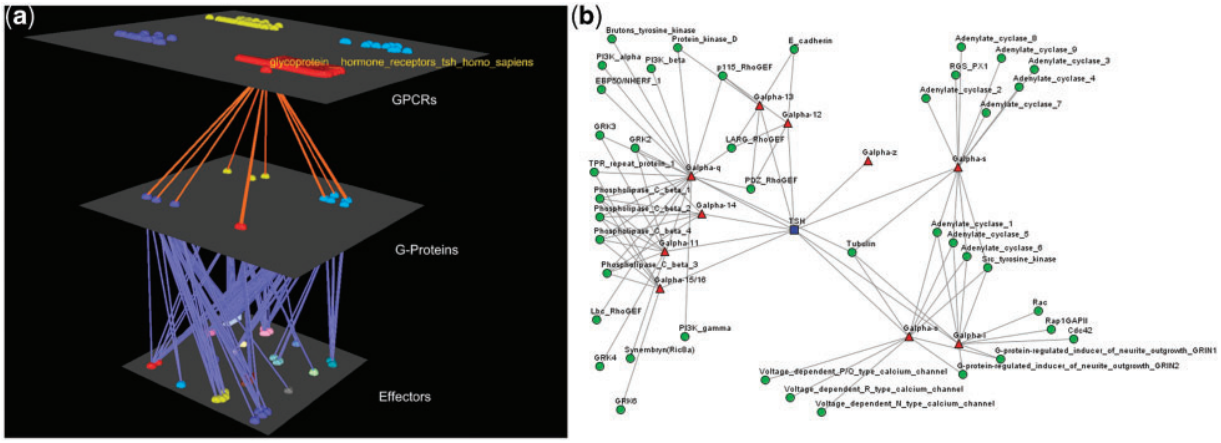


Figure 6. Visualization of human Glucoprotein Hormone TSH receptors' interactions. (**a**) Arena3D Visualization: Human Glucoprotein Hormone TSH receptor targets 13 G-Proteins of 10 different subfamilies, which belong to all four Gα families. These G-proteins target 16 different families of effectors. (**b**) Medusa 2D Visualization: Human Glucoprotein Hormone TSH receptor targets 13 Gα G-Proteins from all four Gα families. More accurately, these Gα G-Proteins belong to 10 respective Gα subfamilies and interact with 19 subfamilies of effectors.

guide in pharmaceutical research and especially in the areas of chemical and drug discovery for human diseases. In the future, the expansion of the current version of the database for other organisms starting from the ones that are evolutionarily closer to Humans is essential.

Availability and requirements

Currently, two Human-gpDB servers are set up, one running at EMBL (http://schneider.embl.de/human_gpdb) and the other running at the Department of Cell Biology and Biophysics of the University of Athens (http://bioinformatics.biol.uoa.gr/human_gpdb/). Both servers hold the same copy of the Human-gpDB database. Concerning the linking to Human-gpDB from external sources, other databases can link to our database by using, for example, the following URLs: http://schneider.embl.de/cgi-bin/human_gpdb.cgi?search=P21918 or http://schneider.embl.de/cgi-bin/human_gpdb.cgi?search=DRD5_HUMAN, based on Uniprot ID or Accession Number.

Acknowledgements

The authors thank the reviewers of this article for their useful criticism. M.C.T. collected the data, V.P.S. developed the web application and data integration, G.A.P. worked on visualization tools, N.C.P., C.K.S., P.G.B. helped in the development of the application, R.S. and S.J.H. supervised the project. All authors are involved in the writing of the article.

Funding

University of Athens, the University of Central Greece and EMBL, Heidelberg. Funding for open access charge: EMBL, Heidelberg.

Conflict of interest. None declared.

References

1. Oldham,W.M. and Hamm,H.E. (2008) Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat. Rev. Mol. Cell. Biol.*, **9**, 60–71.
2. Harmar,A.J., Hills,R.A., Rosser,E.M. *et al.* (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.*, **37**, D680–D685.
3. Horn,F., Bettler,E., Oliveira,L. *et al.* (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
4. Kolakowski,L.F. Jr (1994) GCRDb: a G-protein-coupled receptor database. *Receptors Channels*, **2**, 1–7.
5. Kristiansen,K. (2004) Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol. Ther.*, **103**, 21–80.
6. Pierce,K.L., Premont,R.T. and Lefkowitz,R.J. (2002) Seven-transmembrane receptors. *Nat. Rev. Mol. Cell. Biol.*, **3**, 639–650.
7. McCudden,C.R., Hains,M.D., Kimple,R.J. *et al.* (2005) G-protein signaling: back to the future. *Cell. Mol. Life Sci.*, **62**, 551–577.
8. Cabrera-Vera,T.M., Vanhauwe,J., Thomas,T.O. *et al.* (2003) Insights into G protein structure, function, and regulation. *Endocrine Rev.*, **24**, 765–781.
9. Attwood,T.K. (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol. Sci.*, **22**, 162–165.

10. UniProtConsortium. (2009) The universal protein resource (UniProt) 2009. *Nucleic Acids Res.*, **37**(Database issue), D169–D174.
11. Hubbard,T.J., Aken,B.L., Ayling,S. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**(Database issue), D690–D697.
12. Smedley,D., Haider,S., Ballester,B. *et al*. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
13. Etzold,T. and Verde,G. (1997) Using views for retrieving data from extremely heterogeneous databanks. *Pac. Symp. Biocomput.*134–141
14. Jimenez,R.C., Quinn,A.F., Garcia,A. *et al*. (2008) Dasty2, an Ajax protein DAS client. *Bioinformatics*, **24**, 2119–2121.
15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**(Database issue), D61–D65.
16. Schuler,G.D., Epstein,J.A., Ohkawa,H. *et al*. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
17. Cochrane,G., Aldebert,P., Althorpe,N. *et al*. (2006) EMBL nucleotide sequence database: developments in 2005. *Nucleic Acids Res.*, **34**(Database issue), D10–D15.
18. Maglott,D., Ostell,J., Pruitt,K.D. *et al*. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**(Database issue), D54–D58.
19. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**(Database issue), D229–D232.
20. Mulder,N.J., Apweiler,R., Attwood,T.K. *et al*. (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief. Bioinform.*, **3**, 225–235.
21. Sammut,S.J., Finn,R.D. and Bateman,A. (2008) Pfam 10 years on: 10,000 families and still growing. *Brief. Bioinform.*, **9**, 210–219.
22. Finn,R.D., Tate,J., Mistry,J. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**(Database issue), D281–D288.
23. Berman,H.M., Battistuz,T., Bhat,T.N. *et al*. (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58(Pt 6 No 1)**, 899–907.
24. Sander,C. and Schneider,R. (1993) The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res.*, **21**, 3105–3109.
25. Schafferhans,A., Meyer,J.E. and O'Donoghue,S.I. (2003) The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Res.*, **31**, 494–498.
26. Hamosh,A., Scott,A.F., Amberger,J. *et al*. (2002) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
27. Harris,M.A., Clark,J., Ireland,A. *et al*. (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**(Database issue), D258–D261.
28. Wishart,D.S., Knox,C., Guo,A.C. *et al*. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**(Database issue), D603–D610.
29. Wishart,D.S. (2008) DrugBank and its relevance to pharmacogenomics. *Pharmacogenomics*, **9**, 1155–1162.
30. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
31. Mi,H., Guo,N., Kejariwal,A. *et al*. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.*, **35**(Database issue), D247–D252.
32. Matthews,L., Gopinath,G., Gillespie,M. *et al*. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**(Database issue), D619–D622.
33. Gunther,S., Kuhn,M., Dunkel,M. *et al*. (2008) SuperTarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**(Database issue), D919–D922.
34. Welcome to AKS2. [http://www.bioalma.com/aks2/index.php] (30 July 2010, date last accessed).
35. Pavlopoulos,G.A., O'Donoghue,S.I., Satagopam,V.P. *et al*. (2008) Arena3D: visualization of biological networks in 3D. *BMC Systems Biol.*, **2**, 104.
36. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
37. Kuhn,M., von Mering,C., Campillos,M. *et al*. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**(Database issue), D684–D688.
38. Lopez,R., Silventoinen,V., Robinson,S. *et al*. (2003) WU-Blast2 server at the European bioinformatics institute. *Nucleic Acids Res.*, **31**, 3795–3798.
39. Elefsinioti,A.L., Bagos,P.G., Spyropoulos,I.C. *et al*. (2004) A database for G proteins and their interaction with GPCRs. *BMC bioinformatics*, **5**, 208.
40. Theodoropoulou,M.C., Bagos,P.G., Spyropoulos,I.C. *et al*. (2008) gpDB: a database of GPCRs, G-proteins, effectors and their interactions. *Bioinformatics*, **24**, 1471–1472.

Big Data
Volume 4 Number 2, 2016
Mary Ann Liebert, Inc.
DOI: 10.1089/big.2015.0057

ORIGINAL ARTICLE

Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases

Venkata Satagopam[1,*] Wei Gu[1] Serge Eifes[1,2] Piotr Gawron[1] Marek Ostaszewski[1] Stephan Gebel[1] Adriano Barbosa-Silva[1] Rudi Balling[1] and Reinhard Schneider[1]

Abstract

Translational medicine is a domain turning results of basic life science research into new tools and methods in a clinical environment, for example, as new diagnostics or therapies. Nowadays, the process of translation is supported by large amounts of heterogeneous data ranging from medical data to a whole range of -omics data. It is not only a great opportunity but also a great challenge, as translational medicine big data is difficult to integrate and analyze, and requires the involvement of biomedical experts for the data processing. We show here that visualization and interoperable workflows, combining multiple complex steps, can address at least parts of the challenge. In this article, we present an integrated workflow for exploring, analysis, and interpretation of translational medicine data in the context of human health. Three Web services—tranSMART, a Galaxy Server, and a MINERVA platform—are combined into one big data pipeline. Native visualization capabilities enable the biomedical experts to get a comprehensive overview and control over separate steps of the workflow. The capabilities of tranSMART enable a flexible filtering of multidimensional integrated data sets to create subsets suitable for downstream processing. A Galaxy Server offers visually aided construction of analytical pipelines, with the use of existing or custom components. A MINERVA platform supports the exploration of health and disease-related mechanisms in a contextualized analytical visualization system. We demonstrate the utility of our workflow by illustrating its subsequent steps using an existing data set, for which we propose a filtering scheme, an analytical pipeline, and a corresponding visualization of analytical results. The workflow is available as a sandbox environment, where readers can work with the described setup themselves. Overall, our work shows how visualization and interfacing of big data processing services facilitate exploration, analysis, and interpretation of translational medicine data.

Key words: big data analytics; big data infrastructure design; data acquisition and cleaning; data integration; data mining; disease map

Introduction

Translational medicine capitalizes on advances in basic life sciences to improve clinical research and care. We witness great technological advances in methods characterizing human health and disease, including genetic and environmental factors of our well-being. This is a great opportunity to understand diseases and to find new diagnoses and treatments. However, the progress comes at a cost—translational research data sets nowadays include genomic, imaging, and clinical data sources,[1,2] making them large and heterogeneous. In effect, important steps of the data life cycle in discovery—integration, analysis, and interpretation—are a challenge for biomedical research. Moreover, enabling biomedical experts to efficiently use big data processing pipelines is another challenge.

As translational medicine data become more and more rich and complex, their potential in informing

[1]*Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-Belval, Luxembourg.*
[2]*Information Technology for Translational Medicine (ITTM) S.A., Esch-Belval, Luxembourg.*

*Address correspondence to: *Venkata Satagopam, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, Avenue des Hauts-Fourneaux, Esch-Belval L-4362, Luxembourg*, E-mail: venkata.satagopam@uni.lu

both clinical and basic research grows.[3] With constantly increasing presence of high-throughput molecular profiling, it becomes increasingly important to ensure that data interpretation capabilities follow generation of large-scale biomedical data sets.[4,5] Visualization can support greatly the processing of complex data sets on each of the steps of the data life cycle. This opportunity is actively explored in various domains of biomedical research, including clinical big data[6] or multiscale biomedical ontologies.[7]

Modern translational medicine approaches aim to combine clinical and molecular profiles of the patients to formulate informed hypothesis on the basis of stratified data.[8] Integration of plethora of sources renders these data sets complex and difficult to process. Visualization of such integrated data sets can aid exploration and selection of key dimensions and subsets for downstream analysis. In turn, visually aided data analysis allows to comprehend even complicated workflows and aids interpretation of resulting data.

In this article, we demonstrate a workflow for translational medicine big data, in which visualization is an important component at each step of data processing and exploration. We describe in detail the interfaces allowing the construction of the workflow, followed up by a use case scenario. We conclude with a discussion of the results and an outlook for future development of visualization in biomedical big data exploration.

Related Work

Clinical and molecular (omics) data integration platforms

The rise of personalized medicine and the availability of high-throughput molecular analysis drives the development of storage, analytics, and interpretive methods to enable the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory healthcare.[3,9,10] Key properties of biomedical big data in translational medicine, according to the "5V" classification,[11] besides its volume are variety and veracity. A combination of clinical* and high-throughput molecular profiles ("omics")† creates a very variable heterogeneous data set, where dimensionalities of different data types span several orders of magnitude.[12] Moreover, ensuring veracity, that is, quality, to clinical data is a challenging and time-consuming task.[13,14] This stems from a variety of collection methods, featuring manual data input, nondigital data capture, and nonstandard formats. It needs to be stressed that proper data curation is a mandatory step for accurate analysis of clinical data and proper interpretation of analytical results.

The emergence of big biomedical data sets, covering dozens of thousands of patients,[12] raises questions on infrastructure necessary to host and analyze them. Especially genomic data, generated rapidly due to dropping sequencing costs, pose a problem in terms of storage and analytics. The perspective of cloud computing is postulated as a solution to this challenge, as summarized in recent and extensive reviews.[5,15,16] Nevertheless, due to ethical and legal issues arising in cloud-based scenarios,[17] incorporation of clinical data and processing of sensitive omics are still considered an open question.

Translational medicine platforms integrating clinical and omics data need to ensure a protected environment for sensitive data processing. A number of solutions were developed to address this challenge, as summarized in an excellent review by Canuel et al.[18] Platforms integrating clinical and omics data can be divided into two groups: repositories with an existing infrastructure and solutions requiring deployment. The first group is represented by technologies, such as STRIDE,[19] iDASH,[20] caGRID, and its follow-up, TRIAD[21,22] or BDDS Center.[23] Certain platforms of this group focus on a specific disease, such as cBioPortal[24] or G-DOC[25] for cancer or COPD Knowledge Base[26] for pulmonary dysfunction. The advantage of solutions based on existing computational infrastructure is direct use but at the cost of reduced flexibility in configuration and toolset management. The other group of solutions for translational medicine requires deployment on the user's infrastructure, often requiring substantial storage or high-performance computing (HPC) capabilities. Notable examples in this group are BRISK,[27] tranSMART,[28] and Transmed.[29] Because of their highly configurable nature, such solutions are suitable in projects implicating sensitive data, and where a repository is needed to support ongoing projects, such as in case of longitudinal cohort‡ studies. Informative use cases of such

*__Clinical data:__ Data collected by the characterization of a biomedical research participant by a medical professional, for example, demographics, study-specific questionnaires, or examinations. __Molecular data:__ Data collected by analyzing samples donated by a biomedical research participant using imaging (microscopy) or high-throughput molecular profiling ("omics").

†__"omics":__ Technologies for characterization and quantification of entire pools of biological molecules in a given sample. Data sets generated using omics are highly dimensional, ranging from hundreds to hundreds of thousands of variables per sample. The name "omics" encompasses particular readout methods: genomics (entire genome), transcriptomics (entire gene expression profile), proteomics (entire protein expression profile), metabolomics (entire pool of metabolites), and others.

‡__Cohort:__ A group of people with a shared characteristic. Here, a group of subjects with demographic, clinical, or other characteristics relevant for translational research.

platforms are SHRINE[30] and DARiS,[31] where well-defined demands of clinical research projects drove the design and implementation of infrastructure supporting translational medicine.

Visually aided data exploration is an important component of clinical and omics integration platforms. A notable contributor in this field is the Informatics for Integrating Biology and the Bedside project (i2b2, www.i2b2.org), a scalable framework enabling the use of clinical data for discovery research.[32,33] The i2b2 Hive[34] is a powerful collection of interoperable tools ranging from repository services to basic data conversions provided by i2b2 cells. Importantly, i2b2 Hive does not support directly the analysis of omics data, such as gene expression or whole-genome sequences by itself,[35] but enables key capabilities of clinical data exploration and processing to be used by other platforms.

Bioinformatics workflow management systems

Reusable and interoperable bioinformatics workflows become increasingly important in reproducible analysis of biomedical data and metadata, including clinical, omics, imaging, and sensor data.[36–38] A number of software frameworks were developed to support the scientific community in this goal. In a thorough review and classification of these workflow frameworks, Leipzig[36] groups existing technologies according to their interaction mode into command-line/application programming interface (API) and workbench approaches. The first group includes Snakemake,[39] Yabi,[40] Chipster,[41] or Mobyle[42] and relies on textual workflow construction in a script-like format. Certain tools in this group, such as Chipster, enable Web-based collaborative development of workflows.

The second group of platforms provides the so-called "workbench environment": a GUI enabling visually supported construction of workflows. Usually, workflows are represented as graphs, where nodes correspond to data processing steps, and edges to data flow. Workbench solutions include Galaxy,[43] Taverna,[44] Pipeline Pilot,[45] KNIME,[46] or gUSE.[47] Similar to data integration platforms, these tools need to be deployed on the user-provided infrastructure, and the extent of possible analysis is restrained by available storage and HPC capacities.

Ensuring computational resources may be a challenging task, and cloud computing becomes increasingly more important paradigm in development and execution of bioinformatics workflows. Cloud-oriented workflow management systems offer API support for construction of an analytical pipeline, including open-access solutions, such as Agave[48] or Arvados,[49] or a number of commercial services.[5] Workbench platforms are also available in the computational cloud environment. Interestingly, a number of open-access solutions use Galaxy as a workflow construction engine, including Galaxy Cloud,[50] Tavaxy,[51] or Genomics Virtual Laboratory.[38] Commercial cloud workbenches, such as Seven Bridges (http://sbgenomics.com), are also available. In summary, cloud computing is an attractive scalable option on demand, especially for multisite collaborative research projects in terms of bringing the tools to the data. However, the speed of data transfer to the cloud, flexibility of the configuration of analytical pipelines, and the issues of privacy and security in data analytics remain challenges to address.[15,36]

Platforms for visualization of molecular interaction networks

With the progress of systems biomedicine, molecular interaction networks[§] became a popular form of representing knowledge about molecular mechanisms pertinent to human health.[52] First, such networks provide a necessary format to encode multitude of interactions identified in biomedicine. Second, they provide a good support for visual exploration and analytics of complex knowledge.[53] As such, they have a great potential in aiding the interpretation of analytical outcomes of translational medicine pipelines.

Molecular interaction networks can be constructed in various ways that determine their size and purpose. Experiment-derived networks are established from different types of molecular readouts, allowing, with a certain probability, ascertain physical interaction between molecules, for example, protein–protein interaction[54] or chromatin immunoprecipitation assays.[55] Analysis-inferred networks are constructed by analyzing high-throughput omics data to identify molecules with similar properties or behavior, for example, using coexpression analysis.[56] Finally, knowledge-based networks are established on the basis of existing body of knowledge, usually a set of published articles. Construction of knowledge-based networks is usually accomplished with text mining approaches[57] or expert curation.[58,59]

While experiment-derived and analysis-inferred networks offer a vast amount of unbiased information, they are usually large-scale graphs, requiring sophisticated network analysis to draw meaningful conclusions.

[§]**Molecular interaction networks:** A class of graphs, where nodes represent various biomolecules, and edges represent interactions between them.

Mapping translational medicine data sets on top of these networks may be considered an important step in the analysis[60] but not in the final interpretation of an analytical workflow. In turn, knowledge-based networks are usually established on the basis of low-throughput, in-depth experiments and allow for direct data interpretation. In particular, text mining networks are often used by the scientific community, where a number of commercial solutions, such as Ingenuity Pathway Analysis,[61] Pathway Studio,[62] or MetaCore,[63] offer already established databases. These solutions, however, tend to contain the entire discovery pipeline inside their platforms, greatly reducing data interoperability.

Expert-curated networks are focused resources of high-quality confirmed knowledge and offer the highest degree of data set interpretation to translational medicine researchers. Important resources in the field of expert-curated networks are repositories called "pathway databases," such as KEGG,[64] Reactome,[65] or WikiPathways,[59] which describe general biomolecular mechanisms. In contrast, the other type of networks focuses on representing mechanisms of human diseases as so-called "disease maps."[58,66,67] Detailed representation of domain knowledge and support by domain-related literature makes disease maps a potentially interesting element of translational medicine analytical pipelines. Computational architectures supporting these maps provide dedicated APIs,[68,69] opening an interesting avenue in translational medicine data processing—from storage, through bioinformatics workflow analytics, to interpretation by visualization on the dedicated molecular interaction network.

Approach

A flexible workflow for translational medicine big data needs to provide biomedical experts, such as medical doctors and life scientists, with a possibility to explore high-dimensional data sets. Given the complexity of source data, experts need to be able to flexibly define constraints and filters to focus on the most representative data points for particular health-related questions. Selected data points need to be processed, often in multiple analytical steps, as biomedical data are heterogeneous and represent complex readouts. Finally, biomedical experts need to interpret their findings in the context of biological mechanisms to formulate hypotheses on disease mechanisms.

We decided to focus on translational medicine workflow providing the possibility of visually aided data exploration and informative hypothesis formulation. Therefore, our data integration platform of choice was tranSMART as it is a server-based solution with i2b2 data exploration component. We chose Galaxy as a workflow management system, considering its flexibility and the availability of tools. Finally, to provide informative interpretation of analytical results, we bridged the Galaxy Server with MINERVA platform, allowing overlay of exported data on disease-related mechanisms.

We approached this problem in three steps:

1. Data integration and exploration are handled using tranSMART repository[28]
2. Analysis of tranSMART-provided data is supported by Galaxy Server workflows[43,70,71]
3. Visualization of Galaxy-provided results is enabled via domain-specific knowledge repositories.[58]

The workflow, as illustrated in detail in Figure 1, assumes a biomedical expert supervising each of the steps, while dedicated interfaces support automated data transition between each step.

Integration and exploration of clinical and molecular data in tranSMART database

Translational medicine data sources are heterogeneous and of various granularities,[2,72] and visually aided data exploration[73] is an important enabling technology for biomedical experts. The powerful visualization and interoperability functionalities of i2b2 are coupled together with omics integration in tranSMART[28] platform. tranSMART is a well-established platform enabling translation of preclinical research data into meaningful biological knowledge.[74] It supports integration of low-dimensional clinical data and high-dimensional molecular data in a data warehouse architecture. Although tranSMART by default relies on a relational database technology, it extends toward storing the high-dimensional biological data using NoSQL technology HBase.[75]

The platform features data interoperability connectors, including clinical information collection,[76] imaging data,[77] visual analytics,[78] or bioinformatics workflow management.[79] Finally, tranSMART features built-in data mining and analysis applications based on open-source systems, such as i2b2 and GenePattern,[28] and provides plugins to external tools, such as Dalliance Genome Browser,[80] or APIs for statistical packages, such as R.[81]

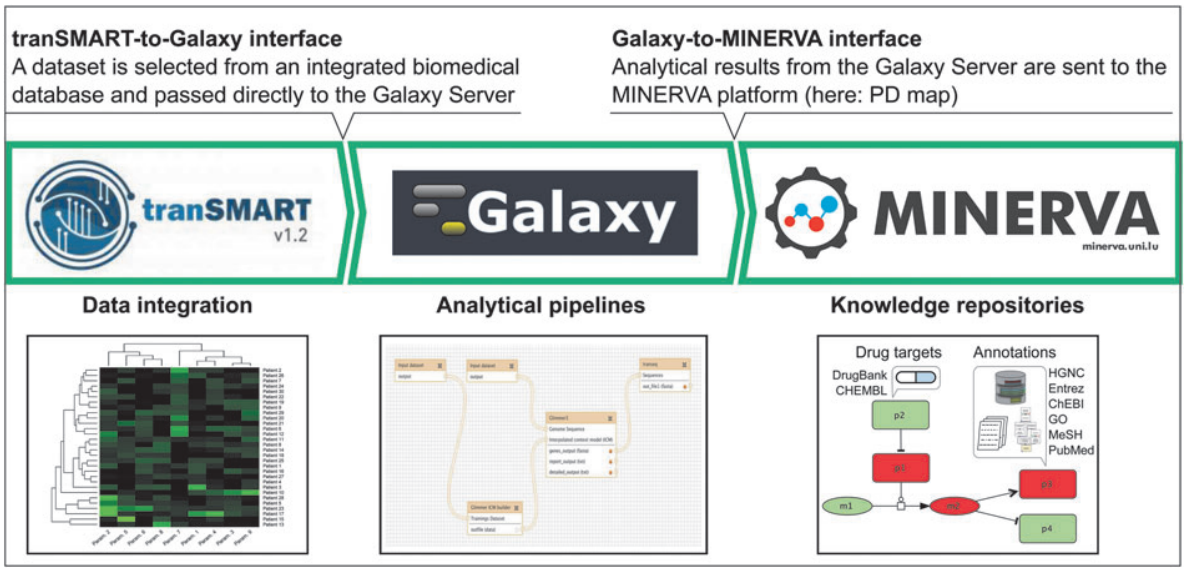


FIG. 1. A workflow for big data analytics in translational medicine. Clinical and "omics" data are integrated in the tranSMART database, allowing their exploration and selection of relevant subsets for downstream analysis. Selected data set is automatically transferred to Galaxy Server as a source for user-defined analytical pipelines. Finally, the results of the analysis are automatically transferred to an associated knowledge repository hosted on MINERVA platform (here: PD map) and displayed on the visualized molecular interaction networks. PD, Parkinson's disease.

For the abovementioned reasons, tranSMART became a technology of choice for European Translational Information and Knowledge Management Services (eTRIKS, www.etriks.org) initiative. eTRIKS provides infrastructure for data staging, exploration, and use in translational research supported by Innovative Medicines Initiative (IMI). IMI is a collaborative scheme, in which academic institutions and pharmaceutical companies in Europe conduct large-scale biomedical research.

To take advantage of the multiple functionalities of tranSMART, the target data sets have to be integrated following strict rules of data harmonization, semantic alignment, and quality checking. The data sets are curated following three common steps:

1. Data extraction: Source raw data files are extracted from either public or private data repositories. This could be a simple FTP transference from a Web repository or a database dump from a database management system, such as MySQL or Oracle™.
2. Data retrieval: Target information from the raw source files is identified and converted as Standard Format Files as defined by tranSMART curation guidelines. At this step, subject-level to sample-level data mapping is established.
3. Data annotation: Completing and standardizing annotations of metadata are also expected for guaranteeing data provenance.

The final product of the abovementioned steps is a set of Standard Format Files, which are used as input by tranSMART's native ETL (Extract, Transform, and Load) scripts. After data curation and loading to tranSMART, features collected for subject and samples become variables of the corresponding data set. These variables, as well as the relationships among them, are represented as a hierarchical parent–child tree control structure (or simply "tree," see Fig. 2). This tree can be gradually expanded, which allows efficient data sets exploration and also the selection of variables from the hierarchy to build customized patient subsets for downstream analysis. Features that characterize desired data points in the tree, such as "age," "gender," or "disease state," could be used as filters to narrow down the selected group. With tranSMART, researchers can pinpoint groups of patients and samples sharing similar

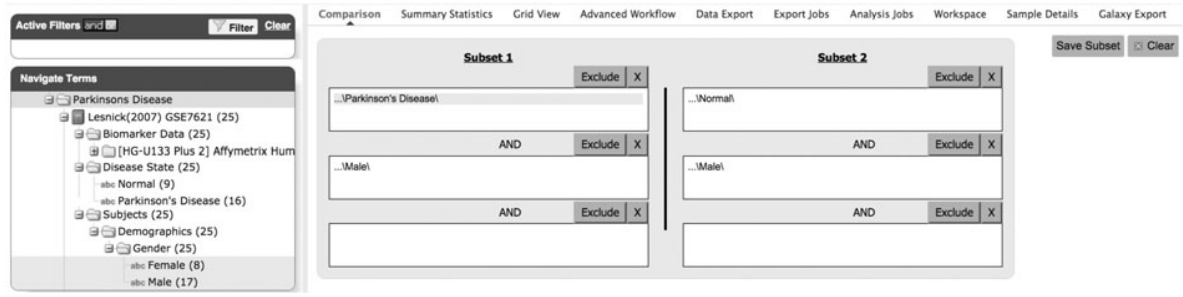


FIG. 2. Cohort/subset definition based on the variables displayed in data tree. Two distinct subsets are defined based on the variables "disease state" and "gender." In the left panel: data tree in tranSMART data set explorer. The data tree for GEO study GSE7621 following curation and loading to tranSMART is shown here. The data leafs correspond to the low- and high-dimensional data variable names. GEO, Gene Expression Omnibus.

characteristics, allowing straightforward hypothesis formulation. Easy identification of such coherent groups is a necessary prerequisite for accurate downstream analysis.

tranSMART platform has certain limitations concerning the size of handled data sets. First, data curation and integration are very time-consuming steps, which are necessary to upload a multidimensional data collection to tranSMART. Only then, users benefit the most from visually aided data exploration and interpretation. Second, considering the growing volume of data collected per patient,[12,15] data storage may become a bottleneck in the proposed architecture. In our experience with tranSMART, even when working with a data set of 15,000 patients each with 2000 clinical variables, the system was responsive (data not shown). However, storing omics for these patients in the native tranSMART database is an issue. NoSQL solutions can be considered to address the problem of both storage and analysis of large data.[75] Final bottleneck we foresee concerns the visualization capabilities of tranSMART. Displaying large amounts of data points via Web browsers is inefficient and may become a burden for large data sets.

Analysis of selected data points using Galaxy Server

The process of selection and filtering of tranSMART data results in a focused subset, which is suitable to answer a particular research question of a biomedical expert. For this to happen, an analytical workflow needs to be designed, pinpointing key characteristics of the selected subset.

Galaxy as a bioinformatics workflow management system is available as both Web server and cloud workbench, offering flexibility in terms of data interoperability and allocation of computational resources.[43,50,51] The Galaxy environment automatically and transparently tracks every detail of the analysis, allows the construction of complex workflows, and permits the results to be documented, shared, and published with complete provenance, guaranteeing transparency and reproducibility.[50] Galaxy Tool Shed[82] is a repository of more than 3000 community-developed tools, allowing easy and versatile establishing of bioinformatics workflows.[70,71] Such workflows may combine different aspects of expert knowledge required in subsequent analytical steps. Basic knowledge about the system is sufficient to use default elements in the workflow construction. These default methods can be modified, where the user has sufficient expertise. Once the workflow step is done, users can easily share and modify it. Analytical results can be directly visualized using embedded functionalities or exported for downstream processing.

Data interoperability and flexibility are important features of Galaxy. The platform is available in both server and cloud-based versions and bridges to the other major bioinformatics workflow management systems—Taverna,[51] KNIME, and gUSE.[37] Such architecture permits transparent and replicable design of analytical workflows for data exploration and formulation of data-driven hypotheses.

Galaxy may face similar data volume-related issues as discussed above on tranSMART. In case of big omics data sets, data transfer and analysis may become time consuming, especially for large subsets chosen for analysis and complicated workflows. A possible solution to

consider in such a case are advanced computational architectures offered by other workflow managers, such as gUSE. This solution is feasible to consider in the light of recent results on KNIME–Galaxy–gUSE workflow translation.[37]

Interpretation of analytical results using contextualized knowledge repositories

High-dimensional translational medicine data sets are difficult material to draw conclusions relevant for human health. Data sets exported preselected from tranSMART database and analyzed using Galaxy will either, in many cases, remain multidimensional data sets or will be reduced to the list of prioritized molecules. Interpretation of such results remains challenging and requires both contextualization and visualization. Galaxy Server allows various export options. As the last step of our pipeline, we propose to interpret the results of analysis of Galaxy Server in the context of dedicated knowledge repositories supported by MINERVA platform, such as Parkinson's disease (PD) map.[58,69] In particular, molecules prioritized by the constructed pipeline are automatically visualized on molecular interaction networks hosted by MINERVA platform.[69]

Knowledge on detailed molecular mechanisms can be assembled in the context of a given biological mechanism or a particular perturbation of this mechanism—a disease. Among others, Systems Biology Graphical Notation (SBGN)[83] is used as a format for such mechanistic descriptions. Importantly, SBGN foresees a diagrammatical description of molecular mechanisms, introducing an important aspect of visualization to their curation. In effect, a "map" of molecular processes can be drawn and then visually explored for a comprehensive understanding of complex interactions. A number of systems biology-oriented maps were established following this paradigm.[84–86] More importantly, the so-called "disease maps" gained interest as a way to assemble an overview of pathways and perturbations specific to a given pathology.[58,67,87,88]

MINERVA platform is a Web server supporting curation and visualization of SBGN-compliant molecular interaction networks. The maps of biological processes can be drawn in editors supporting SBGN notation, such as CellDesigner (www.celldesigner.org) or SBGN-ED (www.sbgn-ed.org), and uploaded to an instance of MINERVA Web server. There, the maps are automatically verified and annotated and become accessible for exploration via Web browser. MINERVA features dedicated functionalities coupled with Google Maps API to enable intuitive visual exploration, interaction with visualized content, advanced search queries, and mapping experimental data on the displayed networks. In turn, drug-targeting interface facilitates health-related interpretation or hypothesis generation.

Results

We have combined three server-based platforms addressing different aspects of data processing in translational medicine research—data integration and exploration, bioinformatics workflow construction, and interpretation of analysis results in the disease context. In our choice of technologies, we focused on two criteria—capability for exploratory hypothesis generation and data interoperability. The platforms of our choice, tranSMART, Galaxy, and MINERVA, can be deployed as a single data processing workflow for translational medicine.

We focused on available PD studies and exercised our workflow as described above, from data set selection and filtering in tranSMART, through analysis in Galaxy Server, to interpretation of results in the PD map—an open-access dedicated knowledge repository. We have established a dedicated Virtual Machine** to demonstrate the functioning of our workflow. To provide data sets for exploration, we have used tranSMART PD data sets we previously curated, which are also available at https://public.etriks.org.

Integration and visual exploration of PD data sets in tranSMART

For the first step of our workflow, we used PD-related studies that are publicly available in the Gene Expression Omnibus (GEO) database.[89] To integrate the GEO studies, data curation was performed to meet the required format of tranSMART,[74] as discussed above. In this use case, we worked with the GSE7621 PD study data[90] for defining two focused cohorts using tranSMART data set explorer.

Study-related variables in tranSMART can be assigned to two broad categories: low- and high-dimensional data. Low-dimensional data correspond mostly to clinical, patient-centric data (e.g., systolic blood pressure) and low-throughput diagnostic measurements (e.g., quantification of a disease-related blood biomarker). The corresponding values of low-dimensional data are usually available as text or numeric values. High-dimensional data, in the majority reflecting "omics" data, are structured as a numeric matrix.

**Available at http://r3lab.uni.lu/web/tgm-pipeline

For the purpose of this work, we used tranSMART for defining two specific patient cohorts based on low-dimensional data. We used tranSMART data set explorer to traverse the data tree displaying the low- and high-dimensional data variables for a given study (Fig. 2). Using associated drag-and-drop functionality, we performed on-the-fly cohort definition. As can be seen in Figure 2, the two cohorts have been defined based on the variables "disease state" and "gender." Having these two cohorts, we were in subsequent steps to export their high-throughput data sets containing gene expression profiles of the patient brain samples for downstream analysis and visual exploration.

Interface: tranSMART to Galaxy Server

Once subcohorts are built using the i2b2 tree, all data related to the two subcohorts can be exported as tab-delimited files. This step is possible as tranSMART data interface enables export of all selected data to be shared with analytical tools. To make the gene expression data available to the Galaxy environment for further analysis, we used tranSMART data export functionality. This connection has been implemented within the collaboration of the eTRIKS consortium and the tranSMART foundation. In particular, exported data can be streamlined automatically to an associated Galaxy Server via the Galaxy plugin to tranSMART (https://github.com/thehyve/transmart-galaxy).

The tranSMART–Galaxy interface uses the export function of tranSMART and transfers the files via Galaxy API to the Galaxy Server. User of Galaxy will then have access to the exported data of the subcohorts built in tranSMART. This way, preselected microarray data then become available in the Galaxy Server Workspace (GSW) for further analysis.

Both tranSMART and Galaxy provide user access rights management functions. Here, we rely on security mechanisms natively provided by the two systems. The interface requires a preconfigured login–password pair to upload data to a dedicated GSW. The login–password pair is then used as a parameter in the interface configuration, such that only users having access rights to both systems can establish the interface and execute the data transfer over it.

Analysis of a selected subset on Galaxy

High-throughput data provided by tranSMART contain gene expression in samples from the two selected cohorts: males with PD (four samples) and age-matched healthy males (eight samples). The data files are automatically available in GSW after their export from tranSMART and can be used as input files.

We have designed a dedicated Galaxy workflow (Fig. 3). The workflow is subdivided into steps from incorporation of the input files taken from tranSMART through performing the differential expression analysis and uploading the obtained results to the PD map hosted on the MINERVA platform and making them accessible for interpretation in the disease-specific context.

A comparison between these two data sets provides insight about disease-related mechanisms that may be cohort specific. This differential gene expression was calculated as predefined method using Bioconductor package "limma" in Galaxy[91,92] (absolute fold change >1.5, p-value <0.05). The resulting list of 3286 differentially regulated genes was uploaded via MINERVA to the PD map for visual interpretation. This process led to the labeling of 224 different genes and/or their related protein products in the PD map.

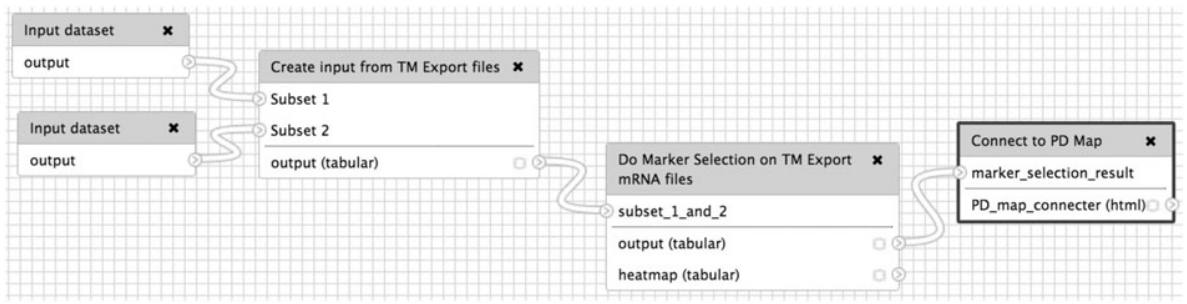


FIG. 3. Visually constructed data flow in the Galaxy Server comparing two cohorts from tranSMART.

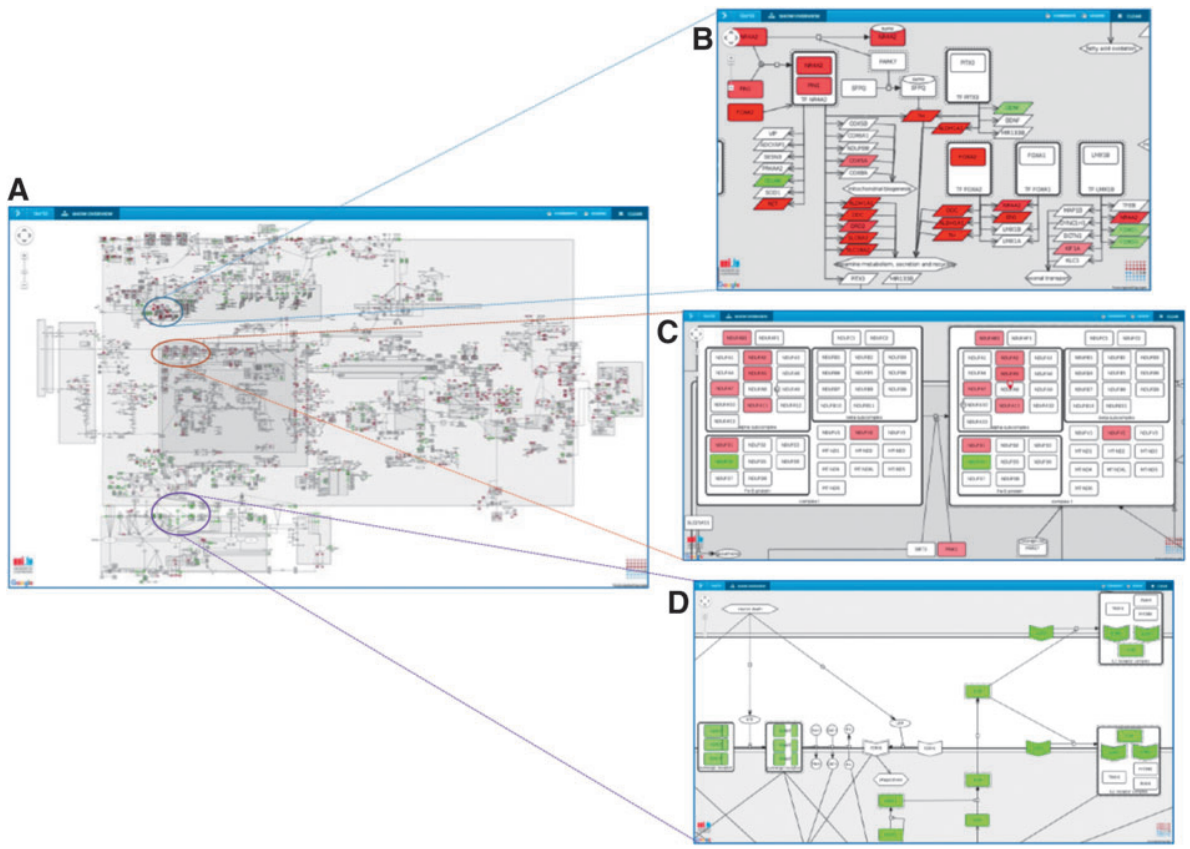


FIG. 4. Data visualization and analysis using PD map. **(A)** Differential gene expression data comparing postmortem brain tissues from male PD patients versus controls are displayed on the PD map (green, upregulated; red, downregulated). Pathways and processes of conspicuous areas (colored circle) could be identified using the pathway and compartment layout view of the PD map. Detailed view on deregulated genes that encode for proteins involved in dopamine metabolism, secretion, and recycling **(B)**, on mitochondrial electron transport chain, in particular elements of complex I **(C)**, and on microglia activation **(D)**.

Interface: Galaxy Server to MINERVA

MINERVA platform accepts POST requests, where the user specifies the target molecular network, user, password, and the data set to be uploaded. To ensure seamless data transfer from Galaxy to MINERVA, we created a step in the GSW called "PD map connector."[††] This step generates a POST request to the associated MINERVA instance—PD map in this case—to generate a custom visualization on the basis of the workflow data.

In the backend of the target MINERVA instance, a temporary session will be created for that particular data set to generate a custom layout, which will be available in the "Layouts" tab after user logs in. The uploaded data set may contain different types of elements, allowing coloring elements corresponding to multiple "omics" or interactions of the visualized network.[69]

By seamlessly connecting Galaxy Server to MINERVA platform, the users can securely transfer analysis results obtained from Galaxy workflows to MINERVA platform without leaving the Galaxy system. As shown below, visualization of the results on the PD map allows the identification of major molecular pathways perturbed in postmortem brain tissue of male Parkinson's patients, as selected in tranSMART and processed in Galaxy.

[††]Code available on the project website, http://r3lab.uni.lu/web/tgm-pipeline

Upload and interpretation of analysis results in the PD map

The data exploration and analysis steps described above created a list of molecules characterizing the PD-related cohort in comparison to the controls. This list is then projected on the PD map, a contextualized knowledge repository of mechanisms relevant for the disease. The repository is hosted using the MINERVA platform, a Web service for visualization of molecular networks, with the capability of custom data upload and mapping.[69] Pathways and processes displayed in the PD map provide disease- and cellular context-related information.[93] More than 1500 molecular interactions displayed in the PD map are from more than 1000 PD-related publications manually curated by experts.[58]

Evaluation of highlighted areas in the PD map shows pronounced alterations in the cell nucleus, in particular a battery of downregulated (red) genes involved in metabolism and secretion of the neurotransmitter dopamine (Fig. 4, blue circle[‡‡]).[94] Another visible perturbation affects the mitochondria, in particular elements of complex I (Fig. 4, red circle[§§]). This process is essential for energy homeostasis, in particular in high energy demanding neurons. Finally, we observe upregulation (green) of processes involved in neuroinflammation (Fig. 4, purple circle[***]).[95] On the basis of this visual exploration, data analyst may get comprehensive insights in molecular processes potentially involved in the disease of this specific patient cohort supporting new insights for diagnosis, prognosis, and therapy. Another approach for visualization is the drug target interface integrated in the MINERVA platform, enabling the mapping of potential drug interactions with elements of the map, suggesting more precise treatments and possibly an improvement in existing therapies.[96]

Conclusions

Visualization is a necessary tool on the interface between the expert and big data processing pipelines. It is especially important in the field of translational medicine, where biomedical experts formulate and test their hypotheses about new diagnostic approaches or treatment. This process can be greatly supported with the available translational medicine big data, including clinical and molecular data sets.[97] Efforts in this direction are reflected with development of disease-oriented knowledge repositories, for example, for pulmonary[26] or neurodegenerative disorders.[98] Nevertheless, these knowledge bases lack seamless data flow and require a number of explicit data transformation steps for exploratory analysis. In turn, less technically versed users are restrained in data-driven hypothesis generation.

Currently, a single person has to master a wide range of skills to perform a complete biomedical data analysis and interpretation. This is one of the reasons that big data integration, analytics, and interpretation become the true bottleneck of translational medicine.[15] We address this issue by seamlessly combining platforms supporting these steps, each of them having strong components of visually aided data exploration and analysis. Our approach is modular and capitalizes on strong points of each of the platforms, promoting data interoperability. Similar efforts have already been proposed,[99] involving tranSMART as data integration platform and a commercial solution GeneData as the analysis and interpretation engine. We believe that our pipeline extends their approach by involving a disease-related knowledge repository and by involving only open-access technologies will be useful for the scientific community.

The platforms of our choice are server based, allowing construction of the entire pipeline in a protected environment, avoiding ethical and legal issues present in the cloud scenarios. Nevertheless, cloud computing paradigm is compelling, especially for researchers having limited storage and HPC capabilities.[5,16] Efforts in this direction are promising[31] and need to be supported by further advances in data interoperability.[12] We believe our work is a step in this direction.

Acknowledgments

We would like to thank the reviewers of this article for their constructive remarks that helped in improving the article. We thank eTRIKS (www.etriks.org) and AETIONOMY (www.aetionomy.org) consortia. This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking AETIONOMY and eTRIKS grants and LCSB, University of Luxembourg.

Author Contributions

VS, WG, SE conceived and designed the project. ABS, WG, SE curated the data and integrated into tranSMART. VS, SE implemented Galaxy workflows. VS, PG developed the Galaxy-MINERVA interface. SG

[‡‡]In the PD map: http://tgm-pipeline.uni.lu/minerva/?search=%22TF%20NR4A2%22
[§§]In the PD map: http://tgm-pipeline.uni.lu/minerva/?search=reaction:re720
[***]In the PD map: http://tgm-pipeline.uni.lu/minerva/?search=IL1%20receptor%20complex,purinergic%20receptor

interpreted the experimental results. VS, RS, RB supervised the project. All the authors wrote and revised the manuscript, and MO coordinated the writing.

Author Disclosure Statement

No competing financial interests exist.

References

1. Topol EJ. The big medical data miss: Challenges in establishing an open medical resource. Nat Rev Genet. 2015;16:253–254.
2. Bender E. Big data in biomedicine: 4 big questions. Nature. 2015;527:S19.
3. Regan K, Payne PRO. From molecules to patients: The clinical applications of translational bioinformatics. Yearb Med Inform. 2015;10:164–169.
4. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2:84.
5. Costa FF. Big data in biomedicine. Drug Discov Today. 2014;19:433–440.
6. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: A systematic review. J Am Med Informatics Assoc. 2015;22:330–339.
7. de Bono B, Grenon P, Sammut SJ. ApiNATOMY: A novel toolkit for visualizing multiscale anatomy schematics with phenotype-related information. Hum Mutat. 2012;33:837–848.
8. Tian Q, Price ND, Hood L. Systems cancer medicine: Towards realization of predictive, preventive, personalized and participatory (P4) medicine. J Intern Med. 2012;271:111–121.
9. Liu Y, Yu C, Zhang X, et al. Impaired long distance functional connectivity and weighted network architecture in Alzheimer's disease. Cereb Cortex. 2014;24:1422–1435.
10. Butte AJ. Translational bioinformatics: Coming of age. J Am Med Inform Assoc. 2008;15:709–714.
11. Andreu-Perez J, Poon CCY, Merrifield RD, et al. Big Data for health. IEEE J Biomed Heal Inform. 2015;19:1193–1208.
12. Wade TD. Traits and types of health data repositories. Health Inf Sci Syst. 2014;2:4.
13. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. Biomed Res Int. 2014;2014:1–13.
14. Stonebraker M, Beskales G, Pagan A, et al. Data curation at scale: The Data Tamer System. In: Proceedings of the 6th Biennial Conference on Innovative Data Systems Research. Asilomar, CA, January 6–9, 2013.
15. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Med Genomics. 2015;8:33.
16. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: A literature review. Biomed Inform Insights. 2016;8:1–10.
17. Dove ES, Joly Y, Tassé A-M, Knoppers BM. Genomic cloud computing: Legal and ethical points to consider. Eur J Hum Genet. 2015;23:1271–1278.
18. Canuel V, Rance B, Avillach P, et al. Translational research platforms integrating clinical and omics data: A review of publicly available solutions. Brief Bioinform. 2015;16:280–290.
19. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE—An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009;2009:391–395.
20. Ohno-Machado L, Bafna V, Boxwala AA, et al. iDASH: Integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc. 2012;19:196–201.
21. Oster S, Langella S, Hastings S, et al. caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. J Am Med Inform Assoc. 2008;15:138–149.
22. Payne P, Ervin D, Dhaval R, et al. TRIAD: The Translational Research Informatics and Data Management Grid. Appl Clin Inform. 2011;2:331–344.
23. Toga AW, Foster I, Kesselman C, et al. Big Biomedical data as the key resource for discovery science. J Am Med Inform Assoc. 2015;22:1126–1131.
24. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–404.
25. Madhavan S, Gauba R, Song L, et al. Platform for personalized oncology: Integrative analyses reveal novel molecular signatures associated with colorectal cancer relapse. AMIA Jt Summits Transl Sci Proc. 2013;2013:118.
26. Cano I, Tényi Á, Schueller C, et al. The COPD Knowledge Base: Enabling data analysis and computational simulation in translational COPD research. J Transl Med. 2014;12:S6.
27. Tan A, Tripp B, Daley D. BRISK—research-oriented storage kit for biology-related data. Bioinformatics. 2011;27:2422–2425.
28. Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational medicine. J Transl Med. 2010;8:68.
29. Saulnier Sholler GL, Ferguson W, Bergendahl G, et al. A pilot trial testing the feasibility of using molecular-guided therapy in patients with recurrent neuroblastoma. J Cancer Ther. 2012;3:602–612.
30. Natter MD, Quan J, Ortiz DM, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc. 2013;20:172–179.
31. Nguyen TD, Raniga P, Barnes DG, Egan GF. Design, implementation and operation of a multimodality research imaging informatics repository. Health Inf Sci Syst. 2015;3:S6.
32. Murphy S, Churchill S, Bry L, et al. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res. 2009;19:1675–1681.
33. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17:124–130.
34. Gainer V, Hackett K, Mendis M, et al. Using the i2b2 hive for clinical discovery: An example. AMIA Annu Symp Proc. 2007;959.
35. Kalaitzopoulos D, Patel K, Younesi E. Advancements in Data Management and Data Mining Approaches. Transl Med. 2016;31:35–53.
36. Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. [Epub ahead of print]; DOI: 10.1093/bib/bbw020.
37. de la Garza L, Veit J, Szolek A, et al. From the desktop to the grid: Scalable bioinformatics via workflow conversion. BMC Bioinformatics. 2016;17:127.
38. Afgan E, Sloggett C, Goonasekera N, et al. Genomics virtual laboratory: A practical bioinformatics workbench for the cloud. PLoS One. 2015;10:e0140829.
39. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2522.
40. Hunter AA, Macgregor AB, Szabo TO, et al. Yabi: An online research environment for grid, high performance and cloud computing. Source Code Biol Med. 2012;7:1.
41. Kallio MA, Tuimala JT, Hupponen T, et al. Chipster: User-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011;12:507.
42. Neron B, Menager H, Maufrais C, et al. Mobyle: A new full web bioinformatics framework. Bioinformatics. 2009;25:3005–3011.
43. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11:R86.
44. Wolstencroft K, Haines R, Fellows D, et al. The Taverna workflow suite: Designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. 2013;41:W557–W561.
45. Warr WA. Scientific workflow systems: Pipeline Pilot and KNIME. J Comput Aided Mol Des. 2012;26:801–804.
46. Jagla B, Wiswedel B, Coppee J-Y. Extending KNIME for next-generation sequencing data analysis. Bioinformatics. 2011;27:2907–2909.
47. Kacsuk P, Farkas Z, Kozlovszky M, et al. WS-PGRADE/gUSE generic DCI gateway framework for a large variety of user communities. J Grid Comput. 2012;10:601–630.
48. Dooley R, Vaughn M, Stanzione D, et al. Software-as-a-Service: The iPlant Foundation API. In: 5th IEEE Workshop on Many-Task Computing Grids and Supercomputers (MTAGS). IEEE, 2012.
49. Arvados. A free and open source platform for big data science. 2013. Available online at http://doc.arvados.org (last accessed April 25, 2016).
50. Afgan E, Baker D, Coraor N, et al. Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol. 2011;29:972–974.
51. Abouelhoda M, Issa S, Ghanem M. Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinformatics. 2012;13:77.
52. Jacunski A, Tatonetti NP. Connecting the dots: Applications of network medicine in pharmacology and disease. Clin Pharmacol Ther. 2013;94:659–669.

53. Gerasch A, Faber D, Küntzer J, et al. BiNA: A visual analytics tool for biological network data. PLoS One. 2014;9:e87397.
54. Pizzuti C, Rombo SE. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics. 2014;30:1343–1352.
55. Kim T-M, Park PJ. Advances in analysis of transcriptional regulatory networks. Wiley Interdiscip Rev Syst Biol Med. 2011;3:21–35.
56. Guo N, Wan Y-W. Network-based identification of biomarkers coexpressed with multiple pathways. Cancer Inform. 2014;13(Suppl 5):37–47.
57. Neves M, Leser U. A survey on annotation tools for the biomedical literature. Brief Bioinform. 2014;15:327–340.
58. Fujita KA, Ostaszewski M, Matsuoka Y, et al. Integrating pathways of Parkinson's disease in a molecular interaction map. Mol Neurobiol. 2014;49:88–102.
59. Kutmon M, Riutta A, Nunes N, et al. WikiPathways: Capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016;44:D488–D494.
60. Glaab E, Schneider R. PathVar: Analysis of gene and protein expression variance in cellular pathways using microarray data. Bioinformatics. 2012;28:446–447.
61. Kramer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30:523–530.
62. Pathway Studio®. Experimental data and disease models at the heart of biological research. 2016. Available online at www.elsevier.com/solutions/pathway-studio-biological-research (last accessed April 25, 2016).
63. MetaCore. MetaCore and Key Pathway Advisor Data-mining and pathway analysis. 2016. Available online at http://ipscience.thomsonreuters.com/product/metacore (last accessed on April 25, 2016).
64. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
65. Croft D, Mundo AF, Haw R, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42:D472–D477.
66. Mizuno S, Iijima R, Ogishima S, et al. AlzPathway: A comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst Biol. 2012;6:52.
67. Kuperstein I, Bonnet E, Nguyen H-A, et al. Atlas of cancer signalling network: A systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis. 2015;4:e160.
68. Bonnet E, Viara E, Kuperstein I, et al. NaviCell Web Service for network-based data visualization. Nucleic Acids Res. 2015;43:W560–W565.
69. Gawron P, Ostaszewski M, Satagopam V, et al. MINERVA—a platform for visualization and curation of molecular interaction networks. 2016. Available online at http://r3lab.uni.lu/web/minerva-website (last accessed April 25, 2016).
70. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: A web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010;Chapter 19:19.10.1–21.
71. Giardine B, Riemer C, Hardison RC, et al. Galaxy: A platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451–1455.
72. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearb Med Inform. 2014;9:14–20.
73. Shneiderman B, Plaisant C, Hesse BW. Improving healthcare with interactive visualization. Computer. 2013;46:58–66.
74. Scheufele E, Aronzon D, Coopersmith R, et al. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Jt Summits Transl Sci Proc. 2014;2014:96–101.
75. Wang S, Pandis I, Wu C, et al. High dimensional biological data retrieval optimization with NoSQL technology. BMC Genomics. 2014;15:S3.
76. Blondé W, de Bruijn F. OpenClinica and RedCap conversion to transMART. 2015. Available online at https://github.com/CTMM-TraIT/trait_odm_to_i2b2 (last accessed April 25, 2016).
77. Vast E. tranSMART XNAT importer. 2015. Available online at https://github.com/evast/transmart-xnat-importer-plugin (last accessed April 25, 2016).
78. Herzinger S. SmartR: A grails plugin for visual analytics of the tranSMART platform using recent web technologies. Available online at https://github.com/transmart/SmartR (last accessed on April 25, 2016).
79. Bierkens M, van der Linden W, van Bochove K, et al. tranSMART. J Clin Bioinforma. 2015;5:S9.
80. Down TA, Piipari M, Hubbard TJP. Dalliance: Interactive genome viewing on the web. Bioinformatics. 2011;27:889–890.
81. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2013.
82. Lazarus R, Kaspi A, Ziemann M. Creating reusable tools from scripts: The Galaxy Tool Factory. Bioinformatics. 2012;28:3139–3140.
83. Le Novère N, Hucka M, Mi H, et al. The systems biology graphical notation. Nat Biotechnol. 2009;27:735–741.
84. Oda K, Matsuoka Y, Funahashi A, Kitano H. A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol. 2005;1:2005.0010.
85. Oda K, Kitano H. A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol. 2006;2:2006.0015.
86. Caron E, Ghosh S, Matsuoka Y, et al. A comprehensive map of the mTOR signaling network. Mol Syst Biol. 2010;6:453.
87. Matsuoka Y, Matsumae H, Katoh M, et al. A comprehensive map of the influenza A virus replication cycle. BMC Syst Biol. 2013;7:97.
88. Mizuno S, Iijima R, Ogishima S, et al. AlzPathway: A comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst Biol. 2012;6:52.
89. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: Archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–D995.
90. Lesnick TG, Papapetropoulos S, Mash DC, et al. A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease. PLoS Genet. 2007;3:e98.
91. Smyth GK. Limma: Linear models for microarray data. In: Gentleman R, Carey V, Huber W, et al. (Eds.): Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer, 2005, pp. 397–420.
92. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
93. Hofmann-Apitius M, Ball G, Gebel S, et al. Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. Int J Mol Sci. 2015;16:29179–29206.
94. Meiser J, Weindl D, Hiller K. Complexity of dopamine metabolism. Cell Commun Signal. 2013;11:34.
95. Glass CK, Saijo K, Winner B, et al. Mechanisms underlying inflammation in neurodegeneration. Cell. 2010;140:918–934.
96. Poletti M, Bonuccelli U. Acute and chronic cognitive effects of levodopa and dopamine agonists on patients with Parkinson's disease: A review. Ther Adv Psychopharmacol. 2013;3:101–113.
97. Dagliati A, Marinoni A, Cerra C, et al. Integration of administrative, clinical, and environmental data to support the management of type 2 diabetes mellitus: From satellites to clinical care. J Diabetes Sci Technol. 2016;10:19–26.
98. Aetionomy Knowledge Base: Organising knowledge about neurodegenerative disease mechanisms for the improvement of drug development and therapy. 2015. Available online at http://aetionomy.scai.fhg.de (last accessed April 25, 2016).
99. Schumacher A, Rujan T, Hoefkens J. A collaborative approach to develop a multi-omics data analytics platform for translational research. Appl Transl Genomics. 2014;3:105–108.

Cite this article as: Satagopam V, Gu W, Eifes S, Gawron P, Ostaszewski M, Gebel S, Barbosa-Silva A, Balling R, Schneider R (2016) Integration and visualization of translational medicine data for better understanding of human diseases. *Big Data* 4:2, 97–108, DOI: 10.1089/big.2015.0057.

Abbreviations Used

API = application programming interface
ETL = Extract, Transform, and Load
eTRIKS = European Translational Information and Knowledge Management Services
GEO = Gene Expression Omnibus
GSW = Galaxy Server Workspace
HPC = high-performance computing
IMI = Innovative Medicines Initiative
PD = Parkinson's disease
SBGN = Systems Biology Graphical Notation

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

The HIV Mutation Browser: A Resource for Human Immunodeficiency Virus Mutagenesis and Polymorphism Data

Norman E. Davey[1,2☯], Venkata P. Satagopam[3☯], Salvador Santiago-Mozos[1☯], Carlos Villacorta-Martin[1], Tanmay A. M. Bharat[1], Reinhard Schneider[3]*, John A. G. Briggs[1,4]*

1 Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, **2** Department of Physiology and Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, United States of America, **3** Luxembourg Centre for Systems Biomedicine, Belval, Luxembourg, **4** Molecular Medicine Partnership Unit, EMBL/Universitätsklinikum Heidelberg, Heidelberg, Germany

Abstract

Huge research effort has been invested over many years to determine the phenotypes of natural or artificial mutations in HIV proteins—interpretation of mutation phenotypes is an invaluable source of new knowledge. The results of this research effort are recorded in the scientific literature, but it is difficult for virologists to rapidly find it. Manually locating data on phenotypic variation within the approximately 270,000 available HIV-related research articles, or the further 1,500 articles that are published each month is a daunting task. Accordingly, the HIV research community would benefit from a resource cataloguing the available HIV mutation literature. We have applied computational text-mining techniques to parse and map mutagenesis and polymorphism information from the HIV literature, have enriched the data with ancillary information and have developed a public, web-based interface through which it can be intuitively explored: the HIV mutation browser. The current release of the HIV mutation browser describes the phenotypes of 7,608 unique mutations at 2,520 sites in the HIV proteome, resulting from the analysis of 120,899 papers. The mutation information for each protein is organised in a residue-centric manner and each residue is linked to the relevant experimental literature. The importance of HIV as a global health burden advocates extensive effort to maximise the efficiency of HIV research. The HIV mutation browser provides a valuable new resource for the research community. The HIV mutation browser is available at: http://hivmut.org.

Citation: Davey NE, Satagopam VP, Santiago-Mozos S, Villacorta-Martin C, Bharat TAM, et al. (2014) The HIV Mutation Browser: A Resource for Human Immunodeficiency Virus Mutagenesis and Polymorphism Data. PLoS Comput Biol 10(12): e1003951. doi:10.1371/journal.pcbi.1003951

Editor: Alan Rein, National Cancer Institute at Frederick, Frederick, Maryland, United States of America

Received June 26, 2014; **Accepted** September 29, 2014; **Published** December 4, 2014

Data Availability: The authors confirm that, for approved reasons, some access restrictions apply to the data underlying the findings. The resource is available at http://hivmut.org. The articles text-mined to create the resource are accessible at the publishers websites. These articles are linked from the resource, however, access to these articles may require a subscription.

Funding: This study was partly funded by the ViroQuant (a FORSYS) project (http://www.viroquant.uni-hd.de/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: reinhard.schneider@uni.lu (RS); john.briggs@embl.de (JAGB)

☯ These authors contributed equally to this work.

Introduction

Human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS), infects millions of people worldwide and, to date, has been responsible for over 25 million deaths [1]. The clinical importance of the virus has prompted substantial funding of HIV/AIDS research across many diverse clinical, therapeutic (drug design, vaccine production) and basic research fields. This research has produced an extensive catalogue of HIV literature and consequently finding literature pertinent to a particular topic is a difficult task. Researchers are often interested in the phenotypic variation resulting from naturally occurring single nucleotide polymorphism or directed mutagenesis in the HIV genome. Traditionally, mutation data for a particular protein or region must be manually collected by trawling literature repositories such as PubMed using author names, protein/gene names, keywords or a mixture of all three. The scale of the HIV literature (over 270,000 articles) makes such an approach inadequate. Several valuable online resources have provided mutation data to researchers by manually curating polymorphism and mutagenesis data from HIV studies. These include the Stanford Drug Resistance database [2], which curates mutations related to drug resistance, the UniProt knowledgebase [3], which manually annotates articles describing mutagenesis of HIV proteins and the Los Alamos HIV Database, which annotates various sources of HIV data including epitope variants and escape mutations (http://www.hiv.lanl.gov/). However, these resources are limited in scope because manual curation cannot feasibly be carried out on all of the available literature.

The technology exists to quickly computationally scan, annotate and organise scientific literature and these techniques should be applied to facilitate the work of HIV researchers [4,5]. Consequently, it is surprising that so few resources are available to access the available literature in an organised and structured way. This incongruity can partly be explained by the strict licensing agreements with scientific publishers that prohibit the bulk download and computational processing of scientific research literature. Fortunately, recent pressure from government and

Author Summary

Naturally occurring mutations within the HIV proteome are of therapeutic interest as they can affect the virulence of the virus or result in drug resistance. Furthermore, directed mutagenesis of specific residues is a common method to investigate the function and mechanism of the viral proteins. We have developed novel computational text-mining tools to analyse over 120,000 HIV research articles, identify data on mutations and work out which amino-acid in which protein has been mutated. We have organised these data and made them available in an online resource—The HIV mutation browser. The resource allows HIV researchers to efficiently access previously completed research related to their region of interest in the HIV proteome. The HIV Mutation Browser complements currently available manually curated HIV resources and is a valuable tool for HIV researchers.

scientific bodies and the rise of open access publishing has softened the stance of publishers and many are now receptive to waiving these restrictions. Such advances will pave the way for many large-scale literature text-mining projects and will likely change the way we access scientific literature.

Here we have applied text-mining techniques to extract data on polymorphisms and mutations from the available HIV literature. We have organised this data in a protein and residue-centric way and have made it available through an online resource, the HIV mutation browser (http://hivmut.org). This publicly available resource will simplify the task of virologists attempting to identify the relevant literature for their research, thereby aiding experimental design and reducing replication of efforts.

Results

Creation of the HIV mutation browser required a number of steps (Figure 1). First, we obtained permission from publishers, identified, and accessed the relevant literature. Second, we established and applied text-mining techniques to retrieve data on mutagenesis and polymorphism from the HIV literature. Third, we associated the mutation data to the appropriate residues within the HIV proteome. Finally, we developed a browser through which the data can be accessed in an intuitive and informative way.

Data mining and mutation statistics

We identified ~270,000 articles containing the search term "HIV" or "Human Immunodeficiency Virus" indexed in the PubMed database (from a total of ~23 million publications). We retained 120,899 of these articles, published across 2,614 journals, representing approximately 45% of the total (see materials and methods, Figure 1). For the remaining ~150,000 citations, permission for computational processing of articles was not obtained from the publisher. The 120,899 articles from participating publishers were text-mined for mutagenesis or polymorphism information, and the mutations were mapped to particular residues within the HIV proteome. This required the development of a method to retrieve the text of these articles, scan the articles for patterns that are widely used to describe directed mutations in mutagenesis experiments or polymorphisms, and to map these mutations to the correct position in the correct protein (see materials and methods). A total of 7,608 distinct mutations (a unique non-wildtype amino acid at a given residue in a given protein) were collected. As each mutation can be described in multiple articles and each article can describe multiple mutations, the 7,608 distinct mutations were defined by 43,264 unique references to 5,267 articles.

The identified mutations shed light on the nature of the HIV research effort of the last decades. On the one hand it has been broad in scope: 2,520 of the 3,118 residues in the HIV proteome have one or more associated references to a mutation in the repository. On the other hand it has been narrow in focus: the coverage is far from uniform and certain regions such as the catalytic sites of the protease and reverse transcriptase, as well as host interaction interfaces, are much more highly studied (Figure 2).

HIV Mutation Browser interface

The above analysis resulted in a database within which each reference to a mutagenesis experiment or polymorphism in a citation is indexed using three pieces of information: the protein in which the mutation is present, the position in the protein which has been mutated, and the non-wildtype amino acid to which the wildtype residue has been mutated. To make this data accessible to virologists in a simple, intuitive and informative manner, we designed the HIV Mutation Browser, as a web-interface that acts as a front end for the database. The browser presents the data in a hierarchically organised manner. The user selects a gene of interest, then a position of interest, and the citations relating to this position are presented to the user grouped by non-wildtype amino acid.

The web interface is organised in three panels: the *navigation panel* at the top; the *protein panel* in the middle; and the *residue panel* at the bottom (Figure 3).

Navigation panel. The navigation panel (Figure 3) contains a set of buttons that allows users to navigate the site. The most important of these buttons allow users to access the mutation information by choosing the gene of interest (tagged *gene*) and one of three different options for visualising the data in the protein panel (tagged *view*). The search button gives users access to a tool to query the resource by full text, author and PubMed identifier. The left hand side of the panel contains links to access the home page, the help page (containing all the information necessary to understand the information displayed on the website) and the about pages (containing up-to-date statistics about the website and information about participating publishers).

Protein panel. The mutagenesis and polymorphism data can be visualised and accessed by the user in three ways corresponding to the Sequence, Feature and Table view. All views are accessible by clicking on the relevant button, tagged "view", in the navigation bar (Figure 3). All views show the primary protein sequence of the selected HIV-1 gene - clicking on an amino acid of interest can access detailed mutagenesis data for that residue. The *Sequence* view displays the amino acid sequence of the protein of interest. Residues with mutation data are displayed in blue boxes. The *Feature* view displays the selected protein sequence annotated and enriched with known functional data, structural data and modular features (See Figure 3). The number of distinct mutations for each residue is displayed as a bar plot below the protein sequence. The *Table* view displays information on conservation, structure and chain, as well as the number of mutagenesis and polymorphism experiments available for each position in a tabulated format.

Residue panel. The residue panel displays the mutation information for the residue currently selected in the protein panel. The left side of the residue panel displays a list of citations that include mutation information for the residue. Each residue can have one or more distinct mutations and citations are grouped by

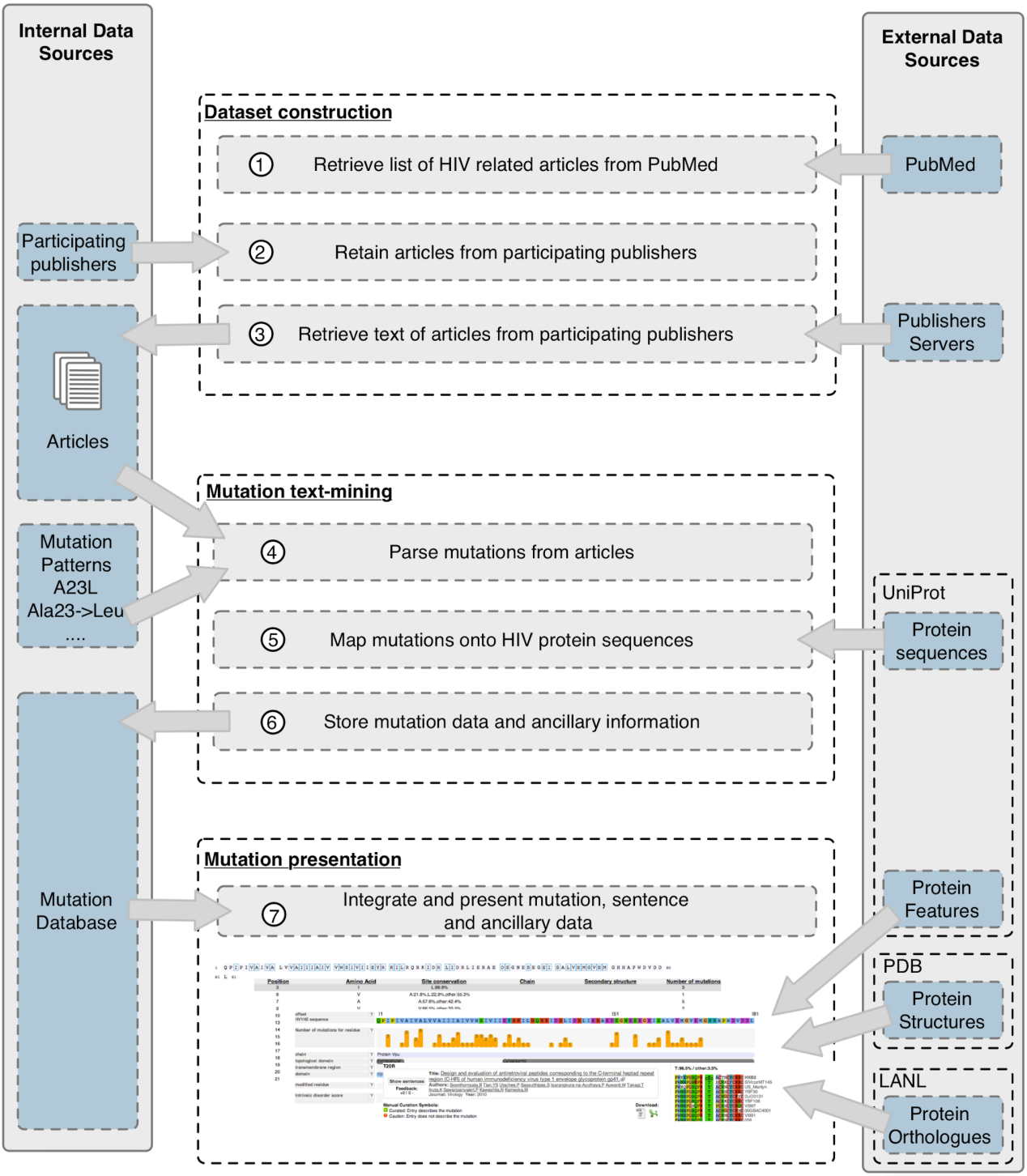


Figure 1. Schema describing article acquisition, mutation curation and data presentation for the HIV mutation browser. A list of HIV-related PubMed article identifiers (PMIDs) are retrieved from PubMed. The publishing journal of each article is compared against a list of participating publishers (i.e. publishers that have given permission for bulk PDF downloading, computational parsing of PDF and display of articles details). Permitted articles are retrieved from the publishers website as PDF files. Retrieved articles are computationally text-mined to parse patterns commonly used in the literature to denote mutations. Each mutation is then mapped onto the HIV proteome. Mutations are stored in a relational database and accessed through a web interface, the HIV mutation browser. The HIV mutation browser organises the data by protein and residue and integrates ancillary information relevant to the users. See methods section for more details.
doi:10.1371/journal.pcbi.1003951.g001

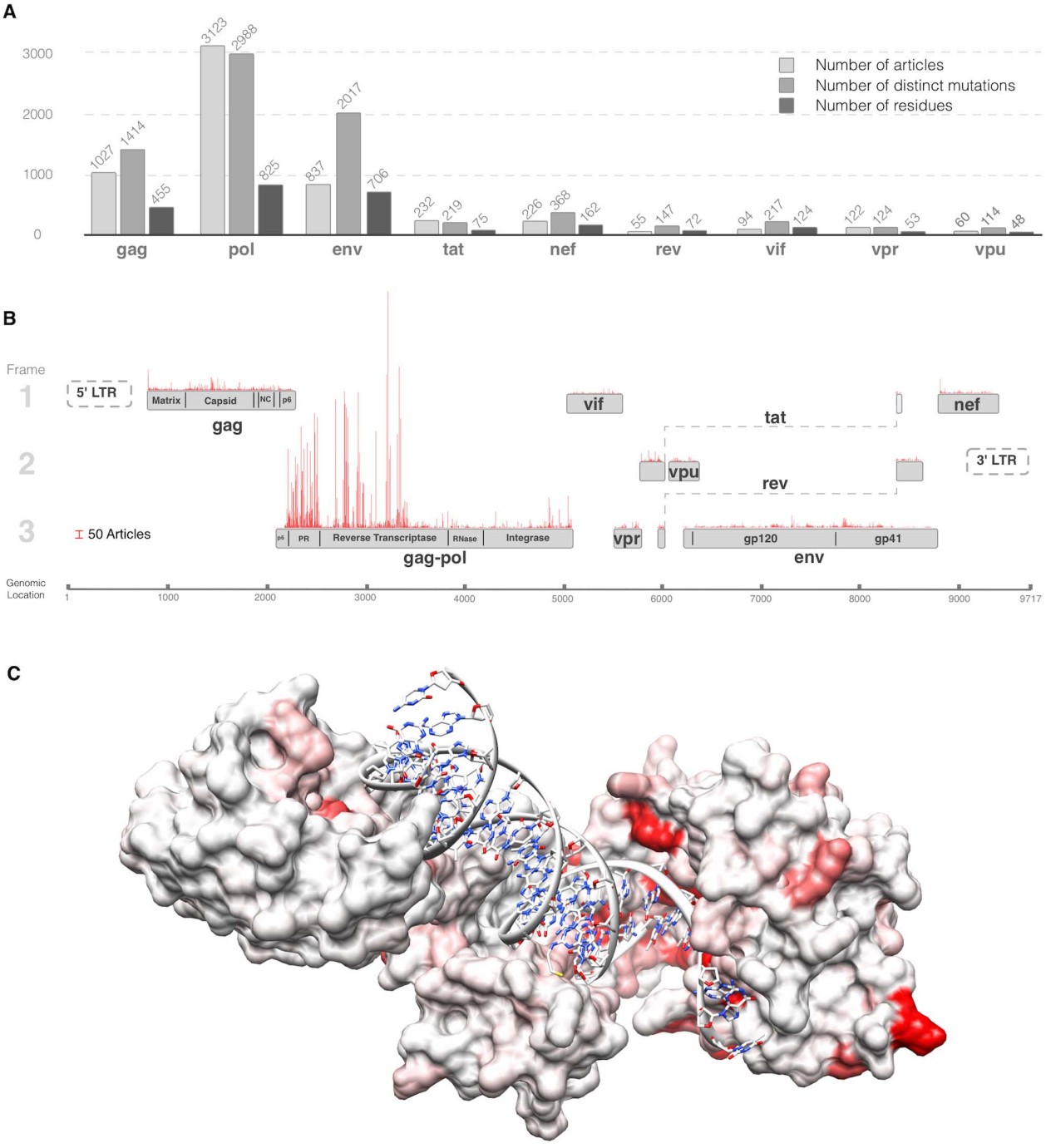


Figure 2. Overview of the distribution of mutation data across the HIV proteome. (A) Barplot of the counts of (i) the number of articles describing mutations, (ii) the number of distinct mutations and (iii) the number of residues with mutation data in the database for each protein in the HIV proteome. (B) Barplot of the counts of the number of curated articles in the database describing mutagenesis experiments or polymorphisms for each residue mapped onto the HIV proteome/genome. (C) Reverse transcriptase p66 subunit with residues coloured by number of articles referring to them. Most highly cited residues are in contact with the nucleotides or are known drug resistance mutations. White denotes no papers, full red denotes 50 or more papers, colouring is linearly scaled between 0 and 50.
doi:10.1371/journal.pcbi.1003951.g002

the non-wild type amino acid to which the wild type amino acid has been mutated. Within these groupings, if more than one citation is present, the citations are presented in an order that is based first on the results of manual curation, then on user feedback (see below) and finally on number of views. Each distinct mutation in each article has a unique citation (i.e. the same article may appear multiple times for the same residue entry but with different non-wild type amino acids).

For each citation, the information presented includes details of the residue: its position in the reference proteome, the wild type

Figure 3. HIV Mutation Browser interface for Vpu residue 51 showing the navigation, protein and residue panels. (1) Options bar for the residue view section of the interface. (2) Mutation information. (3) User feedback buttons. (4) Mutation information download links. (5) Ancillary residue information panel.
doi:10.1371/journal.pcbi.1003951.g003

amino acid and the non-wild type amino acid described in the article; as well as information on the article (title, author list, publishing journal, publication, year, direct link to the article at the publishers website). The 'show sentences' button reveals the sentences from the original article that describe the mutation and phenotype.

The right hand side of the residue panel contains supplementary contextual information. When available, the position of the residue is displayed on a solved structure of the protein domain, retrieved from RCSB Protein Data Bank (PDB) [6]. Hovering the cursor over the structure can rotate it to allow viewing from different directions. Alternatively, the user can choose to display a multiple sequence alignment of homologous reference HIV subtype protein sequences for the 10 flanking residues either side of the residue of interest.

User feedback and curation. The database has been populated by text-mining, and it is therefore unavoidable that the database contains incorrectly assigned citations (see discussion). We have therefore incorporated a user feedback system that allows users to flag the quality of an entry either positively or negatively. This feedback is presented for each entry and influences the order in which citations are presented to the user when multiple papers describing a given mutation are available. The feedback will also be used to guide manual curation by the HIV Mutation Browser team on a quarterly basis coinciding with each new releases of the HIV Mutation Browser. An emphasis will be put on the curation of entries that have received negative feedback and are likely to contain inaccuracies. Manually curated entries are indicated with a green tick (a good entry) or a red cross (a bad entry). This curation relates to the accuracy of the text-mining and does not reflect the quality of the paper.

Download

The available mutagenesis and polymorphism data for a residue can be downloaded in both tab delimited text and Excel formats directly from the web interface.

Discussion

HIV is an important therapeutic target and has been the subject of a major research effort as evidenced by the large catalogue of HIV experimental literature. Appropriate organisation and categorisation of the available HIV literature is necessary to allow efficient and intuitive access to relevant data. In this paper, we have presented the HIV Mutation Browser, a residue-centric resource of HIV mutagenesis and polymorphism literature designed for use by those carrying out basic and applied HIV research. The HIV Mutation Browser is one of the first resources to computationally text-mine mutagenesis and polymorphism data [7,8,9], and the first to apply such methods to the extensive corpus of HIV literature. As such the HIV Mutation Browser will complement the available manually annotated and curated HIV resources such as the Stanford Drug Resistance database [2], the UniProt knowledgebase [3] the Los Alamos HIV Database (http://www.hiv.lanl.gov/). In the coming years, we expect this method or similar methods to be applied to other viral or cellular systems.

The resource will continue to evolve in the following ways. Firstly, HIV literature is produced continuously at a rate of approximately 1,500 articles a month and consequently the HIV Mutation Browser resource will be updated on a quarterly basis. Secondly, while the resource does contain the majority of

important HIV and general interest journals, it is still incomplete, as we did not receive permission from all publishers to text-mine their HIV related articles. Journals from additional publishers will be added when possible. Thirdly, not all mutations can be correctly identified and assigned by the text-mining methods. There are various reasons for this. Many mutations are annotated in an article using non-standard patterns that are not widely used to describe directed mutations in mutagenesis experiments or polymorphisms. For example, consider the following excerpt taken from an article by Mitchell et al., "The phenotype of the combination mutant VpuD51A-S52/56N was indistinguishable from that of Vpu-D51A and Vpu-S52/56N" [10]. The pattern "S52/56N" is a non-canonical construct for describing a mutagenesis experiment and currently will not be discovered by the text-mining method. Furthermore, the position of a mutation in a paper can be ambiguous and as a result mapping of the mutation information to the correct residue and protein can be a difficult task. For example, when multiple proteins are referenced and only a single mutation is discussed (more than one possible mapping can be possible), when unconventional numbering is used (particularly when describing mutations in Gag or Env as both are translated as polypeptide chains and subsequently cleaved) or when unusual strains with insertions and deletions are used (this shifts the numbering of residues). We will continue to improve the methods for text-mining and assignment. We request that members of the community utilise the feedback system for misannotated mutations in the resource and contact us about mutation data that should be in the resource yet is not present. This community input will improve the quality of the annotated data and will pinpoint parts of the text-mining method that require improvement.

In summary, the HIV Mutation Browser is a valuable addition to the currently available HIV resources that will allow researchers to quickly and intuitively access data on mutagenesis and phenotypic variation. We expect the database to aid the process of experimental design and be a key resource for the HIV community.

Materials and Methods

Construction of the HIV literature dataset

A list of HIV-related articles was programmatically retrieved from PubMed using the search terms "HIV" and "Human Immunodeficiency Virus". A list of target journals was constructed based on the number of published HIV-related articles. The licensing agreements of the majority of scientific journals prohibit a licensee from (1) downloading articles in bulk and (2) computationally processing the text of an article. Permission to waive these aspects of the licensing agreement was requested and received from the majority of virology and general interest scientific journals. The text of all HIV-related articles from the participating journals was retrieved programmatically from the publisher's websites to create a *HIV literature dataset*. An up to date list of the participating journals and publishers is available on the HIV Mutation Browsers website.

Mutation text-mining of HIV literature

There is no globally applied nomenclature to define directed mutations in mutagenesis experiments or polymorphisms [5,11,12]. A set of templates that define phrases and shorthand widely used to describe directed mutations in mutagenesis experiments or polymorphisms was created based on the work of Caporaso *et al.* [5] (Figure 4A, see Table S1 for full list). Each article in the *HIV literature dataset* was converted to plain text and scanned using this set of templates. These templates consist of 3 pieces of information: the position in the protein which has been mutated, the amino acid present in that position in the wildtype sequence of the isolate, and the non-wildtype amino acid to which the residue has been mutated. For example, consider the sentence "*As reported previously, S52A and S56A mutations of Vpu had no effect on virus release*" [13]. S52A and S56A refer to the experimental mutagenesis of a serine to an alanine at position 52 and 56 in the Vpu protein.

Mapping of mutations to the HIV proteome

The annotation of a mutation text-mined from papers in the *HIV literature dataset* requires three piece of information: the sequence of the isolate used in the study; the protein containing the mutation; and the position of the mutation within the protein. This information is sufficient to map a mutation to a reference HIV-1 proteome, but cannot always be directly extracted from the text-mined paper. The nomenclature for describing isolates, genes, proteins, chains and domains have not been standardised. Therefore, mapping dictionaries for HIV isolates and HIV proteins were constructed. The *isolate mapping dictionary* was constructed from isolate names and their synonyms retrieved from HIV data within the UniProt [3] and Allie [14] resources (Table S2). The *protein mapping dictionary* was constructed from synonyms for genes, proteins, cleavage products, chains and domain names from the HIV data, also retrieved from the UniProt [3] and Allie [14] resources (Table S3). The highly-studied HIV group M subtype B HXB2 isolate was selected as the reference proteome and all HIV genes, proteins, cleavage products, chains and domain names, and their synonyms, were mapped onto the 9 proteins of the isolate (Gag, Gag-Pol, Env, Tat, Nef, Rev, Vif, Vpr, Vpu). This mapping included normalised start positions to correct the inconsistent numbering schemes of cleavage products, chains and domain (Figure 4B). Both dictionaries were further manually curated to improve upon the computationally retrieved mapping.

Several different *experimental isolates* are commonly used in HIV experiments. Each paper in the *HIV literature dataset* was scanned using the contents of the *isolate mapping dictionary* to identify the *experimental isolate* used in the study (Table S2). If no isolate information was retrieved, the Human immunodeficiency virus type 1 group M subtype B HXB2 isolate was set as the *experimental isolate* for the paper. The numbering of a mutation in the HIV literature can refer to the numbering of a protein, chain, domain or cleavage product, consequently, for a defined mutation numbering a conclusive mapping may not possible. Inconclusively mapped mutations text-mined from the *HIV literature dataset* were mapped to the HIV proteome using a *co-occurrence based approach* (Figure 4C). The *co-occurrence based approach* utilised the Reflect tool for automated tagging of biological entities to scan mutation-containing sentences for protein identifying terms from the *protein mapping dictionary* [4]. Each mutation's position was normalised to the protein-numbering scheme of the full-length protein based on the co-occurring protein identifying terms. If the mutation wildtype amino acid matched the amino acid at the normalised mutation position in the *experimental isolate*, the mutation was retained as a mapped mutation. In the cases where no information relating to the mutated protein was available, all HIV HXB2 proteins were scanned at their full-length protein, chain and domain levels. In the case of chain and domain, a displacement factor was applied to adjust the mutation's position and map the mutation to all possible positions in the proteome. A mutation mapping score (see below) was calculated for each putative mutation mapping and the top scoring mapping was retained as the mapped position of the mutation. In the cases

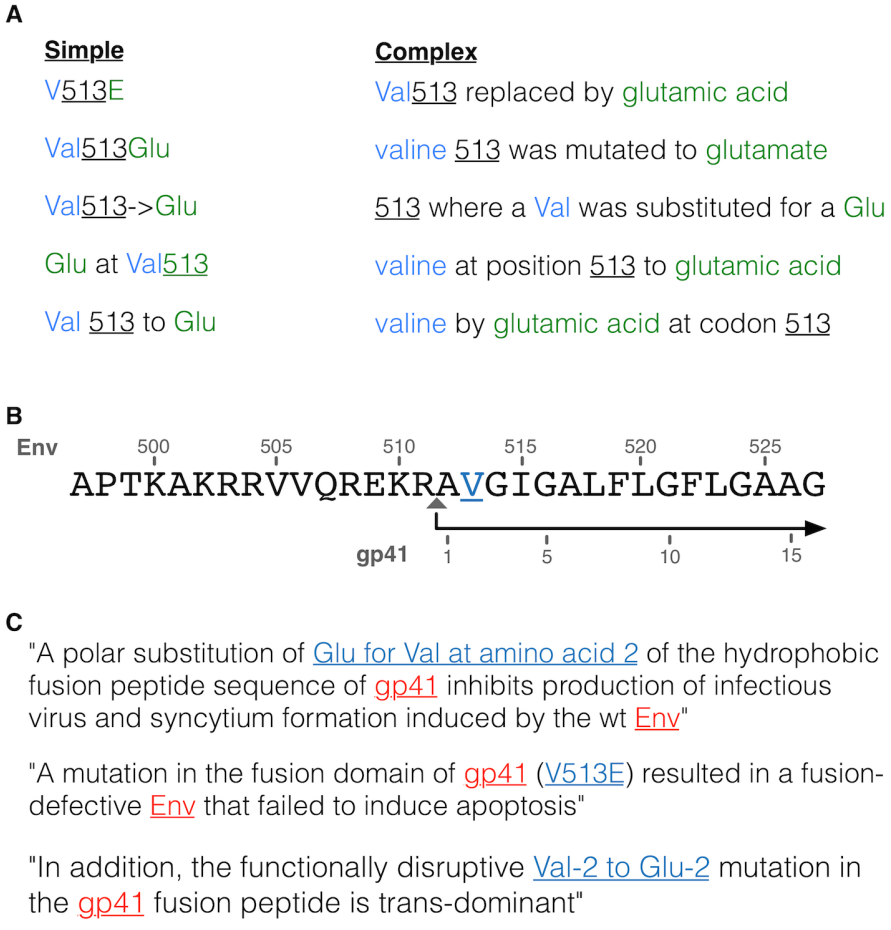


Figure 4. Example illustrations of issues associated with parsing and mapping mutation data. (A) Representative simple and complex examples of sentences recognised by the templates used to perform the mutation text-mining of articles (see Table S1 for complete list). Information of interest representing the wildtype residue (blue) and mutated residue (green) are coloured and position of the mutations are underlined. (B) Illustration of the distinct numbering schemes for different chains of the same protein. The shown peptide sequence is a short region (497–527) of the HIV Envelope glycoprotein gp160 (Env) overlapping the site cleaved by the host furin to produce the Surface protein gp120 and Transmembrane protein gp41 chains. The cleavage site is denoted by a grey triangle. The numbering above the sequence defines the position relative to the start of the gp160 protein and the numbering below the sequence defines the position relative to the start of the gp41 chain. (C) Examples of three sentences from the HIV literature where each article uses a different nomenclature or numbering scheme (blue) to describe the same mutation at the same site, Valine at residue 513 in the gp160 protein. Each article also refers to the protein by the chain name, gp41, rather than the name of the unprocessed protein, Envelope glycoprotein gp160 (Env), used for mapping in the HIV mutation resource. One example sentence refers to the gp41 chain while utilising the numbering for the unprocessed protein.
doi:10.1371/journal.pcbi.1003951.g004

where no matches to the experimental isolate proteome were found, the search was expanded to other commonly studied HIV isolates (Table S2).

Scoring mutation matches

For each mutation mapped using the above approach, a mutation mapping score, S, is calculated. The score is the function of three parameters: the probability of a match by chance; the number of mapped mutations in the paper; and the displacement from the reference protein position numbering scheme. The score ranges from 0 to 1, with values closer to 1 representing high confidence mapping of a mutation. The top-scoring mapping was retained as the mapped position of the mutation.

The score, S, is calculated as:

$$S = a(1-P) + b\left(\frac{M}{N}\right) + c\left(1 - \frac{|d|}{(L-1)}\right)$$

where M is the number of mapped mutations in the paper, N is the total number of mutations mentioned in the paper, d is the distance between the defined mutation position and the mapped position, L is the sequence length of the protein, the values of a, b, and c are constants to weight the contribution of each parameter in the equation ($a = 0.7$, $b = 0.15$ and $c = 0.15$) and P is the probability that the mapped mutation would map to the protein by chance and is calculated as:

$$P = 1 - e(L(\log(1-p^r)))$$

where p is 0.05, the probability of matching an amino acid by chance given a 20 amino acid alphabet and assuming an equal frequency for each amino acid in the HIV proteome, and r is the number of mutations that have been mapped unambiguously to the protein.

Sources of ancillary data

The HIV Mutation Browser interface integrates information from several resources to increase the ease of interpretation of the

available HIV mutation and mutagenesis data. Conservation information is displayed using multiple sequence alignments (aligned using the MAAFT algorithm [15]) retrieved from the HIV Subtype Reference Protein sequences from the Los Alamos National Laboratory (http://www.hiv.lanl.gov/). Structural information is displayed using structures of HIV proteins retrieved from the RCSB Protein Data Bank (PDB) [6]. Intrinsic disorder predictions for the proteins are calculated using the IUPred algorithm [16]. Enzymatic active sites, sites of post-translational moiety addition, sites of proteolytic cleavage and other sites of functional importance are retrieved from the UniProt resource [3]. Short linear motif interaction interfaces are retrieved from the ELM databases [17]. An up to date list of the ancillary information used and displayed is available on the HIV Mutation Browsers website.

Supporting Information

Table S1 Regular expressions for mutation identification and harmonization. A list of templates/regular expressions, based on the work of Caporaso et al. [5], used to search the publications for sentences describing mutagenesis experiments or polymorphisms.
(XLSX)

Table S2 The nomenclature for HIV isolates. The dictionary of HIV strains/isolates and their corresponding synonyms, used to identify the HIV isolates mentioned in the publications.
(XLSX)

Table S3 The nomenclature for HIV genes, proteins, chains and domains. The dictionary of HIV genes, proteins, cleavage products, chains, domains and their synonyms.
(XLSX)

Acknowledgments

We would like to thank members of the Schneider and Briggs groups, and Aino I Järvelin for their helpful comments on the manuscript. We acknowledge Ricardo Santiago-Mozos, Björn Karges, Vladimir Kuryshev, Barbara Müller, Marc Johnson, Oliver Fackler, Mark Marsh and Anna Schneider for testing the server and their useful suggestions. We thank the participating publishers who have given permission to text-mine their collection of HIV-related literature. We are grateful to Tobias Sack for assisting us in our negotiations with the various publishers.

Author Contributions

Conceived and designed the experiments: JAGB RS. Performed the experiments: NED VPS SSM. Analyzed the data: VPS SSM. Contributed reagents/materials/analysis tools: NED VPS SSM CVM. Wrote the paper: NED VPS JAGB. Conceived and constructed the mutation retrieval and mapping software: SSM VPS RS. Designed and constructed the article retrieval software: NED VPS SSM. Designed and constructed the HIV mutation database: VPS SSM. Designed and constructed the HIV mutation browser website: NED CV TAMB. Negotiated permission for copyrighted articles: JAGB NED VPS RS.

References

1. Kallings LO (2008) The first postmodern pandemic: 25 years of HIV/AIDS. J Intern Med Mar;263(3):218–43.
2. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res Jan 1;31(1):298–303.
3. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res Jan;42(Database issue):D191–8.
4. Pafilis E, O'Donoghue SI, Jensen LJ, Horn H, Kuhn M, et al. (2009) Reflect: augmented browsing for the life scientist. Nat Biotechnol Jun;27(6):508–10.
5. Caporaso JG, Baumgartner WA Jr, Randolph DA, Cohen KB, Hunter L (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics. Jul 15;23(14):1862–5.
6. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, et al. (2013) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res Jan;41(Database issue):D475–82.
7. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, et al. (2011) Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics Feb 1;27(3):408–15.
8. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, et al. (2010) Algorithms and semantic infrastructure for mutation impact extraction and grounding. BMC Genomics Dec 2;11 Suppl 4:S24.
9. Krallinger M, Izarzugaza JM, Rodriguez-Penagos C, Valencia A (2009) Extraction of human kinase mutations from literature, databases and genotyping studies. BMC Bioinformatics Aug 27;10 Suppl 8:S1.
10. Mitchell RS, Katsura C, Skasko MA, Fitzpatrick K, Lau D, et al. (2009) Vpu antagonizes BST-2-mediated restriction of HIV-1 release via beta-TrCP and endo-lysosomal trafficking. PLoS Pathog May;5(5):e1000450.
11. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, et al. (2004) Automatic extraction of mutations from Medline and cross-validation with OMIM. Nucleic Acids Res Jan 2;32(1):135–42.
12. Witte R, Baker CJ (2007) Towards a systematic evaluation of protein mutation extraction systems. J Bioinform Comput Biol Dec;5(6):1339–59.
13. Estrabaud E, Le Rouzic E, Lopez-Vergès S, Morel M, Belaïdouni N, et al. (2007) Regulated degradation of the HIV-1 Vpu protein through a betaTrCP-independent pathway limits the release of viral particles. PLoS Pathog Jul 27;3(7):e104.
14. Yamamoto Y, Yamaguchi A, Bono H, Takagi T (2011) Allie: a database and a search service of abbreviations and long forms. Database (Oxford) Apr 15;2011:bar013.
15. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol Apr;30(4):772–80.
16. Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol Apr 8;347(4):827–39.
17. Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, et al. (2014) The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res Jan;42(Database issue):D259–66.

References

Abouelhoda, M., Issa, S. A., and Ghanem, M. (2012). Tavaxy: integrating Taverna and Galaxy workflows with cloud computing support. *BMC Bioinformatics*, 13:77. [PubMed Central:PMC3583125] [DOI:10.1186/1471-2105-13-77] [PubMed:22559942]. 29, 162

Affeldt, S., Verny, L., and Isambert, H. (2016). 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics*, 17 Suppl 2:12. [PubMed Central:PMC4959376] [DOI:10.1186/s12859-015-0856-x] [PubMed:26823190]. 27

Afgan, E., Baker, D., Coraor, N., Goto, H., Paul, I. M., Makova, K. D., Nekrutenko, A., and Taylor, J. (2011). Harnessing cloud computing with Galaxy Cloud. *Nat. Biotechnol.*, 29(11):972–974. [PubMed Central:PMC3868438] [DOI:10.1038/nbt.2028] [PubMed:22068528]. 29, 162

Afgan, E., Sloggett, C., Goonasekera, N., Makunin, I., Benson, D., Crowe, M., Gladman, S., Kowsar, Y., Pheasant, M., Horst, R., and Lonie, A. (2015). Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud. *PLoS ONE*, 10(10):e0140829. [PubMed Central:PMC4621043] [DOI:10.1371/journal.pone.0140829] [PubMed:26501966]. 27, 29

Ahluwalia, I. (2016). What Is E2B(R3)? `https://blogs.perficient.com/lifesciences/2016/06/23/what-is-e2br3/`. [Online; accessed 01-November-2017]. 22

Alfresco (2017). Free enterprise content management system for Microsoft Win-

dows and Unix-like operating systems. https://www.alfresco.com. [Online; accessed 09-November-2017]. 10

Allan, C., Burel, J. M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., Macdonald, D., Moore, W. J., Neves, C., Patterson, A., Porter, M., Tarkowska, A., Loranger, B., Avondo, J., Lagerstedt, I., Lianas, L., Leo, S., Hands, K., Hay, R. T., Patwardhan, A., Best, C., Kleywegt, G. J., Zanetti, G., and Swedlow, J. R. (2012). OMERO: flexible, model-driven data management for experimental biology. *Nat. Methods*, 9(3):245–253. [PubMed Central:PMC3437820] [DOI:10.1038/nmeth.1896] [PubMed:22373911]. 12

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics*, 8:33. [PubMed Central:PMC4482045] [DOI:10.1186/s12920-015-0108-y] [PubMed:26112054]. 25, 29

Apostol, B. L., Illes, K., Pallos, J., Bodai, L., Wu, J., Strand, A., Schweitzer, E. S., Olson, J. M., Kazantsev, A., Marsh, J. L., and Thompson, L. M. (2006). Mutant huntingtin alters MAPK signaling pathways in PC12 and striatal cells: ERK1/2 protects against mutant huntingtin-associated toxicity. *Hum. Mol. Genet.*, 15(2):273–285. [DOI:10.1093/hmg/ddi443] [PubMed:16330479]. 123

Arita, M. (2009). A pitfall of wiki solution for biological databases. *Brief. Bioinformatics*, 10(3):295–296. [DOI:10.1093/bib/bbn053] [PubMed:19060305]. 11

Arvados (2013). Arvados. A free and open source platform for big data science. http://doc.arvados.org. [Online; accessed 41-November-2017]. 29

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29. [PubMed Central:PMC3037419] [DOI:10.1038/75556] [PubMed:10802651]. 4, 42, 69, 137

Ashkenazi, A., Bento, C. F., Ricketts, T., Vicinanza, M., Siddiqi, F., Pavel, M., Squitieri, F., Hardenberg, M. C., Imarisio, S., Menzies, F. M., and Rubinsztein, D. C. (2017). Polyglutamine tracts regulate beclin 1-dependent autophagy. *Nature*, 545(7652):108–111. [PubMed Central:PMC5420314] [DOI:10.1038/nature22078] [PubMed:28445460]. 32

Athey, B. D., Braxenthaler, M., Haas, M., and Guo, Y. (2013). tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc*, 2013:6–8. [PubMed Central:PMC3814495] [PubMed:24303286]. 161

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., Roma-Mateo, C., Theodosiou, A., and Mitchell, A. L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford)*, 2012:bas019. [PubMed Central:PMC3326521] [DOI:10.1093/database/bas019] [PubMed:22508994]. 3, 42

Bakay, M., Wang, Z., Melcon, G., Schiltz, L., Xuan, J., Zhao, P., Sartorelli, V., Seo, J., Pegoraro, E., Angelini, C., Shneiderman, B., Escolar, D., Chen, Y. W., Winokur, S. T., Pachman, L. M., Fan, C., Mandler, R., Nevo, Y., Gordon, E., Zhu, Y., Dong, Y., Wang, Y., and Hoffman, E. P. (2006). Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. *Brain*, 129(Pt 4):996–1013. [DOI:10.1093/brain/awl023] [PubMed:16478798]. 32, 100, 108, 109, 110, 113

Bali, K. K., Selvaraj, D., Satagopam, V. P., Lu, J., Schneider, R., and Kuner, R. (2013). Genome-wide identification and functional analyses of microRNA signatures associated with cancer pain. *EMBO Mol Med*, 5(11):1740–1758. [PubMed Central:PMC3840489] [DOI:10.1002/emmm.201302797] [PubMed:24039159]. 82

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* [DOI:10.1093/nar/gks1193] [PubMed:23193258]. 53, 164

Barriot, R., Sherman, D. J., and Dutour, I. (2007). How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. *BMC Bioinformatics*, 8:332. [PubMed Central:PMC2206060] [DOI:10.1186/1471-2105-8-332] [PubMed:17848190]. 69, 70, 71

Bart, T. (2003). Comparison of electronic data capture with paper data collection. `http://www.dreamslab.it/doc/eclinica.pdf`. [Online; accessed 01-November-2017]. 21

BaseCamp (2017). Web-based project collaboration and management. `https://basecamp.com`. [Online; accessed 09-November-2017]. 10

Bates, G. (2003). Huntingtin aggregation and toxicity in Huntington's disease. *Lancet*, 361(9369):1642–1644. [DOI:10.1016/S0140-6736(03)13304-1] [PubMed:12747895]. 116

Bauch, A., Adamczyk, I., Buczek, P., Elmer, F. J., Enimanev, K., Glyzewski, P., Kohler, M., Pylak, T., Quandt, A., Ramakrishnan, C., Beisel, C., Malmstrom, L., Aebersold, R., and Rinn, B. (2011). openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, 12:468. [PubMed Central:PMC3275639] [DOI:10.1186/1471-2105-12-468] [PubMed:22151573]. 12

Bellary, S., Krishnankutty, B., and Latha, M. S. (2014). Basics of case report form designing in clinical research. *Perspect Clin Res*, 5(4):159–166. [PubMed Central:PMC4170533] [DOI:10.4103/2229-3485.140555] [PubMed:25276625]. 20

Bender, E. (2015). Big data in biomedicine: 4 big questions. *Nature*, 527(7576):S19. [DOI:10.1038/527S19a] [PubMed:26536221]. 17, 159, 161

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300. [MathSciNet:MR1869245]. 66, 71

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). GenBank. *Nucleic Acids Res.*, 39(Database issue):D32–37. [PubMed Central:PMC3013681] [DOI:10.1093/nar/gkq1079] [PubMed:21071399]. 3, 42

Bergmann, F. T., Cooper, J., Le Novere, N., Nickerson, D., and Waltemath, D. (2015). Simulation Experiment Description Markup Language (SED-ML) Level 1 Version 2. *J Integr Bioinform*, 12(2):262. [DOI:10.2390/biecoll-jib-2015-262] [PubMed:26528560]. 5

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, 35(Database issue):D301–303. [PubMed Central:PMC1669775] [DOI:10.1093/nar/gkl971] [PubMed:17142228]. 3, 42, 121, 137

betaJUDO (2015a). Beta-cell function in juvenile diabetes and obesity. `http://www.betajudo.org/joomla`. [Online; accessed 22-August-2015]. 82, 86

betaJUDO (2015b). Beta-cell function in juvenile diabetes and obesity. `http://ec.europa.eu/research/health/medical-research/diabetes-and-obesity/projects/beta-judo_en.html`. [Online; accessed 24-August-2015]. 85

Beulke, D. (2011). Big Data Impacts Data Management: The 5 Vs of Big Data. [URL:The 5 Vs of Big Data]. 6

Bierkens, M., van der Linden, W., van Bochove, K. and. Ward, W., Remond, F. J. A., Rita, A., Jan-Willem, B., Jeroen, B., and Gerrit, M. A. (2015). tranSMART. *J Clin Bioinforma*, 5:S9. 12, 27, 161

Bioalma (2010). Bioalma. `http://healthinformatics.wikispaces.com/Bioalma`. [Online; accessed 10-August-2015]. 42

BioMedBridges (2014). BioMedBridges: Building data bridges from biology to medicine in Europe. `http://www.biomedbridges.eu`. [Online; accessed 31-October-2017]. 18

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T. (2009). Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 10–18, Stroudsburg, PA, USA. Association for Computational Linguistics. 14

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:1–21. [PubMed Central:PMC4264107] [DOI:10.1002/0471142727.mb1910s89] [PubMed:20069535]. 160, 163

Blond, W. and de Bruijn, F. (2015). OpenClinica and RedCap conversion to transMART. `https://github.com/CTMM-TraIT/trait_odm_to_i2b2`. [Online; accessed 23-August-2017]. 27, 161

Bolton, E., Wang, Y., Thiessen, P., and Bryant, S. (2008). *PubChem: Integrated Platform of Small Molecules and Biological Activities*, volume 4, chapter Chapter 12 IN Annual Reports in Computational Chemistry. American Chemical Society, Washington, DC. [URL:PubChem]. 3, 42, 121

Bonnet, E., Viara, E., Kuperstein, I., Calzone, L., Cohen, D. P., Barillot, E., and Zinovyev, A. (2015). NaviCell Web Service for network-based data visualization. *Nucleic Acids Res.*, 43(W1):W560–565. [PubMed Central:PMC4489283] [DOI:10.1093/nar/gkv450] [PubMed:25958393]. 31

Borfitz, D. (2007). Connor: 2007 will be tipping point for EDC. `http://www.bio-itworld.com/newsitems/2007/may/05-22-07-edc-forecast`. [Online; accessed 01-November-2017]. 22

Borfitz, D. (2017). Forecast: EDC Money-Making Shifts to Phase II Trials. `http://www.bio-itworld.com/bioit_article.aspx?id=35322`. [Online; accessed 01-November-2017]. 22

Borner, K., Hermle, J., Sommer, C., Brown, N. P., Knapp, B., Glass, B., Kunkel, J., Torralba, G., Reymann, J., Beil, N., Beneke, J., Pepperkok, R., Schneider, R., Ludwig, T., Hausmann, M., Hamprecht, F., Erfle, H., Kaderali, L.,

Krausslich, H. G., and Lehmann, M. J. (2010). From experimental setup to bioinformatics: an RNAi screening platform to identify host factors involved in HIV-1 replication. *Biotechnol J*, 5(1):39–49. [DOI:10.1002/biot.200900226] [PubMed:20013946]. 82

Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H. D., Hersch, S. M., Hogarth, P., Bouzou, B., Jensen, R. V., and Krainc, D. (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc. Natl. Acad. Sci. U.S.A.*, 102(31):11023–11028. [PubMed Central:PMC1182457] [DOI:10.1073/pnas.0504921102] [PubMed:16043692]. 123

Brazas, M. D., Yamada, J. T., and Ouellette, B. F. (2010). Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, 38(Web Server issue):3–6. [PubMed Central:PMC2896181] [DOI:10.1093/nar/gkq553] [PubMed:20542914]. 4

Breitwieser, F. P., Muller, A., Dayon, L., Kocher, T., Hainard, A., Pichler, P., Schmidt-Erfurth, U., Superti-Furga, G., Sanchez, J. C., Mechtler, K., Bennett, K. L., and Colinge, J. (2011). General statistical modeling of data from protein relative expression isobaric tags. *J. Proteome Res.*, 10(6):2758–2766. [DOI:10.1021/pr1012784] [PubMed:21526793]. 88

Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, 36(Database issue):D102–106. [PubMed Central:PMC2238834] [DOI:10.1093/nar/gkm955] [PubMed:18006571]. 43, 72

BSCW (2017). Be Smart Cooperate Worldwide. `https://public.bscw.de/pub/`. [Online; accessed 09-November-2017]. 10

Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., Naik, C., Pavlidis, P., Schriml, L., Good, B. M., and Su, A. I. (2016). Wikidata as a semantic framework for the Gene Wiki initiative. *Database (Oxford)*, 2016. [PubMed Central:PMC4795929] [DOI:10.1093/database/baw015] [PubMed:26989148]. 10

Cano, I., Tenyi, A., Schueller, C., Wolff, M., Huertas Miguelanez, M. M., Gomez-Cabrero, D., Antczak, P., Roca, J., Cascante, M., Falciani, F., and Maier, D. (2014). The COPD Knowledge Base: enabling data analysis and computational simulation in translational COPD research. *J Transl Med*, 12 Suppl 2:S6. [PubMed Central:PMC4255911] [DOI:10.1186/1479-5876-12-S2-S6] [PubMed:25471253]. 26

Canuel, V., Rance, B., Avillach, P., Degoulet, P., and Burgun, A. (2015). Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief. Bioinformatics*, 16(2):280–290. [PubMed Central:PMC4364065] [DOI:10.1093/bib/bbu006] [PubMed:24608524]. 17, 26, 159, 184

Capell, B. C., Erdos, M. R., Madigan, J. P., Fiordalisi, J. J., Varga, R., Conneely, K. N., Gordon, L. B., Der, C. J., Cox, A. D., and Collins, F. S. (2005). Inhibiting farnesylation of progerin prevents the characteristic nuclear blebbing of Hutchinson-Gilford progeria syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 102(36):12879–12884. [PubMed Central:PMC1200293] [DOI:10.1073/pnas.0506001102] [PubMed:16129833]. 102, 112

Capell, B. C., Olive, M., Erdos, M. R., Cao, K., Faddah, D. A., Tavarez, U. L., Conneely, K. N., Qu, X., San, H., Ganesh, S. K., Chen, X., Avallone, H., Kolodgie, F. D., Virmani, R., Nabel, E. G., and Collins, F. S. (2008). A farnesyltransferase inhibitor prevents both the onset and late progression of cardiovascular disease in a progeria mouse model. *Proc. Natl. Acad. Sci. U.S.A.*, 105(41):15902–15907. [PubMed Central:PMC2562418] [DOI:10.1073/pnas.0807840105] [PubMed:18838683]. 102, 112

Caporaso, J. G., Baumgartner, W. A., Randolph, D. A., Cohen, K. B., and Hunter, L. (2007). MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865. [PubMed Central:PMC2516306] [DOI:10.1093/bioinformatics/btm235] [PubMed:17495998]. 15, 147

Cattaneo, E., Rigamonti, D., Goffredo, D., Zuccato, C., Squitieri, F., and Sipione,

S. (2001). Loss of normal huntingtin function: new developments in Huntington's disease research. *Trends Neurosci.*, 24(3):182–188. [PubMed:11182459]. 116

CDISC (2017). CDISC - Clinical Data Interchange Standards Consortium. `https://www.cdisc.org`. [Online; accessed 31-Oct-2017]. 18, 20, 22

CDISC-Consortium (2017). SDTM - Study Data Tabulation Model. `https://www.cdisc.org/standards/foundational/sdtm`. [Online; accessed 12-Sep-2017]. 10

Cejuela, J. M., Bojchevski, A., Uhlig, C., Bekmukhametov, R., Kumar Karn, S., Mahmuti, S., Baghudana, A., Dubey, A., Satagopam, V. P., and Rost, B. (2017). nala: text mining natural language mutation mentions. *Bioinformatics*, 33(12):1852–1858. [DOI:10.1093/bioinformatics/btx083] [PubMed:28200120]. 14, 15

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., and Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5):401–404. [PubMed Central:PMC3956037] [DOI:10.1158/2159-8290.CD-12-0095] [PubMed:22588877]. 12, 26

Chalmel, F. and Primig, M. (2008). The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. *BMC Bioinformatics*, 9:86. [PubMed Central:PMC2375118] [DOI:10.1186/1471-2105-9-86] [PubMed:18254954]. 12

Chowdhury, A., Satagopam, V. P., Manukyan, L., Artemenko, K. A., Fung, Y. M., Schneider, R., Bergquist, J., and Bergsten, P. (2013). Signaling in insulin-secreting MIN6 pseudoislets and monolayer cells. *J. Proteome Res.*, 12(12):5954–5962. [DOI:10.1021/pr400864w] [PubMed:24006944]. 82

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., Jang, M., Juhos, S.,

Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Plaister, S., Radhakrishnan, R., Robinson, S., Sobhany, S., Hoopen, P. T., Vaughan, R., Zalunin, V., and Birney, E. (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37(Database issue):19–25. [PubMed Central:PMC2686451] [DOI:10.1093/nar/gkn765] [PubMed:18978013]. 3, 42, 137

Cohen, K. B. and Hunter, L. (2008). Getting started in text mining. *PLoS Comput. Biol.*, 4(1):e20. [PubMed Central:PMC2217579] [DOI:10.1371/journal.pcbi.0040020] [PubMed:18225946]. 14

Cohrs, R. J., Martin, T., Ghahramani, P., Bidaut, L., Higgins, P. J., and Shahzad, A. (2015). Translational medicine definition by the european society for translational medicine. *New Horizons in Translational Medicine*, 2(3):86–88. [DOI:10.1016/j.nhtm.2014.12.002]. 16, 158, 183

Costa, F. F. (2014). Big data in biomedicine. *Drug Discov. Today*, 19(4):433–440. [DOI:10.1016/j.drudis.2013.10.012] [PubMed:24183925]. 17, 25, 29, 159, 184

Courtot, M., Juty, N., Knupfer, C., Waltemath, D., Zhukova, A., Drager, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D. B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., Pocock, M., Rodriguez, N., Villeger, A., Wilkinson, D. J., Wimalaratne, S., Laibe, C., Hucka, M., and Le Novere, N. (2011). Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.*, 7:543. [PubMed Central:PMC3261705] [DOI:10.1038/msb.2011.77] [PubMed:22027554]. 13

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.*, 42(Database issue):D472–477. [PubMed Central:PMC3965010] [DOI:10.1093/nar/gkt1102] [PubMed:24243840]. 31

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue):D691–697. [PubMed Central:PMC3013646] [DOI:10.1093/nar/gkq1018] [PubMed:21067998]. 4, 42, 64, 121, 137

Crook, Z. R. and Housman, D. E. (2013). Surveying the landscape of Huntington's disease mechanisms, measurements, and medicines. *J Huntingtons Dis*, 2(4):405–436. [DOI:10.3233/JHD-130072] [PubMed:25062729]. 115

Csoka, A. B., English, S. B., Simkevich, C. P., Ginzinger, D. G., Butte, A. J., Schatten, G. P., Rothman, F. G., and Sedivy, J. M. (2004). Genome-scale expression profiling of Hutchinson-Gilford progeria syndrome reveals widespread transcriptional misregulation leading to mesodermal/mesenchymal defects and accelerated atherosclerosis. *Aging Cell*, 3(4):235–243. [DOI:10.1111/j.1474-9728.2004.00105.x] [PubMed:15268757]. 103, 108, 109, 110, 113

Cuellar, A., Hedley, W., Nelson, M., Lloyd, C., Halstead, M., Bullivant, D., Nickerson, D., Hunter, P., and Nielsen, P. (2015). The CellML 1.1 Specification. *J Integr Bioinform*, 12(2):259. [DOI:10.2390/biecoll-jib-2015-259] [PubMed:26528557]. 5

Dada, J. O., Spasi?, I., Paton, N. W., and Mendes, P. (2010). SBRML: a markup language for associating systems biology data with models. *Bioinformatics*, 26(7):932–938. [DOI:10.1093/bioinformatics/btq069] [PubMed:20176582]. 5

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, A., and Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences.*, 32(3):244–255. [DOI:10.1021/ci00007a012]. 69

DATATRAK (2017). Electronic Data Capture and Medical Coding . `http:`

//www.datatrak.com/home/solutions/edc-medical/. [Online; accessed 01-November-2017]. 23

Davey, N. E., Satagopam, V. P., Santiago-Mozos, S., Villacorta-Martin, C., Bharat, T. A., Schneider, R., and Briggs, J. A. (2014). The HIV mutation browser: a resource for human immunodeficiency virus mutagenesis and polymorphism data. *PLoS Comput. Biol.*, 10(12):e1003951. [PubMed Central:PMC4256008] [DOI:10.1371/journal.pcbi.1003951] [PubMed:25474213]. 149, 150, 151, 152, 154, 156, 157

Davies, B. S., Fong, L. G., Yang, S. H., Coffinier, C., and Young, S. G. (2009). The posttranslational processing of prelamin A and disease. *Annu Rev Genomics Hum Genet*, 10:153–174. [PubMed Central:PMC2846822] [DOI:10.1146/annurev-genom-082908-150150] [PubMed:19453251]. 102

Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R., and Sanchez, J. C. (2008). Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Anal. Chem.*, 80(8):2921–2931. [DOI:10.1021/ac702422x] [PubMed:18312001]. 87

de Bono, B., Grenon, P., and Sammut, S. J. (2012). ApiNATOMY: a novel toolkit for visualizing multiscale anatomy schematics with phenotype-related information. *Hum. Mutat.*, 33(5):837–848. [PubMed:22616108]. 17, 159, 184

De Sandre-Giovannoli, A., Bernard, R., Cau, P., Navarro, C., Amiel, J., Boccaccio, I., Lyonnet, S., Stewart, C. L., Munnich, A., Le Merrer, M., and Levy, N. (2003). Lamin a truncation in Hutchinson-Gilford progeria. *Science*, 300(5628):2055. [DOI:10.1126/science.1084125] [PubMed:12702809]. 101, 102

de Wilde, J., Mohren, R., van den Berg, S., Boekschoten, M., Dijk, K. W., de Groot, P., Muller, M., Mariman, E., and Smit, E. (2008). Short-term high fat-feeding results in morphological and metabolic adaptations in the skeletal muscle of C57BL/6J mice. *Physiol. Genomics*, 32(3):360–369. [DOI:10.1152/physiolgenomics.00219.2007] [PubMed:18042831]. 56

Dechat, T., Shimi, T., Adam, S. A., Rusinol, A. E., Andres, D. A., Spielmann, H. P., Sinensky, M. S., and Goldman, R. D. (2007). Alterations in mitosis

and cell cycle progression caused by a mutant lamin A known to accelerate human aging. *Proc. Natl. Acad. Sci. U.S.A.*, 104(12):4955–4960. [PubMed Central:PMC1829246] [DOI:10.1073/pnas.0700854104] [PubMed:17360326]. 112

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36(Database issue):D344–350. [PubMed Central:PMC2238832] [DOI:10.1093/nar/gkm791] [PubMed:17932057]. 3, 42

Di Prospero, N. A. and Fischbeck, K. H. (2005). Therapeutics development for triplet repeat expansion diseases. *Nat. Rev. Genet.*, 6(10):756–765. [DOI:10.1038/nrg1690] [PubMed:16205715]. 115

Dominguez, V., Raimondi, C., Somanath, S., Bugliani, M., Loder, M. K., Edling, C. E., Divecha, N., da Silva-Xavier, G., Marselli, L., Persaud, S. J., Turner, M. D., Rutter, G. A., Marchetti, P., Falasca, M., and Maffucci, T. (2011). Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic beta cells. *J. Biol. Chem.*, 286(6):4216–4225. [PubMed Central:PMC3039383] [DOI:10.1074/jbc.M110.200295] [PubMed:21127054]. 53, 55

Dooley, R., Vaughn, M., Stanzione, D., Terry, S., and Skidmore, E. (2012). Software-as-a-Service: The iPlant Foundation API. . *In:5th IEEE Workshop on Many-Task Computing Grids and Supercomputers (MTAGS)*. 29

Doughty, E., Kertesz-Farkas, A., Bodenreider, O., Thompson, G., Adadey, A., Peterson, T., and Kann, M. G. (2011). Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, 27(3):408–415. [PubMed Central:PMC3031038] [DOI:10.1093/bioinformatics/btq667] [PubMed:21138947]. 155, 182

Dove, E. S., Joly, Y., Tasse, A. M., Knoppers, B. M., Burton, P., Chisholm, R., Fortier, I., Goodwin, P., Harris, J., Hveem, K., Kaye, J., Kent, A., Knoppers, B. M., Lindpaintner, K., Little, J., Riegman, P., Ripatti, S., Stolk, R., Bobrow, M., Cambon-Thomsen, A., Dressler, L., Joly, Y., Kato, K., Knoppers,

B. M., Rodriguez, L. L., McPherson, T., Nicolas, P., Ouellette, F., Romeo-Casabona, C., Sarin, R., Wallace, S., Wiesner, G., Wilson, J., Zeps, N., Simkevitz, H., and De Rienzo, A. (2015). Genomic cloud computing: legal and ethical points to consider. *Eur. J. Hum. Genet.*, 23(10):1271–1278. [PubMed Central:PMC4592072] [DOI:10.1038/ejhg.2014.196] [PubMed:25248396]. 26

Down, T. A., Piipari, M., and Hubbard, T. J. (2011). Dalliance: interactive genome viewing on the web. *Bioinformatics*, 27(6):889–890. [PubMed Central:PMC3051325] [DOI:10.1093/bioinformatics/btr020] [PubMed:21252075]. 27, 161

Dreher, F., Kreitler, T., Hardt, C., Kamburov, A., Yildirimman, R., Schellander, K., Lehrach, H., Lange, B. M., and Herwig, R. (2012). DIPSBC–data integration platform for systems biology collaborations. *BMC Bioinformatics*, 13:85. [PubMed Central:PMC3424966] [DOI:10.1186/1471-2105-13-85] [PubMed:22568834]. 12

Dropbox (2017). File hosting service operated by American company Dropbox, Inc. `https://www.dropbox.com`. [Online; accessed 09-November-2017]. 10

Drupal (2017). Drupal - Open Source Content Management Systems. `https://www.drupal.org`. [Online; accessed 09-November-2017]. 10

EGroupware (2017). Free open source groupware software intended for businesses from small to enterprises. `http://www.egroupware.org`. [Online; accessed 09-November-2017]. 10

El Emam, K., Jonker, E., Sampson, M., Krleza-JeriÄ, K., and Neisa, A. (2009). The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials. *J. Med. Internet Res.*, 11(1):e8. [PubMed Central:PMC2762772] [DOI:10.2196/jmir.1120] [PubMed:19275984]. 22

Elefsinioti, A. L., Bagos, P. G., Spyropoulos, I. C., and Hamodrakas, S. J. (2004). A database for G proteins and their interaction with GPCRs. *BMC Bioinformatics*, 5:208. [PubMed Central:PMC544346] [DOI:10.1186/1471-2105-5-208] [PubMed:15619328]. 133, 179

EMIF (2013). IMI-EMIF: Innovative Medicines Initiative, European Medical Information Framework. http://www.emif.eu. [Online; accessed 31-October-2017]. 18

ENCODE-Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, 9(4):e1001046. [PubMed Central:PMC3079585] [DOI:10.1371/journal.pbio.1001046] [PubMed:21526222]. 41, 98, 180

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584. [PubMed Central:PMC101833] [PubMed:11917018]. 60, 64

Ensembl (2009). Protein trees and orthologies. http://www.ensembl.org/info/genome/compara/homology_method.html. [Online; accessed 17-August-2015]. 63

Eriksson, M., Brown, W. T., Gordon, L. B., Glynn, M. W., Singer, J., Scott, L., Erdos, M. R., Robbins, C. M., Moses, T. Y., Berglund, P., Dutra, A., Pak, E., Durkin, S., Csoka, A. B., Boehnke, M., Glover, T. W., and Collins, F. S. (2003). Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature*, 423(6937):293–298. [DOI:10.1038/nature01629] [PubMed:12714972]. 101, 102

eTRIKS consortium (2017). European Translational Information and Knowledge Management Services. https://www.etriks.org. [Online; accessed 09-November-2017]. 12

Etzold, T. and Verde, G. (1997). Using views for retrieving data from extremely heterogeneous databanks. *Pac Symp Biocomput*, pages 134–141. [PubMed:9390286]. 8, 39, 137, 178

FDA (2002). General Principles of Software Validation; Final Guidance for Industry and FDA Staff. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm085371.pdf. [Online; accessed 01-November-2017]. 22

FDA (2003). Part 11, Electronic Records; Electronic Signatures ? Scope and Application. https://www.fda.gov/RegulatoryInformation/Guidances/ucm125067.htm. [Online; accessed 01-November-2017]. 22

FDA (2007). Guidance for Industry Computerized Systems Used in Clinical Investigations . https://www.fda.gov/OHRMS/DOCKETS/98fr/04d-0440-gdl0002.pdf. [Online; accessed 01-November-2017]. 22

FDA (2014). E2B(R3) Electronic Transmission of Individual Case Safety Reports (ICSRs) Implementation Guide - Data Elements and Message Specification. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm275638.pdf. [Online; accessed 01-November-2017]. 22

FDA (2017). CFR - Code of Federal Regulations Title 21. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=11. [Online; accessed 01-November-2017]. 22

Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., and Westbrook, J. (2004). Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13):2153–2155. [DOI:10.1093/bioinformatics/bth214] [PubMed:15059838]. 3, 42

Fisher, R. A. (1922). On the interpretation of ?2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94. [DOI:10.2307/2340521]. 66, 71

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J., Parker, A., Proctor,

G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. (2012). Ensembl 2012. *Nucleic Acids Res.*, 40(Database issue):84–90. [PubMed Central:PMC3245178] [DOI:10.1093/nar/gkr991] [PubMed:22086963]. 3, 39, 42, 137

Fong, L. G., Frost, D., Meta, M., Qiao, X., Yang, S. H., Coffinier, C., and Young, S. G. (2006). A protein farnesyltransferase inhibitor ameliorates disease in a mouse model of progeria. *Science*, 311(5767):1621–1623. [DOI:10.1126/science.1124875] [PubMed:16484451]. 102

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D. (2015). How big data can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165:234 – 246. 6

Fuchs, T. (2010). script.aculo.us JavaScript library. `http://script.aculo.us`. [Online; accessed 14-August-2015]. 45

Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M., Diederich, N., Buttini, M., Kodama, A., Satagopam, V. P., Eifes, S., Del Sol, A., Schneider, R., Kitano, H., and Balling, R. (2014). Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol. Neurobiol.*, 49(1):88–102. [PubMed Central:PMC4153395] [DOI:10.1007/s12035-013-8489-4] [PubMed:23832570]. 30, 31, 135, 136, 139, 160, 161, 163, 178

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, 39(Database issue):D876–882. [PubMed Central:PMC3242726] [DOI:10.1093/nar/gkq963] [PubMed:20959295]. 3, 42, 137

Gainer, V., Hackett, K., Mendis, M., Kuttan, R., Pan, W., Phillips, L. C., Chueh, H. C., and Murphy, S. (2007). Using the i2b2 hive for clinical discovery: an example. *AMIA Annu Symp Proc*, page 959. [PubMed:18694059]. 27

Galperin, M. Y., Fernandez-Suarez, X. M., and Rigden, D. J. (2017). The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res.*, 45(D1):D1–D11. [PubMed Central:PMC5210597] [DOI:10.1093/nar/gkw1188] [PubMed:28053160]. 4, 137

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., and Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*, 6(269):pl1. [PubMed Central:PMC4160307] [DOI:10.1126/scisignal.2004088] [PubMed:23550210]. 12

Garrett, J. J. (2005). Ajax: A New Approach to Web Applications. `https://web.archive.org/web/20080702075113/http://www.adaptivepath.com/ideas/essays/archives/000385.php`. [Online; accessed 14-August-2015]. 45

Garuda (2015). GARUDA - The way biology connects. `http://www.garuda-alliance.org`. [Online; accessed 24-August-2015]. 82, 84

Garvey, T. D., Lincoln, P., Pedersen, C. J., Martin, D., and Johnson, M. (2003). BioSPICE: access to the most current computational tools for biologists. *OMICS*, 7(4):411–420. [DOI:10.1089/153623103322637715] [PubMed:14683613]. 12

Gattiker, A., Hermida, L., Liechti, R., Xenarios, I., Collin, O., Rougemont, J., and Primig, M. (2009). MIMAS 3.0 is a Multiomics Information Management and Annotation System. *BMC Bioinformatics*, 10:151. [PubMed Central:PMC2694794] [DOI:10.1186/1471-2105-10-151] [PubMed:19450266]. 12

Gawron, P., Ostaszewski, M., Satagopam, V., Gebel, S., Mazein, A., Kuzma, M., Zorzan, S., McGee, F., Otjacques, B., Balling, R., and Schneider, R.

(2016). MINERVA-a platform for visualization and curation of molecular interaction networks. *NPJ Syst Biol Appl*, 2:16020. [PubMed Central:PMC5516855] [DOI:10.1038/npjsba.2016.20] [PubMed:28725475]. 31, 135, 136, 141, 161, 163, 178

Gawronska, B., Erlendsson, B., and Olsson, B. (2005). Tracking biological relations in texts: a referent grammar based approach. In *Proceedings of the Workshop Biomedical Ontologies and Text Processing, ECCB 2005*, pages 15–22. 14

GenomicsEngland (2017). The 100,000 Genomes Project by numbers. `https://www.genomicsengland.co.uk/the-100000-genomes-project-by-numbers/`. [Online; accessed 17-October-2017]. 8

Gerasch, A., Faber, D., Kuntzer, J., Niermann, P., Kohlbacher, O., Lenhof, H. P., and Kaufmann, M. (2014). BiNA: a visual analytics tool for biological network data. *PLoS ONE*, 9(2):e87397. [PubMed Central:PMC3923765] [DOI:10.1371/journal.pone.0087397] [PubMed:24551056]. 30

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K. Y., and Kitano, H. (2011). Software for systems biology: from tools to integrated platforms. *Nat. Rev. Genet.*, 12(12):821–832. [DOI:10.1038/nrg3096] [PubMed:22048662]. 83, 176

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, 15(10):1451–1455. [PubMed Central:PMC1240089] [DOI:10.1101/gr.4086505] [PubMed:16169926]. 160, 163

Giles, J. (2007). Key biology databases go wiki. *Nature*, 445(7129):691. [DOI:10.1038/445691a] [PubMed:17301755]. 10

Gish, W. (1996). WU-BLAST. `http://blast.wustl.edu`. [Online; accessed 17-August-2015]. 59, 74

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457. [PubMed Central:PMC3436816] [DOI:10.1093/bioinformatics/bts389] [PubMed:22962466]. 79

Glaab, E. and Schneider, R. (2012). PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data. *Bioinformatics*, 28(3):446–447. [PubMed Central:PMC3268235] [DOI:10.1093/bioinformatics/btr656] [PubMed:22123829]. 30

Glass, C. K., Saijo, K., Winner, B., Marchetto, M. C., and Gage, F. H. (2010). Mechanisms underlying inflammation in neurodegeneration. *Cell*, 140(6):918–934. [PubMed Central:PMC2873093] [DOI:10.1016/j.cell.2010.02.016] [PubMed:20303880]. 168

Gluck, F., Hoogland, C., Antinori, P., Robin, X., Nikitin, F., Zufferey, A., Pasquarello, C., Fetaud, V., Dayon, L., Muller, M., Lisacek, F., Geiser, L., Hochstrasser, D., Sanchez, J. C., and Scherl, A. (2013). EasyProt–an easy-to-use graphical platform for proteomics data analysis. *J Proteomics*, 79:146–160. [DOI:10.1016/j.jprot.2012.12.012] [PubMed:23277275]. 88

Glynn, M. W. and Glover, T. W. (2005). Incomplete processing of mutant lamin A in Hutchinson-Gilford progeria leads to nuclear abnormalities, which are reversed by farnesyltransferase inhibition. *Hum. Mol. Genet.*, 14(20):2959–2969. [DOI:10.1093/hmg/ddi326] [PubMed:16126733]. 112

Goecks, J., Nekrutenko, A., Taylor, J., Afgan, E., Ananda, G., Baker, D., Blankenberg, D., Chakrabarty, R., Coraor, N., Goecks, J., Von Kuster, G., Lazarus, R., Li, K., Nekrutenko, A., Taylor, J., and Vincent, K. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11(8):R86. [PubMed Central:PMC2945788] [DOI:10.1186/gb-2010-11-8-r86] [PubMed:20738864]. 29, 160, 162

Good, B. M., Clarke, E. L., Loguercio, S., and Su, A. I. (2012). Building a biomedical semantic network in Wikipedia with Semantic Wiki

Links. *Database (Oxford)*, 2012:bar060. [PubMed Central:PMC3308151] [DOI:10.1093/database/bar060] [PubMed:22434829]. 10

Google (2015a). Google analytics standard. `http://www.google.com/analytics/`. [Online; accessed 23-August-2015]. 83

Google (2015b). Interactive charts for browsers and mobile devices. `https://developers.google.com/chart/`. [Online; accessed 18-August-2015]. 72

Google (2017). File storage and synchronization service developed by Google. `https://www.google.com/drive/`. [Online; accessed 09-November-2017]. 10

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, 20(5):565–577. [PubMed Central:PMC2860159] [DOI:10.1101/gr.104471.109] [PubMed:20363979]. 72

Guberman, J. M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R. J., Di Genova, A., Forbes, S., Fujisawa, T., Gadaleta, E., Goodstein, D. M., Gundem, G., Haggarty, B., Haider, S., Hall, M., Harris, T., Haw, R., Hu, S., Hubbard, S., Hsu, J., Iyer, V., Jones, P., Katayama, T., Kinsella, R., Kong, L., Lawson, D., Liang, Y., Lopez-Bigas, N., Luo, J., Lush, M., Mason, J., Moreews, F., Ndegwa, N., Oakley, D., Perez-Llamas, C., Primig, M., Rivkin, E., Rosanoff, S., Shepherd, R., Simon, R., Skarnes, B., Smedley, D., Sperling, L., Spooner, W., Stevenson, P., Stone, K., Teague, J., Wang, J., Wang, J., Whitty, B., Wong, D. T., Wong-Erasmus, M., Yao, L., Youens-Clark, K., Yung, C., Zhang, J., and Kasprzyk, A. (2011). BioMart Central Portal: an open database network for the biological community. *Database (Oxford)*, 2011:bar041. [PubMed Central:PMC3263598] [DOI:10.1093/database/bar041] [PubMed:21930507]. 39

Gunther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., Schneider, R., Skoblo, R., Russell, R. B., Bourne, P. E., Bork, P., and Preissner, R. (2008). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic*

Acids Res., 36(Database issue):D919–922. [PubMed Central:PMC2238858] [DOI:10.1093/nar/gkm862] [PubMed:17942422]. 3, 42

Guo, N. L. and Wan, Y. W. (2014). Network-based identification of biomarkers co-expressed with multiple pathways. *Cancer Inform*, 13(Suppl 5):37–47. [PubMed Central:PMC4218687] [DOI:10.4137/CIN.S14054] [PubMed:25392692]. 30

H2020 (2016). Open research data in horizon 2020. https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf. [Online; accessed 07-November-2017]. 3

Haider, N. (2012a). mol2png.pl, a perl script reads MDL MOL or SDF files, extracts the molecular structures and generates 2D graphical images in PNG format by piping the data through the mol2ps utility program and Ghostscript. http://merian.pch.univie.ac.at/~nhaider/cheminf/mol2png.pl.txt. [Online; accessed 21-May-2012]. 69

Haider, N. (2012b). mol2ps, a freeware tool for 2D depiction of molecular structures. http://merian.pch.univie.ac.at/~nhaider/cheminf/mol2ps.html. [Online; accessed 21-May-2012]. 69

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res.*, 33(Database issue):D514–517. [PubMed Central:PMC539987] [DOI:10.1093/nar/gki033] [PubMed:15608251]. 3, 42

Harmar, A. J., Hills, R. A., Rosser, E. M., Jones, M., Buneman, O. P., Dunbar, D. R., Greenhill, S. D., Hale, V. A., Sharman, J. L., Bonner, T. I., Catterall, W. A., Davenport, A. P., Delagrange, P., Dollery, C. T., Foord, S. M., Gutman, G. A., Laudet, V., Neubig, R. R., Ohlstein, E. H., Olsen, R. W., Peters, J., Pin, J. P., Ruffolo, R. R., Searls, D. B., Wright, M. W., and Spedding, M. (2009). IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.*, 37(Database issue):D680–685. [PubMed Central:PMC2686540] [DOI:10.1093/nar/gkn728] [PubMed:18948278]. 130

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*, 42(2):377–381. [PubMed Central:PMC2700030] [DOI:10.1016/j.jbi.2008.08.010] [PubMed:18929686]. 12, 23

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, 11(7):476–486. [PubMed Central:PMC3321268] [DOI:10.1038/nrg2795] [PubMed:20531367]. 17, 159, 184

Hennekam, R. C. (2006). Hutchinson-Gilford progeria syndrome: review of the phenotype. *Am. J. Med. Genet. A*, 140(23):2603–2624. [DOI:10.1002/ajmg.a.31346] [PubMed:16838330]. 32, 100

Herzinger, S. (2015). tranSMART XNAT importer. `https://github.com/transmart/SmartR`. [Online; accessed 23-August-2017]. 27, 161

Herzinger, S., Gu, W., Satagopam, V., Eifes, S., Rege, K., Barbosa Da Silva, A., and Schneider, R. (2017). SmartR: An open-source platform for interactive visual analytics for translational research data. *Bioinformatics*. [DOI:10.1093/bioinformatics/btx137] [PubMed:28334291]. 27, 161

Hirschman, L., Morgan, A. A., and Yeh, A. S. (2002). Rutabaga by any other name: extracting biological names. *J Biomed Inform*, 35(4):247–259. [PubMed:12755519]. 14

HL7 (2017). HL7: Health Level Seven International. `http://www.hl7.org/implement/standards`. [Online; accessed 31-October-2017]. 18

Ho Sui, S. J., Begley, K., Reilly, D., Chapman, B., McGovern, R., Rocca-Sera, P., Maguire, E., Altschuler, G. M., Hansen, T. A., Sompallae, R., Krivtsov, A., Shivdasani, R. A., Armstrong, S. A., Culhane, A. C., Correll, M., Sansone, S. A., Hofmann, O., and Hide, W. (2012). The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acids Res.*, 40(Database issue):D984–991. [PubMed Central:PMC3245064] [DOI:10.1093/nar/gkr1051] [PubMed:22121217]. 10

Hodges, A., Hughes, G., Brooks, S., Elliston, L., Holmans, P., Dunnett, S. B., and Jones, L. (2008). Brain gene expression correlates with changes in behavior in the R6/1 mouse model of Huntington's disease. *Genes Brain Behav.*, 7(3):288–299. [DOI:10.1111/j.1601-183X.2007.00350.x] [PubMed:17696994]. 123

Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C., and Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE*, 2005(283):pe21. [DOI:10.1126/stke.2832005pe21] [PubMed:15886388]. 14

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., and Verspoor, K. (2014). *Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges*, page 271?300. Springer Berlin Heidelberg. 14

Hood, L. and Flores, M. (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol*, 29(6):613–624. [DOI:10.1016/j.nbt.2012.03.004] [PubMed:22450380]. 16, 158

Hooper, S. D. and Bork, P. (2005). Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, 21(24):4432–4433. [DOI:10.1093/bioinformatics/bti696] [PubMed:16188923]. 79, 131

Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nat Rev Drug Discov*, 1(9):727–730. [DOI:10.1038/nrd892] [PubMed:12209152]. 124

HPA (2015). The Human Protein Atlas website. `http://www.proteinatlas.org/humanproteome`. [Online; accessed 26-August-2015]. 89

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novere, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D.,

Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531. [PubMed:12611808]. 5, 137

Hunter, A. A., Macgregor, A. B., Szabo, T. O., Wellington, C. A., and Bellgard, M. I. (2012a). Yabi: An online research environment for grid, high performance and cloud computing. *Source Code Biol Med*, 7(1):1. [PubMed Central:PMC3298538] [DOI:10.1186/1751-0473-7-1] [PubMed:22333270]. 28

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C., and Yong, S. Y. (2012b). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, 40(Database issue):D306–312. [PubMed Central:PMC3245097] [DOI:10.1093/nar/gkr948] [PubMed:22096229]. 3, 42, 137

ICH (2017). ICH: The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. `http://www.ich.org/about/mission.html`. [Online; accessed 01-November-2017]. 22

ICH-Guidelines (1996). ICH Guidance E6: Good Clinical Practice: Consolidated guideline. US HHS, US FDA, CDER, CBER. `http://www.fda.gov/downloads/Drugs/Guidances/ucm073122.pdf`. [Online; accessed 31-October-2017]. 20

Jacunski, A. and Tatonetti, N. P. (2013). Connecting the dots: applications of network medicine in pharmacology and disease. *Clin. Pharmacol. Ther.*, 94(6):659–669. [DOI:10.1038/clpt.2013.168] [PubMed:23995266]. 30

Jagla, B., Wiswedel, B., and Coppee, J. Y. (2011). Extending KNIME for next-generation sequencing data analysis. *Bioinformatics*, 27(20):2907–2909. [DOI:10.1093/bioinformatics/btr478] [PubMed:21873641]. 29

Jimenez, R. C., Quinn, A. F., Garcia, A., Labarga, A., O'Neill, K., Martinez, F., Salazar, G. A., and Hermjakob, H. (2008). Dasty2, an Ajax protein DAS client. *Bioinformatics*, 24(18):2119–2121. [DOI:10.1093/bioinformatics/btn387] [PubMed:18694895]. 56, 130

Jimenez-Sanchez, M., Lam, W., Hannus, M., Sonnichsen, B., Imarisio, S., Fleming, A., Tarditi, A., Menzies, F., Ed Dami, T., Xu, C., Gonzalez-Couto, E., Lazzeroni, G., Heitz, F., Diamanti, D., Massai, L., Satagopam, V. P., Marconi, G., Caramelli, C., Nencini, A., Andreini, M., Sardone, G. L., Caradonna, N. P., Porcari, V., Scali, C., Schneider, R., Pollio, G., O'Kane, C. J., Caricasole, A., and Rubinsztein, D. C. (2015a). siRNA screen identifies QPCT as a druggable target for Huntington's disease. *Nat. Chem. Biol.*, 11(5):347–354. [DOI:10.1038/nchembio.1790] [PubMed:25848931]. 82, 117, 118, 122, 126, 127, 128

Jimenez-Sanchez, M., Lam, W., Hannus, M., Sonnichsen, B., Imarisio, S., Fleming, A., Tarditi, A., Menzies, F., Ed Dami, T., Xu, C., Gonzalez-Couto, E., Lazzeroni, G., Heitz, F., Diamanti, D., Massai, L., Satagopam, V. P., Marconi, G., Caramelli, C., Nencini, A., Andreini, M., Sardone, G. L., Caradonna, N. P., Porcari, V., Scali, C., Schneider, R., Pollio, G., O'Kane, C. J., Caricasole, A., and Rubinsztein, D. C. (2015b). Supplementary Note 1 of siRNA screen identifies QPCT as a druggable target for Huntington's disease. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4696152/bin/NIHMS62501-supplement-Supplementary_Note_1.pdf`. [Online; accessed 26-October-2017][DOI:10.1038/nchembio.1790] [PubMed:25848931]. 117

Jimenez-Sanchez, M., Lam, W., Hannus, M., Sonnichsen, B., Imarisio, S., Fleming, A., Tarditi, A., Menzies, F., Ed Dami, T., Xu, C., Gonzalez-Couto, E., Lazzeroni, G., Heitz, F., Diamanti, D., Massai, L., Satagopam, V. P., Marconi, G., Caramelli, C., Nencini, A., Andreini, M., Sardone, G. L., Caradonna, N. P., Porcari, V., Scali, C., Schneider, R., Pollio, G.,

O'Kane, C. J., Caricasole, A., and Rubinsztein, D. C. (2015c). Supplementary DataSet 1 of siRNA screen identifies QPCT as a druggable target for Huntington's disease. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4696152/bin/NIHMS62501-supplement-DataSet1.pdf. [Online; accessed 26-October-2017][DOI:10.1038/nchembio.1790] [PubMed:25848931]. 117, 118, 127

Jimenez-Sanchez, M., Lam, W., Hannus, M., Sonnichsen, B., Imarisio, S., Fleming, A., Tarditi, A., Menzies, F., Ed Dami, T., Xu, C., Gonzalez-Couto, E., Lazzeroni, G., Heitz, F., Diamanti, D., Massai, L., Satagopam, V. P., Marconi, G., Caramelli, C., Nencini, A., Andreini, M., Sardone, G. L., Caradonna, N. P., Porcari, V., Scali, C., Schneider, R., Pollio, G., O'Kane, C. J., Caricasole, A., and Rubinsztein, D. C. (2015d). Supplementary DataSet 2 of siRNA screen identifies QPCT as a druggable target for Huntington's disease. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4696152/bin/NIHMS62501-supplement-DataSet2.pdf. [Online; accessed 26-October-2017][DOI:10.1038/nchembio.1790] [PubMed:25848931]. 118

Jimenez-Sanchez, M., Lam, W., Hannus, M., Sonnichsen, B., Imarisio, S., Fleming, A., Tarditi, A., Menzies, F., Ed Dami, T., Xu, C., Gonzalez-Couto, E., Lazzeroni, G., Heitz, F., Diamanti, D., Massai, L., Satagopam, V. P., Marconi, G., Caramelli, C., Nencini, A., Andreini, M., Sardone, G. L., Caradonna, N. P., Porcari, V., Scali, C., Schneider, R., Pollio, G., O'Kane, C. J., Caricasole, A., and Rubinsztein, D. C. (2015e). Supplementary Information of siRNA screen identifies QPCT as a druggable target for Huntington's disease. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4696152/bin/NIHMS62501-supplement-Supplementary_Information.pdf. [Online; accessed 26-October-2017][DOI:10.1038/nchembio.1790] [PubMed:25848931]. 118, 122, 126

Jimeno Yepes, A. and Verspoor, K. (2014). Literature mining of genetic variants for curation: quantifying the importance of supplementary material. *Database (Oxford)*, 2014:bau003. [PubMed Central:PMC3920087] [DOI:10.1093/database/bau003] [PubMed:24520105]. 14

Joomla (2017). Free and open-source content management system for publishing

web content. https://www.joomla.org. [Online; accessed 09-November-2017]. 10

Kacsuk, P., Farkas, Z., Kozlovszky, M., Hermann, G., Balasko, A., Karoczkai, K., and Marton, I. (2012). Ws-pgrade/guse generic dci gateway framework for a large variety of user communities. *Journal of Grid Computing*, 10(4):601?630. 29

Kallings, L. O. (2008). The first postmodern pandemic: 25 years of HIV/ AIDS. *J. Intern. Med.*, 263(3):218–243. [DOI:10.1111/j.1365-2796.2007.01910.x] [PubMed:18205765]. 146

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemela, P., Gentile, M., Scheinin, I., Koski, M., Kaki, J., and Korpelainen, E. I. (2011). Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, 12:507. [PubMed Central:PMC3215701] [DOI:10.1186/1471-2164-12-507] [PubMed:21999641]. 28

Kamal, J., Liu, J., Ostrander, M., Santangelo, J., Dyta, R., Rogers, P., and Mekhjian, H. S. (2010). Information warehouse - a comprehensive informatics platform for business, clinical, and research applications. *AMIA Annu Symp Proc*, 2010:452–456. [PubMed Central:PMC3041278] [PubMed:21347019]. 26

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30. 31

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, 38(Database issue):D355–360. [PubMed Central:PMC2808910] [DOI:10.1093/nar/gkp896] [PubMed:19880382]. 42

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res.*, 30(1):42–46. [PubMed Central:PMC99091] [PubMed:11752249]. 3, 42

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic*

Acids Res., 40(Database issue):D109–114. [PubMed Central:PMC3245020] [DOI:10.1093/nar/gkr988] [PubMed:22080510]. 4, 42, 64, 121, 137

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database (Oxford)*, 2011:bar049. [PubMed Central:PMC3215098] [DOI:10.1093/database/bar049] [PubMed:22083790]. 12

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988. [DOI:10.1002/pmic.200300721] [PubMed:15221759]. 3, 42

Kilicoglu, H. and Bergler, S. (2009). Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 119–127, Stroudsburg, PA, USA. Association for Computational Linguistics. 14

Kim, T. M. and Park, P. J. (2011). Advances in analysis of transcriptional regulatory networks. *Wiley Interdiscip Rev Syst Biol Med*, 3(1):21–35. [DOI:10.1002/wsbm.105] [PubMed:21069662]. 30

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., and Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011:bar030. [PubMed Central:PMC3170168] [DOI:10.1093/database/bar030] [PubMed:21785142]. 39, 43

Koenig, T., Menze, B. H., Kirchner, M., Monigatti, F., Parker, K. C., Patterson, T., Steen, J. J., Hamprecht, F. A., and Steen, H. (2008). Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *J. Proteome Res.*, 7(9):3708–3717. [DOI:10.1021/pr700859x] [PubMed:18707158]. 88

Koop, A. and Mosges, R. (2002). The use of handheld computers in clinical trials. *Control Clin Trials*, 23(5):469–480. [PubMed:12392861]. 21

Koster, J. and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522. [DOI:10.1093/bioinformatics/bts480] [PubMed:22908215]. 28

Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39(Database issue):D152–157. [PubMed Central:PMC3013655] [DOI:10.1093/nar/gkq1027] [PubMed:21037258]. 4, 43

Krallinger, M., Izarzugaza, J. M., Rodriguez-Penagos, C., and Valencia, A. (2009). Extraction of human kinase mutations from literature, databases and genotyping studies. *BMC Bioinformatics*, 10 Suppl 8:S1. [PubMed Central:PMC2745582] [DOI:10.1186/1471-2105-10-S8-S1] [PubMed:19758464]. 155, 182

Krallinger, M., Valencia, A., and Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, 9 Suppl 2:S8. [PubMed Central:PMC2559992] [DOI:10.1186/gb-2008-9-s2-s8] [PubMed:18834499]. 15

Kramer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–530. [PubMed Central:PMC3928520] [DOI:10.1093/bioinformatics/btt703] [PubMed:24336805]. 30

Kuhn, A., Goldstein, D. R., Hodges, A., Strand, A. D., Sengstag, T., Kooperberg, C., Becanovic, K., Pouladi, M. A., Sathasivam, K., Cha, J. H., Hannan, A. J., Hayden, M. R., Leavitt, B. R., Dunnett, S. B., Ferrante, R. J., Albin, R., Shelbourne, P., Delorenzi, M., Augood, S. J., Faull, R. L., Olson, J. M., Bates, G. P., Jones, L., and Luthi-Carter, R. (2007). Mutant huntingtin's effects on striatal gene expression in mice recapitulate changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage. *Hum. Mol. Genet.*, 16(15):1845–1861. [DOI:10.1093/hmg/ddm133] [PubMed:17519223]. 123

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., and Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, 36(Database issue):D684–688. [PubMed Central:PMC2238848] [DOI:10.1093/nar/gkm795] [PubMed:18084021]. 3, 4, 42, 43, 76, 121

Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., Dutreix, M., Barillot, E., and Zinovyev, A. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4:e160. [PubMed Central:PMC4521180] [DOI:10.1038/oncsis.2015.19] [PubMed:26192618]. 31

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Melius, J., Waagmeester, A., Sinha, S. R., Miller, R., Coort, S. L., Cirillo, E., Smeets, B., Evelo, C. T., and Pico, A. R. (2016). WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.*, 44(D1):D488–494. [PubMed Central:PMC4702772] [DOI:10.1093/nar/gkv1024] [PubMed:26481357]. 30, 31

Lajoie, P. and Snapp, E. L. (2010). Formation and toxicity of soluble polyglutamine oligomers in living cells. *PLoS ONE*, 5(12):e15245. [PubMed Central:PMC3011017] [DOI:10.1371/journal.pone.0015245] [PubMed:21209946]. 127

Landles, C. and Bates, G. P. (2004). Huntingtin and the molecular pathogenesis of Huntington's disease. Fourth in molecular medicine review series. *EMBO Rep.*, 5(10):958–963. [PubMed Central:PMC1299150] [DOI:10.1038/sj.embor.7400250] [PubMed:15459747]. 32, 115

Latimer, P. (2008). Case report form insanity. *J Clin Res Best Pract*, 4. [Available from: PDF][Online; accessed 31-October-2017]. 20

Laurila, J. B., Naderi, N., Witte, R., Riazanov, A., Kouznetsov, A., and Baker, C. J. (2010). Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC Genomics*, 11 Suppl

4:S24. [PubMed Central:PMC3005927] [DOI:10.1186/1471-2164-11-S4-S24] [PubMed:21143808]. 155, 182

Lazarus, R., Kaspi, A., and Ziemann, M. (2012). Creating reusable tools from scripts: the Galaxy Tool Factory. *Bioinformatics*, 28(23):3139–3140. [PubMed Central:PMC3509488] [DOI:10.1093/bioinformatics/bts573] [PubMed:23024011]. 162

Le Novere, N. (2006). Model storage, exchange and integration. *BMC Neurosci*, 7 Suppl 1:S11. [PubMed Central:PMC1775041] [DOI:10.1186/1471-2202-7-S1-S11] [PubMed:17118155]. 5

Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villeger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (2009). The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27(8):735–741. [DOI:10.1038/nbt.1558] [PubMed:19668183]. 137

Leaman, R. and Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput*, pages 652–663. [PubMed:18229723]. 14

Lee, J. M., Ivanova, E. V., Seong, I. S., Cashorali, T., Kohane, I., Gusella, J. F., and MacDonald, M. E. (2007). Unbiased gene expression analysis implicates the huntingtin polyglutamine tract in extra-mitochondrial energy metabolism. *PLoS Genet.*, 3(8):e135. [PubMed Central:PMC1950164] [DOI:10.1371/journal.pgen.0030135] [PubMed:17708681]. 123

Leegwater-Kim, J. and Cha, J. H. (2004). The paradigm of Huntington's disease: therapeutic opportunities in neurodegeneration. *NeuroRx*, 1(1):128–138. [PubMed Central:PMC534918] [DOI:10.1602/neurorx.1.1.128] [PubMed:15717013]. 32, 116

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Brief. Bioinformatics*, 18(3):530–536. [PubMed Central:PMC5429012] [DOI:10.1093/bib/bbw020] [PubMed:27013646]. 27, 28, 29

Lejeune, F. X., Mesrob, L., Parmentier, F., Bicep, C., Vazquez-Manrique, R. P., Parker, J. A., Vert, J. P., Tourette, C., and Neri, C. (2012). Large-scale functional RNAi screen in C. elegans identifies genes that regulate the dysfunction of mutant polyglutamine neurons. *BMC Genomics*, 13:91. [PubMed Central:PMC3331833] [DOI:10.1186/1471-2164-13-91] [PubMed:22413862]. 126

Lesnick, T. G., Papapetropoulos, S., Mash, D. C., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J. R., Rocca, W. A., Ahlskog, J. E., and Maraganore, D. M. (2007). A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.*, 3(6):e98. [PubMed Central:PMC1904362] [DOI:10.1371/journal.pgen.0030098] [PubMed:17571925]. 164

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic Acids Res.*, 37(Database issue):D229–232. [PubMed Central:PMC2686533] [DOI:10.1093/nar/gkn808] [PubMed:18978020]. 3, 42, 121

Li, J., Zhu, X., and Chen, J. Y. (2009a). Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput. Biol.*, 5(7):e1000450. [PubMed Central:PMC2709445] [DOI:10.1371/journal.pcbi.1000450] [PubMed:19649302]. 14

Li, M., Huang, Y., Ma, A. A., Lin, E., and Diamond, M. I. (2009b). Y-27632 improves rotarod performance and reduces huntingtin levels in R6/2 mice. *Neurobiol. Dis.*, 36(3):413–420. [DOI:10.1016/j.nbd.2009.06.011] [PubMed:19591939]. 124

Li, S. H. and Li, X. J. (2004). Huntingtin-protein interactions and the pathogenesis of Huntington's disease. *Trends Genet.*, 20(3):146–154. [DOI:10.1016/j.tig.2004.01.008] [PubMed:15036808]. 116

Liekens, A. M., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., and Del-Favero, J. (2011). BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol.*, 12(6):R57. [PubMed Central:PMC3218845] [DOI:10.1186/gb-2011-12-6-r57] [PubMed:21696594]. 14

Litchfield, J., Freeman, J., Schou, H., Elsley, M., Fuller, R., and Chubb, B. (2005). Is the future for clinical trials internet-based? A cluster randomized clinical trial. *Clin Trials*, 2(1):72–79. [DOI:10.1191/1740774505cn069oa] [PubMed:16279581]. 21

Lloyd, C. M., Halstead, M. D., and Nielsen, P. F. (2004). CellML: its future, present and past. *Prog. Biophys. Mol. Biol.*, 85(2-3):433–450. [DOI:10.1016/j.pbiomolbio.2004.01.004] [PubMed:15142756]. 5

Lomax, J. (2005). Get ready to GO! A biologist's guide to the Gene Ontology. *Brief. Bioinformatics*, 6(3):298–304. [PubMed:16212777]. 13

Lonzer, J. (2014). Transformative Genomics: England Begins Daunting Task of Sequencing 100,000 Genomes by 2017. `https://innovatemedtec.com/content/transformative-genomics-england-begins-daunting-task-of-sequencing-100000-genc` [Online; accessed 17-October-2017]. 8

Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., and Gish, W. (2003). WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, 31(13):3795–3798. [PubMed Central:PMC168979] [PubMed:12824421]. 59, 74

Lowe, H. J., Ferris, T. A., Hernandez, P. M., and Weber, S. C. (2009). STRIDE–An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*, 2009:391–395. [PubMed Central:PMC2815452] [PubMed:20351886]. 26

Lu, Z. and Su, J. (2010). Clinical data management: Current status, challenges, and future directions from industry perspectives. *Open Access J Clin Trials*, 2:93–105. [DOI:10.2147/OAJCT.S8172]. 20

Lucas, A. (2010). *amap: Another Multidimensional Analysis Package*. R package version 0.8-5. 74, 89

Luo, J., Wu, M., Gopukumar, D., and Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights*, 8:1–10. [PubMed Central:PMC4720168] [DOI:10.4137/BII.S31559] [PubMed:26843812]. 25

Luo, S., Mizuta, H., and Rubinsztein, D. C. (2008). p21-activated kinase 1 promotes soluble mutant huntingtin self-interaction and enhances toxicity. *Hum. Mol. Genet.*, 17(6):895–905. [DOI:10.1093/hmg/ddm362] [PubMed:18065495]. 127

Luthi-Carter, R., Strand, A. D., Hanson, S. A., Kooperberg, C., Schilling, G., La Spada, A. R., Merry, D. E., Young, A. B., Ross, C. A., Borchelt, D. R., and Olson, J. M. (2002). Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects. *Hum. Mol. Genet.*, 11(17):1927–1937. [PubMed:12165555]. 123

Ly, D. H., Lockhart, D. J., Lerner, R. A., and Schultz, P. G. (2000). Mitotic misregulation and human aging. *Science*, 287(5462):2486–2492. [PubMed:10741968]. 103

MacDonald, M. E., Gines, S., Gusella, J. F., and Wheeler, V. C. (2003). Huntington's disease. *Neuromolecular Med.*, 4(1-2):7–20. [DOI:10.1385/NMM:4:1-2:7] [PubMed:14528049]. 116

Madhavan, S., Gusev, Y., Harris, M. A., Tanenbaum, D. M., Gauba, R., Bhuvaneshwar, K., Shinohara, A., Rosso, K., Carabet, L. A., Song, L., Riggins, R. B., Dakshanamurthy, S., Wang, Y., Byers, S. W., Clarke, R., and Weiner, L. W. (2011). G-code: enabling systems medicine through innovative informatics. *Genome Biology*, 12(Suppl 1):P38. 26

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 39(Database

issue):D52–57. [PubMed Central:PMC3013746] [DOI:10.1093/nar/gkq1237] [PubMed:21115458]. 3, 42, 137

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, 2011:bar009. [PubMed Central:PMC3070428] [DOI:10.1093/database/bar009] [PubMed:21447597]. 3, 15, 42, 137

Mallampalli, M. P., Huyer, G., Bendale, P., Gelb, M. H., and Michaelis, S. (2005). Inhibiting farnesylation reverses the nuclear morphology defect in a HeLa cell model for Hutchinson-Gilford progeria syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 102(40):14416–14421. [PubMed Central:PMC1242289] [DOI:10.1073/pnas.0503712102] [PubMed:16186497]. 112

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., and Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8):1112–1118. [PubMed Central:PMC2853691] [DOI:10.1093/bioinformatics/btq099] [PubMed:20200009]. 13

Mancini, M. A., Shan, B., Nickerson, J. A., Penman, S., and Lee, W. H. (1994). The retinoblastoma gene product is a cell cycle-dependent, nuclear matrix-associated protein. *Proc. Natl. Acad. Sci. U.S.A.*, 91(1):418–422. [PubMed Central:PMC42959] [PubMed:8278403]. 108, 176

Mardis, E. R. (2010). The $1,000 *genome*, *the* 100,000 analysis? *Genome Med*, 2(11):84. [PubMed Central:PMC3016626] [DOI:10.1186/gm205] [PubMed:21114804]. 17, 159, 184

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., Mollenhauer, B., Sherer, T., Frasier, M., Meunier, C., Rudolph, A., Casaceli, C., Seibyl, J., Mendick, S., Schuff, N., Zhang, Y., Toga, A., Crawford, K., Ansbach, A., De Blasio, P., Piovella, M., Trojanowski, J., Shaw, L., Singleton, A., Hawkins, K., Eberling, J., Brooks, D., Russell, D., Leary, L., Factor, S., Sommerfeld, B., Hogarth, P., Pighetti, E., Williams, K., Standaert, D., Guthrie, S., Hauser, R.,

Delgado, H., Jankovic, J., Hunter, C., Stern, M., Tran, B., Leverenz, J., Baca, M., Frank, S., Thomas, C. A., Richard, I., Deeley, C., Rees, L., Sprenger, F., Lang, E., Shill, H., Obradov, S., Fernandez, H., Winters, A., Berg, D., Gauss, K., Galasko, D., Fontaine, D., Mari, Z., Gerstenhaber, M., Brooks, D., Malloy, S., Barone, P., Longo, K., Comery, T., Ravina, B., Grachev, I., Gallagher, K., Collins, M., Widnell, K. L., Ostrowizki, S., Fontoura, P., La-Roche, F., Ho, T., Luthman, J., van der Brug, M., Reith, A. D., and Taylor, P. (2011). The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.*, 95(4):629–635. [DOI:10.1016/j.pneurobio.2011.09.005] [PubMed:21930184]. 18

Marji, J., O'Donoghue, S. I., McClintock, D., Satagopam, V. P., Schneider, R., Ratner, D., Worman, H. J., Gordon, L. B., and Djabali, K. (2010). Defective lamin A-Rb signaling in Hutchinson-Gilford Progeria Syndrome and reversal by farnesyltransferase inhibition. *PLoS ONE*, 5(6):e11132. [PubMed Central:PMC2886113] [DOI:10.1371/journal.pone.0011132] [PubMed:20559568]. 31, 34, 82, 100, 105, 107, 108, 112, 114

Martin-Sanchez, F. and Verspoor, K. (2014). Big data in medicine is driving big changes. *Yearb Med Inform*, 9:14–20. [PubMed Central:PMC4287083] [DOI:10.15265/IY-2014-0020] [PubMed:25123716]. 161

Marzolf, B., Deutsch, E. W., Moss, P., Campbell, D., Johnson, M. H., and Galitski, T. (2006). SBEAMS-Microarray: database software supporting genomic expression analyses for systems biology. *BMC Bioinformatics*, 7:286. [PubMed Central:PMC1524999] [DOI:10.1186/1471-2105-7-286] [PubMed:16756676]. 12

Mayer, G. (2009a). Data management in systems biology I - Overview and bibliography. *CoRR*, abs/0908.0411. [URL:arXiv:0908.0411v3]. 6, 10, 11, 39

Mayer, G. (2009b). Data management in Systems biology II - Outlook towards the semantic web. *CoRR*, abs/0912.2822. [URL:arXiv:0912.2822]. 13

McCudden, C. R., Hains, M. D., Kimple, R. J., Siderovski, D. P., and Willard, F. S. (2005). G-protein signaling: back to the future. *Cell. Mol. Life Sci.*, 62(5):551–577. [PubMed Central:PMC2794341] [DOI:10.1007/s00018-004-4462-3] [PubMed:15747061]. 33

McDonagh, E. M., Whirl-Carrillo, M., Garten, Y., Altman, R. B., and Klein, T. E. (2011). From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark Med*, 5(6):795–806. 137

Medidata (2013). Capturing the Value of EDC. https://www.mdsol.com/sites/default/files/RAVE_Capturing-Value-EDC_20131130_Medidata_White-Paper.pdf. [Online; accessed 01-November-2017]. 22

Meiser, J., Weindl, D., and Hiller, K. (2013). Complexity of dopamine metabolism. *Cell Commun. Signal*, 11(1):34. [PubMed Central:PMC3693914] [DOI:10.1186/1478-811X-11-34] [PubMed:23683503]. 168

Melone, M. A., Jori, F. P., and Peluso, G. (2005). Huntington's disease: new frontiers for molecular and cell therapy. *Curr Drug Targets*, 6(1):43–56. [PubMed:15720212]. 116

Merelli, I., Prez-Snchez, H., Gesing, S., and D?Agostino, D. (2014). Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *BioMed Research International*, 2014:1?13. 25

Merideth, M. A., Gordon, L. B., Clauss, S., Sachdev, V., Smith, A. C., Perry, M. B., Brewer, C. C., Zalewski, C., Kim, H. J., Solomon, B., Brooks, B. P., Gerber, L. H., Turner, M. L., Domingo, D. L., Hart, T. C., Graf, J., Reynolds, J. C., Gropman, A., Yanovski, J. A., Gerhard-Herman, M., Collins, F. S., Nabel, E. G., Cannon, R. O., Gahl, W. A., and Introne, W. J. (2008). Phenotype and course of Hutchinson-Gilford progeria syndrome. *N. Engl. J. Med.*, 358(6):592–604. [PubMed Central:PMC2940940] [DOI:10.1056/NEJMoa0706898] [PubMed:18256394]. 114

MetaCore (2016). MetaCore and Key Pathway Advisor Data-mining and pathway analysis. http://ipscience.thomsonreuters.com/product/metacore. [Online; accessed 04-November-2017]. 30

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H., and

Thomas, P. D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, 33(Database issue):D284–288. [PubMed Central:PMC540032] [DOI:10.1093/nar/gki078] [PubMed:15608197]. 4, 42, 64, 121, 137

Miller, J. P., Holcomb, J., Al-Ramahi, I., de Haro, M., Gafni, J., Zhang, N., Kim, E., Sanhueza, M., Torcassi, C., Kwak, S., Botas, J., Hughes, R. E., and Ellerby, L. M. (2010). Matrix metalloproteinases are modifiers of huntingtin proteolysis and toxicity in Huntington's disease. *Neuron*, 67(2):199–212. [PubMed Central:PMC3098887] [DOI:10.1016/j.neuron.2010.06.021] [PubMed:20670829]. 126, 127

Miller, J. P., Yates, B. E., Al-Ramahi, I., Berman, A. E., Sanhueza, M., Kim, E., de Haro, M., DeGiacomo, F., Torcassi, C., Holcomb, J., Gafni, J., Mooney, S. D., Botas, J., Ellerby, L. M., and Hughes, R. E. (2012). A genome-scale RNA-interference screen identifies RRAS signaling as a pathologic feature of Huntington's disease. *PLoS Genet.*, 8(11):e1003042. [PubMed Central:PMC3510027] [DOI:10.1371/journal.pgen.1003042] [PubMed:23209424]. 126

Mitchell, R. S., Katsura, C., Skasko, M. A., Fitzpatrick, K., Lau, D., Ruiz, A., Stephens, E. B., Margottin-Goguet, F., Benarous, R., and Guatelli, J. C. (2009). Vpu antagonizes BST-2-mediated restriction of HIV-1 release via beta-TrCP and endo-lysosomal trafficking. *PLoS Pathog.*, 5(5):e1000450. [PubMed Central:PMC2679223] [DOI:10.1371/journal.ppat.1000450] [PubMed:19478868]. 156

Miyoshi, N. S., Pinheiro, D. G., Silva, W. A., and Felipe, J. C. (2013). Computational framework to support integration of biomolecular and clinical data within a translational approach. *BMC Bioinformatics*, 14:180. [PubMed Central:PMC3688149] [DOI:10.1186/1471-2105-14-180] [PubMed:23742129]. 26

Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., Miyamoto, T., Miyashita, A., Kuwano, R., and Tanaka, H. (2012). AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Syst Biol*, 6:52. [PubMed Central:PMC3411424] [DOI:10.1186/1752-0509-6-52] [PubMed:22647208]. 31

Morton, A. J., Hunt, M. J., Hodges, A. K., Lewis, P. D., Redfern, A. J., Dunnett, S. B., and Jones, L. (2005). A combination drug therapy improves cognition and reverses gene expression changes in a mouse model of Huntington's disease. *Eur. J. Neurosci.*, 21(4):855–870. [DOI:10.1111/j.1460-9568.2005.03895.x] [PubMed:15787692]. 123

MRC (2013). UK Medical Research Council (MRC) policy on sharing of research data from population and patient studies. https://www.mrc.ac.uk/publications/browse/mrc-policy-and-guidance-on-sharing-of-research-data-from-population-and-patie [Online; accessed 31-October-2017]. 18

Murphy, S., Churchill, S., Bry, L., Chueh, H., Weiss, S., Lazarus, R., Zeng, Q., Dubey, A., Gainer, V., Mendis, M., Glaser, J., and Kohane, I. (2009). Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.*, 19(9):1675–1681. [PubMed Central:PMC2752136] [DOI:10.1101/gr.094615.109] [PubMed:19602638]. 12, 27, 160

Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., and Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*, 17(2):124–130. [PubMed Central:PMC3000779] [DOI:10.1136/jamia.2009.000893] [PubMed:20190053]. 12, 27, 160

Myers, R. H. (2004). Huntington's disease genetics. *NeuroRx*, 1(2):255–262. [PubMed Central:PMC534940] [DOI:10.1602/neurorx.1.2.255] [PubMed:15717026]. 116

Nagasaki, M., Saito, A., Li, C., Jeong, E., and Miyano, S. (2008). Systematic reconstruction of TRANSPATH data into cell system markup language. *BMC Syst Biol*, 2:53. [PubMed Central:PMC2474843] [DOI:10.1186/1752-0509-2-53] [PubMed:18570683]. 5

Nahm, M., Shepherd, J., Buzenberg, A., Rostami, R., Corcoran, A., McCall, J., and Pietrobon, R. (2011). Design and implementation of an institutional case

report form library. *Clin Trials*, 8(1):94–102. [PubMed Central:PMC3494996] [DOI:10.1177/1740774510391916] [PubMed:21163853]. 20

Natter, M. D., Quan, J., Ortiz, D. M., Bousvaros, A., Ilowite, N. T., Inman, C. J., Marsolo, K., McMurry, A. J., Sandborg, C. I., Schanberg, L. E., Wallace, C. A., Warren, R. W., Weber, G. M., and Mandl, K. D. (2013). An i2b2-based, generalizable, open source, self-scaling chronic disease registry. *J Am Med Inform Assoc*, 20(1):172–179. [PubMed Central:PMC3555330] [DOI:10.1136/amiajnl-2012-001042] [PubMed:22733975]. 26

Nelson, E. K., Piehler, B., Eckels, J., Rauch, A., Bellew, M., Hussey, P., Ramsay, S., Nathe, C., Lum, K., Krouse, K., Stearns, D., Connolly, B., Skillman, T., and Igra, M. (2011). LabKey Server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics*, 12:71. [PubMed Central:PMC3062597] [DOI:10.1186/1471-2105-12-71] [PubMed:21385461]. 12

Neron, B., Menager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P., and Letondal, C. (2009). Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–3011. [PubMed Central:PMC2773253] [DOI:10.1093/bioinformatics/btp493] [PubMed:19689959]. 28

Neumann, B., Walter, T., Heriche, J. K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., Cetin, C., Sieckmann, F., Pau, G., Kabbe, R., Wunsche, A., Satagopam, V., Schmitz, M. H., Chapuis, C., Gerlich, D. W., Schneider, R., Eils, R., Huber, W., Peters, J. M., Hyman, A. A., Durbin, R., Pepperkok, R., and Ellenberg, J. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727. [PubMed Central:PMC3108885] [DOI:10.1038/nature08869] [PubMed:20360735]. 82

Neves, M. and Leser, U. (2014). A survey on annotation tools for the biomedical literature. *Brief. Bioinformatics*, 15(2):327–340. [DOI:10.1093/bib/bbs084] [PubMed:23255168]. 30

Nguyen, L., Shah, A., Harker, M., Martins, H., McCready, M., Menezes, A., Jacobs, D. O., and Pietrobon, R. (2006). DADOS-Prospective: an open

source application for Web-based prospective data collection. *Source Code Biol Med*, 1:7. [PubMed Central:PMC1679801] [DOI:10.1186/1751-0473-1-7] [PubMed:17147787]. 23, 25

Nguyen, T. D., Raniga, P., Barnes, D. G., and Egan, G. F. (2015). Design, implementation and operation of a multimodality research imaging informatics repository. *Health Inf Sci Syst*, 3(Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con):S6. [PubMed Central:PMC4383058] [DOI:10.1186/2047-2501-3-S1-S6] [PubMed:25870760]. 26

NIH (2014). Medical Subject Headings. `https://www.nlm.nih.gov/mesh/MBrowser.html`. [Online; accessed 26-August-2015]. 91

NIH (2016). National Institutes of Health (NIH) - Open access policy and funding information. `https://www.nature.com/openresearch/funding/nih-open-access-policy-funding/`. [Online; accessed 07-November-2017]. 3

O'Donoghue, S. I., Horn, H., Pafilis, E., Haag, S., Kuhn, M., Satagopam, V. P., Schneider, R., and Jensen, L. J. (2010). Reflect: A practical approach to web semantics. *JOURNAL OF WEB SEMANTICS*, 8(2-3):182–189. [DOI:j.websem.2010.03.003]. 40, 44, 96, 174

Ohno-Machado, L., Bafna, V., Boxwala, A. A., Chapman, B. E., Chapman, W. W., Chaudhuri, K., Day, M. E., Farcas, C., Heintzman, N. D., Jiang, X., Kim, H., Kim, J., Matheny, M. E., Resnic, F. S., Vinterbo, S. A., Armstrong, W., Balac, N., Burns, J., Chen, J., Chisholm, R., Cope, R., Dasgupta, S., Dwork, C., El-Kareh, R., Fitzhenry, F., Gamst, A., Gentili, A., Good, P., Gupta, A., Inoue, M., Joyce, R., Krueger, I., Kuo, G., Larkin, J., Messer, K., Nookala, L., Norman, G., Norris, K., Patel, K., Paul, P., Pevzner, P., Patrick, K., Pond, S., Que, J., Rathbun, S., Robbins, S., Sarwate, A., Shimizu, C., Sofia, H., Tarczy-Hornoch, P., Thornton, D., Vaida, F., Valafar, F., Varghese, G., Wolter, N., Wong, C., Wong, M., and Zambon, A. (2012). iDASH: integrating data for analysis, anonymization, and sharing. *J Am Med Inform Assoc*, 19(2):196–201. [PubMed Central:PMC3277627] [DOI:10.1136/amiajnl-2011-000538] [PubMed:22081224]. 26

Oldham, W. M. and Hamm, H. E. (2008). Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat. Rev. Mol. Cell Biol.*, 9(1):60–71. [DOI:10.1038/nrm2299] [PubMed:18043707]. 33

OpenClinica (2017). OpenClinica: Open source Electronic Data Capturing and data management solution. https://www.openclinica.com/community-edition-open-source-edc/. [Online; accessed 01-November-2017]. 12, 23, 24

Oracle (2017a). Oracle Health Sciences InForm. http://www.oracle.com/us/products/applications/health-sciences/inform-medication-adherence/medication-adherence-overview-2319774.html. [Online; accessed 01-November-2017]. 23

Oracle (2017b). Oracle Clinical. http://www.oracle.com/us/products/applications/health-sciences/e-clinical/clinical/index.html. [Online; accessed 01-November-2017]. 23

Orechia, J., Pathak, A., Shi, Y., Nawani, A., Belozerov, A., Fontes, C., Lakhiani, C., Jawale, C., Patel, C., Quinn, D., Botvinnik, D., Mei, E., Cotter, E., Byleckie, J., Ullman-Cullere, M., Chhetri, P., Chalasani, P., Karnam, P., Beaudoin, R., Sahu, S., Belozerova, Y., and Mathew, J. P. (2015). Oncdrs: An integrative clinical and genomic data platform for enabling translational research and precision medicine. *Applied & Translational Genomics*, 6:18?25. 26

Oster, S., Langella, S., Hastings, S., Ervin, D., Madduri, R., Kurc, T., Siebenlist, F., Covitz, P., Shanbhag, K., Foster, I., and Saltz, J. (2007). caGrid 1.0: a Grid enterprise architecture for cancer research. *AMIA Annu Symp Proc*, pages 573–577. [PubMed Central:PMC2655925] [PubMed:18693901]. 26

ownCloud (2017). A suite of client?server software for creating file hosting services and using them. https://owncloud.org. [Online; accessed 09-November-2017]. 10

Ozaki, T., Saijo, M., Murakami, K., Enomoto, H., Taya, Y., and Sakiyama, S. (1994). Complex formation between lamin A and the retinoblastoma gene product: identification of the domain on lamin A required for its interaction. *Oncogene*, 9(9):2649–2653. [PubMed:8058329]. 108, 176

Pace, L. E. and Keating, N. L. (2014). A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA*, 311(13):1327–1335. [DOI:10.1001/jama.2014.1398] [PubMed:24691608].

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Arvanitidis, C., and Jensen, L. J. (2013). The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS ONE*, 8(6):e65390. [PubMed Central:PMC3688812] [DOI:10.1371/journal.pone.0065390] [PubMed:23823062]. 14

Pafilis, E., O'Donoghue, S. I., Jensen, L. J., Horn, H., Kuhn, M., Brown, N. P., and Schneider, R. (2009). Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, 27(6):508–510. [DOI:10.1038/nbt0609-508] [PubMed:19513049]. 15, 44, 93, 147

Park, W. Y., Hwang, C. I., Kang, M. J., Seo, J. Y., Chung, J. H., Kim, Y. S., Lee, J. H., Kim, H., Kim, K. A., Yoo, H. J., and Seo, J. S. (2001). Gene profile of replicative senescence is different from progeria or elderly donor. *Biochem. Biophys. Res. Commun.*, 282(4):934–939. [DOI:10.1006/bbrc.2001.4632] [PubMed:11352641]. 103

Pathway-Studio (2016). Pathway Studio. Experimental data and disease models at the heart of biological research. `www.elsevier.com/solutions/pathway-studio-biological-research`. [Online; accessed 04-November-2017]. 30

Paul, J., Seib, R., and Prescott, T. (2005). The Internet and clinical trials: background, online resources, examples and issues. *J. Med. Internet Res.*, 7(1):e5. [PubMed Central:PMC1550630] [DOI:10.2196/jmir.7.1.e5] [PubMed:15829477]. 22

Pavlidis, P. and Noble, W. S. (2003). Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, 19(2):295–296. [PubMed:12538257]. 74, 89

Pavlopoulos, G. A., O'Donoghue, S. I., Satagopam, V. P., Soldatos, T. G., Pafilis, E., and Schneider, R. (2008). Arena3D: visualization of biological networks in 3D. *BMC Syst Biol*, 2:104. [PubMed Central:PMC2637860] [DOI:10.1186/1752-0509-2-104] [PubMed:19040715]. 79, 131

Payne, P., Ervin, D., Dhaval, R., Borlawsky, T., and Lai, A. (2011). TRIAD: The Translational Research Informatics and Data Management Grid. *Appl Clin Inform*, 2(3):331–344. [PubMed Central:PMC3631927] [DOI:10.4338/ACI-2011-02-RA-0014] [PubMed:23616879]. 26

Peng, Y., Gupta, S., Wu, C., and Vijay-Shanker, K. (2015). An extended dependency graph for relation extraction in biomedical texts. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)*, pages 21–30. 14

Pinsky, P. F., Prorok, P. C., and Kramer, B. S. (2017). Prostate Cancer Screening - A Perspective on the Current State of the Evidence. *N. Engl. J. Med.*, 376(13):1285–1289. [DOI:10.1056/NEJMsb1616281] [PubMed:28355509].

Pizzuti, C. and Rombo, S. E. (2014). Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352. [DOI:10.1093/bioinformatics/btu034] [PubMed:24458952]. 30

Poletti, M. and Bonuccelli, U. (2013). Acute and chronic cognitive effects of levodopa and dopamine agonists on patients with Parkinson's disease: a review. *Ther Adv Psychopharmacol*, 3(2):101–113. [PubMed Central:PMC3805397] [DOI:10.1177/2045125312470130] [PubMed:24167681]. 168

Pollex, R. L. and Hegele, R. A. (2004). Hutchinson-Gilford progeria syndrome. *Clin. Genet.*, 66(5):375–381. [DOI:10.1111/j.1399-0004.2004.00315.x] [PubMed:15479179]. 100, 101

Prokscha, S. (2011). *Practical Guide to Clinical Data Management. , Second Edition*. CRC Press. 23

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, 40(Database issue):D130–135. [PubMed Central:PMC3245008] [DOI:10.1093/nar/gkr1079] [PubMed:22121212]. 3, 42, 137

PubMed (2015). PubMed. `http://www.ncbi.nlm.nih.gov/pubmed`. [Online; accessed 14-August-2015]. 4, 43

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Res.*, 40(Database issue):290–301. [PubMed Central:PMC3245129] [DOI:10.1093/nar/gkr1065] [PubMed:22127870]. 3, 42, 121, 137

Rayner, T. F., Rocca-Serra, P., Spellman, P. T., Causton, H. C., Farne, A., Holloway, E., Irizarry, R. A., Liu, J., Maier, D. S., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert, C. J., White, J., Whetzel, P. L., Wymore, F., Parkinson, H., Sarkans, U., Ball, C. A., and Brazma, A. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7:489. [PubMed Central:PMC1687205] [DOI:10.1186/1471-2105-7-489] [PubMed:17087822]. 10

Rebholz-Schuhmann, D., Jimeno Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J. B., Baker, C. J., Kuo, C. J., Clematide, S., Rinaldi, F., Farkas, R., Mora, G., Hara, K., Furlong, L. I., Rautschka, M., Neves, M. L., Pascual-Montano, A., Wei, Q., Collier, N., Chowdhury, M. F., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J. L., van Mulligen, E., Kors, J., and Hahn, U. (2011). Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J Biomed*

Semantics, 2 Suppl 5:S11. [PubMed Central:PMC3239301] [DOI:10.1186/2041-1480-2-S5-S11] [PubMed:22166494]. 14

REDCap (2017). REDCap: Research Electronic Data Capture. https://projectredcap.org. [Online; accessed 01-November-2017]. 12, 23, 24

Regan, K. and Payne, P. R. (2015). From Molecules to Patients: The Clinical Applications of Translational Bioinformatics. *Yearb Med Inform*, 10(1):164–169. [PubMed Central:PMC4587059] [DOI:10.15265/IY-2015-005] [PubMed:26293863]. 17, 159, 184

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, 31(1):298–303. [PubMed Central:PMC165547] [PubMed:12520007]. 15, 146, 155, 182

RISC (2016). REDCap Getting Started: Compliance. https://rc.partners.org/kb/article/2732. [Online; accessed 01-November-2017]. 22

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47. [PubMed Central:PMC4402510] [DOI:10.1093/nar/gkv007] [PubMed:25605792]. 165

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., and Sansone, S. A. (2010). ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356. [PubMed Central:PMC2935443] [DOI:10.1093/bioinformatics/btq415] [PubMed:20679334]. 10, 12

Roomp, K., Kristinsson, H., Schvartz, D., Ubhayasekera, K., Sargsyan, E., Manukyan, L., Chowdhury, A., Manell, H., Satagopam, V., Groebe, K., Schneider, R., Bergquist, J., Sanchez, J. C., and Bergsten, P. (2017). Combined lipidomic and proteomic analysis of isolated human islets exposed to palmitate reveals time-dependent changes in insulin secretion and lipid

metabolism. *PLoS ONE*, 12(4):e0176391. [PubMed Central:PMC5407795] [DOI:10.1371/journal.pone.0176391] [PubMed:28448538]. 82, 176

Rosario, B. and Hearst, M. (2004). Classifying semantic relations in bioscience texts. *InProceeding ACL ?04 Proceedings of the 42nd Annual Meeting on Association forComputational Linguistics*. 14

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Meth. Enzymol.*, 266:525–539. [PubMed:8743704]. 14

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003). Automatic prediction of protein function. *Cell. Mol. Life Sci.*, 60(12):2637–2650. [DOI:10.1007/s00018-003-3114-8] [PubMed:14685688]. 14

Runne, H., Regulier, E., Kuhn, A., Zala, D., Gokce, O., Perrin, V., Sick, B., Aebischer, P., Deglon, N., and Luthi-Carter, R. (2008). Dysregulation of gene expression in primary neuron models of Huntington's disease shows that polyglutamine-related effects on the striatal transcriptome may not be dependent on brain circuitry. *J. Neurosci.*, 28(39):9723–9731. [DOI:10.1523/JNEUROSCI.3044-08.2008] [PubMed:18815258]. 123

Russell, J. and McHale, P. (2010). Taking the Pain out of EDC Deployment. `https://www.parexel.com/solutions/informatics/electronic-data-capture/datalabs-edc/taking-pain-out-edc-deployment`. [Online; accessed 01-November-2017]. 22

Russom, P. (2011). The three Vs of big data analytics. *TDWI Best Practices Report, Fourth Quarter*, 18:1–35. 6

Rzhetsky, A., Seringhaus, M., and Gerstein, M. B. (2009). Getting started in text mining: part two. *PLoS Comput. Biol.*, 5(7):e1000411. [PubMed Central:PMC2709911] [DOI:10.1371/journal.pcbi.1000411] [PubMed:19649304]. 14

Sachse, C., Krausz, E., Kronke, A., Hannus, M., Walsh, A., Grabner, A., Ovcharenko, D., Dorris, D., Trudel, C., Sonnichsen, B., and Echeverri, C. J. (2005). High-throughput RNA interference strategies for target discovery

and validation by using synthetic short interfering RNAs: functional genomics investigations of biological pathways. *Meth. Enzymol.*, 392:242–277. [DOI:10.1016/S0076-6879(04)92015-0] [PubMed:15644186]. 117

Sadri-Vakili, G., Bouzou, B., Benn, C. L., Kim, M. O., Chawla, P., Overland, R. P., Glajch, K. E., Xia, E., Qiu, Z., Hersch, S. M., Clark, T. W., Yohrling, G. J., and Cha, J. H. (2007). Histones associated with downregulated genes are hypo-acetylated in Huntington's disease models. *Hum. Mol. Genet.*, 16(11):1293–1306. [DOI:10.1093/hmg/ddm078] [PubMed:17409194]. 123

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford)*, 2010:baq020. [PubMed Central:PMC2938269] [DOI:10.1093/database/baq020] [PubMed:20689021]. 3, 42, 137

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523. [DOI:10.1016/0306-4573(88)90021-0]. 61

Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L. A., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C. T., Forster, M. J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J. L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Ho Sui, S. J., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C. E., Shang, C. A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., and Hide, W. (2012). Toward interoperable bioscience data. *Nat. Genet.*, 44(2):121–126. [PubMed Central:PMC3428019] [DOI:10.1038/ng.1054] [PubMed:22281772]. 10

Sasaki, Y., Tsuruoka, Y., McNaught, J., and Ananiadou, S. (2008). How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9

Suppl 11:S5. [PubMed Central:PMC2586754] [DOI:10.1186/1471-2105-9-S11-S5] [PubMed:19025691]. 14

Satagopam, V., Gu, W., Eifes, S., Gawron, P., Ostaszewski, M., Gebel, S., Barbosa-Silva, A., Balling, R., and Schneider, R. (2016). Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases. *Big Data*, 4(2):97–108. [PubMed Central:PMC4932659] [DOI:10.1089/big.2015.0057] [PubMed:27441714]. 25, 26, 27, 28, 29, 30, 31, 160, 168, 169, 171

Satagopam, V. and Schneider, R. (2016). PD map connector. `http://r3lab.uni.lu/web/tgm-pipeline/code/PD%20map%20connector.pl`. [Online; accessed 24-August-2017]. 167

Satagopam, V. P., Theodoropoulou, M. C., Stampolakis, C. K., Pavlopoulos, G. A., Papandreou, N. C., Bagos, P. G., Schneider, R., and Hamodrakas, S. J. (2010). GPCRs, G-proteins, effectors and their interactions: human-gpDB, a database employing visualization tools and data integration techniques. *Database (Oxford)*, 2010:baq019. [PubMed Central:PMC2931634] [DOI:10.1093/database/baq019] [PubMed:20689020]. 32, 33, 131, 132, 134

Sauro, H. M., Hucka, M., Finney, A., Wellock, C., Bolouri, H., Doyle, J., and Kitano, H. (2003). Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS*, 7(4):355–372. [DOI:10.1089/153623103322637670] [PubMed:14683609]. 12

Sawyer, S. A., Parsch, J., Zhang, Z., and Hartl, D. L. (2007). Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, 104(16):6504–6510. [PubMed Central:PMC1871816] [DOI:10.1073/pnas.0701572104] [PubMed:17409186]. 14

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A.,

Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 40(Database issue):13–25. [PubMed Central:PMC3245031] [DOI:10.1093/nar/gkr1184] [PubMed:22140104]. 3, 42

Scaffidi, P. and Misteli, T. (2008). Lamin A-dependent misregulation of adult stem cells associated with accelerated ageing. *Nat. Cell Biol.*, 10(4):452–459. [PubMed Central:PMC2396576] [DOI:10.1038/ncb1708] [PubMed:18311132]. 103, 108, 109, 110, 113, 114

Schafferhans, A., Meyer, J. E., and O'Donoghue, S. I. (2003). The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Res.*, 31(1):494–498. [PubMed Central:PMC165557] [PubMed:12520061]. 3, 42

Scheufele, E., Aronzon, D., Coopersmith, R., McDuffie, M. T., Kapoor, M., Uhrich, C. A., Avitabile, J. E., Liu, J., Housman, D., and Palchuk, M. B. (2014). tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. *AMIA Jt Summits Transl Sci Proc*, 2014:96–101. [PubMed Central:PMC4333702] [PubMed:25717408]. 27, 161

Schneider, R. and Sander, C. (1996). The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, 24(1):201–205. [PubMed Central:PMC145595] [PubMed:8594579]. 3, 42

Schrimpe-Rutledge, A. C., Fontes, G., Gritsenko, M. A., Norbeck, A. D., Anderson, D. J., Waters, K. M., Adkins, J. N., Smith, R. D., Poitout, V., and Metz, T. O. (2012). Discovery of novel glucose-regulated proteins in isolated human pancreatic islets using LC-MS/MS-based proteomics. *J. Proteome Res.*, 11(7):3520–3532. [PubMed Central:PMC3391329] [DOI:10.1021/pr3002996] [PubMed:22578083]. 56

Schuler, G. D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, 75(10):694–698. [PubMed:9382993]. 3, 42, 137

Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W., and Bruford, E. A. (2011). genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, 39(Database issue):D514–519. [PubMed Central:PMC3013772] [DOI:10.1093/nar/gkq892] [PubMed:20929869]. 3, 42, 137

Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192. [DOI:10.1093/bioinformatics/bti475] [PubMed:15860559]. 14

Shah, A. R., Singhal, M., Klicker, K. R., Stephan, E. G., Wiley, H. S., and Waters, K. M. (2007). Enabling high-throughput data management for systems biology: the Bioinformatics Resource Manager. *Bioinformatics*, 23(7):906–909. [DOI:10.1093/bioinformatics/btm031] [PubMed:17324940]. 12

Shah, J., Rajgor, D., Pradhan, S., McCready, M., Zaveri, A., and Pietrobon, R. (2010). Electronic data capture for registries and clinical trials in orthopaedic surgery: open source versus commercial systems. *Clin. Orthop. Relat. Res.*, 468(10):2664–2671. [PubMed Central:PMC3049639] [DOI:10.1007/s11999-010-1469-3] [PubMed:20635174]. 12, 21, 24, 25

Shah, N. H., Jonquet, C., Chiang, A. P., Butte, A. J., Chen, R., and Musen, M. A. (2009). Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*, 10 Suppl 2:S1. [PubMed Central:PMC2646250] [DOI:10.1186/1471-2105-10-S2-S1] [PubMed:19208184]. 14

Shannon, P. T., Reiss, D. J., Bonneau, R., and Baliga, N. S. (2006). The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, 7:176. [PubMed Central:PMC1464137] [DOI:10.1186/1471-2105-7-176] [PubMed:16569235]. 12

Shneiderman, B., Plaisant, C., and Hesse, B. (2013). Improving healthcare with interactive visualization. *Computer*, 46:58–66. 161

Shteynberg, D., Slagel, J., Mendoza, L., Hoopmann, M., Eng, J., Lam, H., Nesvizhskii, A., and Pratt, B. (2015). Trans-Proteomic Pipeline. `http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP`. [Online; accessed 25-August-2015]. 88

Simonetti, M., Hagenston, A. M., Vardeh, D., Freitag, H. E., Mauceri, D., Lu, J., Satagopam, V. P., Schneider, R., Costigan, M., Bading, H., and Kuner, R. (2013). Nuclear calcium signaling in spinal neurons drives a genomic program required for persistent inflammatory pain. *Neuron*, 77(1):43–57. [PubMed Central:PMC3593630] [DOI:10.1016/j.neuron.2012.10.037] [PubMed:23312515]. 82

Singh, M., Bhartiya, D., Maini, J., Sharma, M., Singh, A. R., Kadarkaraisamy, S., Rana, R., Sabharwal, A., Nanda, S., Ramachandran, A., Mittal, A., Kapoor, S., Sehgal, P., Asad, Z., Kaushik, K., Vellarikkal, S. K., Jagga, D., Muthuswami, M., Chauhan, R. K., Leonard, E., Priyadarshini, R., Halimani, M., Malhotra, S., Patowary, A., Vishwakarma, H., Joshi, P., Bhardwaj, V., Bhaumik, A., Bhatt, B., Jha, A., Kumar, A., Budakoti, P., Lalwani, M. K., Meli, R., Jalali, S., Joshi, K., Pal, K., Dhiman, H., Laddha, S. V., Jadhav, V., Singh, N., Pandey, V., Sachidanandan, C., Ekker, S. C., Klee, E. W., Scaria, V., and Sivasubbu, S. (2014). The Zebrafish GenomeWiki: a crowdsourcing approach to connect the long tail for zebrafish gene annotation. *Database (Oxford)*, 2014:bau011. [PubMed Central:PMC3936183] [DOI:10.1093/database/bau011] [PubMed:24578356]. 10

Siwach, P. and Ganesh, S. (2008). Tandem repeats in human disorders: mechanisms and evolution. *Front. Biosci.*, 13:4467–4484. [PubMed:18508523]. 32

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3. [DOI:10.2202/1544-6115.1027] [PubMed:16646809]. 53

Sonnichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A. M., Artelt, J., Bettencourt, P., Cassin, E., Hewitson, M., Holz, C., Khan, M., Lazik, S., Martin, C., Nitzsche, B., Ruer, M., Stamford, J., Winzi, M., Heinkel, R., Roder, M., Finell, J., Hantsch, H., Jones, S. J., Jones, M., Piano, F., Gunsalus, K. C., Oegema, K., Gonczy, P., Coulson, A., Hyman, A. A., and Echeverri, C. J. (2005). Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature*, 434(7032):462–469. [DOI:10.1038/nature03353] [PubMed:15791247]. 117

Spasi, I., Schober, D., Sansone, S. A., Rebholz-Schuhmann, D., Kell, D. B., and Paton, N. W. (2008). Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*, 9 Suppl 5:S5. [PubMed Central:PMC2367623] [DOI:10.1186/1471-2105-9-S5-S5] [PubMed:18460187]. 14

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biol.*, 13(7):e1002195. [PubMed Central:PMC4494865] [DOI:10.1371/journal.pbio.1002195] [PubMed:26151137]. 8

Stonebraker, M., Beskales, G., Pagan, A., Bruckner, D., Cherniack, M., Xu, S., Ilyas, I., and Zdonik, S. (2013). Data curation at scale: The Data Tamer System. *In: Proceedings of the 6th Biennial Conference on Innovative Data Systems Research. Asilomar, CA*. [Online; accessed 01-November-2017 PDF]. 25

Strand, A. D., Aragaki, A. K., Baquet, Z. C., Hodges, A., Cunningham, P., Holmans, P., Jones, K. R., Jones, L., Kooperberg, C., and Olson, J. M. (2007a). Conservation of regional gene expression in mouse and human brain. *PLoS Genet.*, 3(4):e59. [PubMed Central:PMC1853119] [DOI:10.1371/journal.pgen.0030059] [PubMed:17447843]. 123

Strand, A. D., Baquet, Z. C., Aragaki, A. K., Holmans, P., Yang, L., Cleren, C., Beal, M. F., Jones, L., Kooperberg, C., Olson, J. M., and Jones, K. R. (2007b). Expression profiling of Huntington's disease models suggests that brain-derived neurotrophic factor depletion plays a major role in striatal degeneration. *J. Neurosci.*, 27(43):11758–11768. [DOI:10.1523/JNEUROSCI.2461-07.2007] [PubMed:17959817]. 123

Sugars, K. L. and Rubinsztein, D. C. (2003). Transcriptional abnormalities in Huntington disease. *Trends Genet.*, 19(5):233–238. [DOI:10.1016/S0168-9525(03)00074-X] [PubMed:12711212]. 116

Sulè, A. and Lapeyra, L. (2016). Introduction to the Semantic web and Linked

data . http://dlis.hypotheses.org/788. [Online; accessed 10-November-2017]. 13

Systec (2010). Systec:FANCI consortium. http://systec.embl.de/index.php?option=com_content&view=article&id=46&Itemid=55. [Online; accessed 22-August-2015]. 82

Szalma, S., Koka, V., Khasanova, T., and Perakslis, E. D. (2010). Effective knowledge management in translational medicine. *J Transl Med*, 8:68. [PubMed Central:PMC2914663] [DOI:10.1186/1479-5876-8-68] [PubMed:20642836]. 12, 26, 27, 160, 161

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39(Database issue):D561–568. [PubMed Central:PMC3013807] [DOI:10.1093/nar/gkq973] [PubMed:21045058]. 4, 14, 43, 121

Takahashi, T., Kikuchi, S., Katada, S., Nagai, Y., Nishizawa, M., and Onodera, O. (2008). Soluble polyglutamine oligomers formed prior to inclusion body formation are cytotoxic. *Hum. Mol. Genet.*, 17(3):345–356. [DOI:10.1093/hmg/ddm311] [PubMed:17947294]. 127

Tan, A., Tripp, B., and Daley, D. (2011). BRISK–research-oriented storage kit for biology-related data. *Bioinformatics*, 27(17):2422–2425. [DOI:10.1093/bioinformatics/btr389] [PubMed:21712248]. 26

TCGA (2017). TCGA by numbers, program overview. https://cancergenome.nih.gov/abouttcga/overview. [Online; accessed 17-October-2017]. 8

Team, R. D. C. (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 72

Theodoropoulou, M. C., Bagos, P. G., Spyropoulos, I. C., and Hamodrakas, S. J. (2008). gpDB: a database of GPCRs, G-proteins, effectors and their interac-

tions. *Bioinformatics*, 24(12):1471–1472. [DOI:10.1093/bioinformatics/btn206] [PubMed:18441001]. 133, 179

Thomas, E. A., Coppola, G., Desplats, P. A., Tang, B., Soragni, E., Burnett, R., Gao, F., Fitzgerald, K. M., Borok, J. F., Herman, D., Geschwind, D. H., and Gottesfeld, J. M. (2008). The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice. *Proc. Natl. Acad. Sci. U.S.A.*, 105(40):15564–15569. [PubMed Central:PMC2563081] [DOI:10.1073/pnas.0804249105] [PubMed:18829438]. 123

Tian, Q., Price, N. D., and Hood, L. (2012). Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J. Intern. Med.*, 271(2):111–121. [PubMed Central:PMC3978383] [DOI:10.1111/j.1365-2796.2011.02498.x] [PubMed:22142401]. 18, 159, 184

Toga, A. W., Foster, I., Kesselman, C., Madduri, R., Chard, K., Deutsch, E. W., Price, N. D., Glusman, G., Heavner, B. D., Dinov, I. D., Ames, J., Van Horn, J., Kramer, R., and Hood, L. (2015). Big biomedical data as the key resource for discovery science. *J Am Med Inform Assoc*, 22(6):1126–1131. [PubMed Central:PMC5009918] [DOI:10.1093/jamia/ocv077] [PubMed:26198305]. 26

Tomlinson, C., Thimma, M., Alexandrakis, S., Castillo, T., Dennis, J. L., Brooks, A., Bradley, T., Turnbull, C., Blaveri, E., Barton, G., Chiba, N., Maratou, K., Soutter, P., Aitman, T., and Game, L. (2008). MiMiR–an integrated platform for microarray data sharing, mining and analysis. *BMC Bioinformatics*, 9:379. [PubMed Central:PMC2572073] [DOI:10.1186/1471-2105-9-379] [PubMed:18801157]. 12

Tomlinson, C. D., Barton, G. R., Woodbridge, M., and Butcher, S. A. (2013). XperimentR: painless annotation of a biological experiment for the laboratory scientist. *BMC Bioinformatics*, 14:8. [PubMed Central:PMC3571946] [DOI:10.1186/1471-2105-14-8] [PubMed:23323856]. 12

Topol, E. J. (2015). The big medical data miss: challenges in establishing an open medical resource. *Nat. Rev. Genet.*, 16(5):253–254. [PubMed:26065035]. 17, 159

Toth, J. I., Yang, S. H., Qiao, X., Beigneux, A. P., Gelb, M. H., Moulson, C. L., Miner, J. H., Young, S. G., and Fong, L. G. (2005). Blocking protein farnesyltransferase improves nuclear shape in fibroblasts from humans with progeroid syndromes. *Proc. Natl. Acad. Sci. U.S.A.*, 102(36):12873–12878. [PubMed Central:PMC1193538] [DOI:10.1073/pnas.0505767102] [PubMed:16129834]. 102, 112

Tradigo, G., Veneziano, C., Greco, S., and Veltri, P. (2014). An architecture for integrating genetic and clinical data. *Procedia Computer Science*, 29:1959?1969. 26

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science*, 347(6220):1260419. [DOI:10.1126/science.1260419] [PubMed:25613900]. 89

UniProt-Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 42(Database issue):D191–198. 15, 147, 155, 182

University of Oxford, D. o. C. S. (2014). Architecture to solve the genome data mountain. `https://www.cs.ox.ac.uk/news/845-full.html`. [Online; accessed 17-October-2017]. 8

USA-HHS (1996). The Health Insurance Portability and Accountability Act. `https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/HIPAA-ACA`. [Online; accessed 01-November-2017]. 23

USA-Homeland-Security (2014). Federal Information Security Modernization Act. `https://www.dhs.gov/fisma`. [Online; accessed 01-November-2017]. 23

Vallon-Christersson, J., Nordborg, N., Svensson, M., and Hakkinen, J. (2009). BASE–2nd generation software for microarray data management and analysis. *BMC Bioinformatics*, 10:330. [PubMed Central:PMC2768720] [DOI:10.1186/1471-2105-10-330] [PubMed:19822003]. 12

Van Deun, K., Smilde, A. K., van der Werf, M. J., Kiers, H. A., and Van Mechelen, I. (2009). A structured overview of simultaneous component based data integration. *BMC Bioinformatics*, 10:246. [PubMed Central:PMC2752463] [DOI:10.1186/1471-2105-10-246] [PubMed:19671149]. 6

Van Landeghem, S., Hakala, K., Ronnqvist, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations. *Adv Bioinformatics*, 2012:582765. [PubMed Central:PMC3375141] [DOI:10.1155/2012/582765] [PubMed:22719757]. 14

van Roon-Mom, W. M., Pepers, B. A., 't Hoen, P. A., Verwijmeren, C. A., den Dunnen, J. T., Dorsman, J. C., and van Ommen, G. B. (2008). Mutant huntingtin activates Nrf2-responsive genes and impairs dopamine synthesis in a PC12 model of Huntington's disease. *BMC Mol. Biol.*, 9:84. [PubMed Central:PMC2588454] [DOI:10.1186/1471-2199-9-84] [PubMed:18844975]. 123

Varela, I., Pereira, S., Ugalde, A. P., Navarro, C. L., Suarez, M. F., Cau, P., Cadinanos, J., Osorio, F. G., Foray, N., Cobo, J., de Carlos, F., Levy, N., Freije, J. M., and Lopez-Otin, C. (2008). Combined treatment with statins and aminobisphosphonates extends longevity in a mouse model of human premature aging. *Nat. Med.*, 14(7):767–772. [DOI:10.1038/nm1786] [PubMed:18587406]. 102

Vast, E. (2015). tranSMART XNAT importer. `https://github.com/evast/transmart-xnat-importer-plugin`. [Online; accessed 23-August-2017]. 27, 161

Velikova, G., Wright, E. P., Smith, A. B., Cull, A., Gould, A., Forman, D., Perren, T., Stead, M., Brown, J., and Selby, P. J. (1999). Automated collection

of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. *J. Clin. Oncol.*, 17(3):998–1007. [DOI:10.1200/JCO.1999.17.3.998] [PubMed:10071295]. 21, 22

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19(2):327–335. [PubMed Central:PMC2652215] [DOI:10.1101/gr.073585.107] [PubMed:19029536]. 43, 45, 63

Wade, T. D. (2014). Traits and types of health data repositories. *Health Inf Sci Syst*, 2:4. [PubMed Central:PMC4340801] [DOI:10.1186/2047-2501-2-4] [PubMed:25825668]. 25

Wang, S., Pandis, I., Wu, C., He, S., Johnson, D., Emam, I., Guitton, F., and Guo, Y. (2014). High dimensional biological data retrieval optimization with NoSQL technology. *BMC Genomics*, 15 Suppl 8:S3. [PubMed Central:PMC4248814] [DOI:10.1186/1471-2164-15-S8-S3] [PubMed:25435347]. 27, 161

Wang, Y., Panteleyev, A. A., Owens, D. M., Djabali, K., Stewart, C. L., and Worman, H. J. (2008). Epidermal expression of the truncated prelamin A causing Hutchinson-Gilford progeria syndrome: effects on keratinocytes, hair and skin. *Hum. Mol. Genet.*, 17(15):2357–2369. [PubMed Central:PMC2733813] [DOI:10.1093/hmg/ddn136] [PubMed:18442998]. 112

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *J. Comput. Aided Mol. Des.*, 26(7):801–804. [PubMed Central:PMC3414708] [DOI:10.1007/s10822-012-9577-7] [PubMed:22644661]. 29

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, S. N., Chu, A., Chuah, E., Chun, H. J., Dhalla, N., Guin, R., Hirst, M., Hirst, C., Holt, R. A., Jones, S. J., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Robertson,

A. G., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Varhol, R. J., Beroukhim, R., Bhatt, A. S., Brooks, A. N., Cherniack, A. D., Freeman, S. S., Gabriel, S. B., Helman, E., Jung, J., Meyerson, M., Ojesina, A. I., Pedamallu, C. S., Saksena, G., Schumacher, S. E., Tabak, B., Zack, T., Lander, E. S., Bristow, C. A., Hadjipanayis, A., Haseley, P., Kucherlapati, R., Lee, S., Lee, E., Luquette, L. J., Mahadeshwar, H. S., Pantazi, A., Parfenov, M., Park, P. J., Protopopov, A., Ren, X., Santoso, N., Seidman, J., Seth, S., Song, X., Tang, J., Xi, R., Xu, A. W., Yang, L., Zeng, D., Auman, J. T., Balu, S., Buda, E., Fan, C., Hoadley, K. A., Jones, C. D., Meng, S., Mieczkowski, P. A., Parker, J. S., Perou, C. M., Roach, J., Shi, Y., Silva, G. O., Tan, D., Veluvolu, U., Waring, S., Wilkerson, M. D., Wu, J., Zhao, W., Bodenheimer, T., Hayes, D. N., Hoyle, A. P., Jeffreys, S. R., Mose, L. E., Simons, J. V., Soloway, M. G., Baylin, S. B., Berman, B. P., Bootwalla, M. S., Danilova, L., Herman, J. G., Hinoue, T., Laird, P. W., Rhie, S. K., Shen, H., Triche, T., Weisenberger, D. J., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Sougnez, C., Wang, M., Saksena, G., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Dinh, H., Doddapaneni, H. V., Gibbs, R., Gunaratne, P., Han, Y., Kalra, D., Kovar, C., Lewis, L., Morgan, M., Morton, D., Muzny, D., Reid, J., Xi, L., Cho, J., DiCara, D., Frazer, S., Gehlenborg, N., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Liu, Y., Noble, M. S., Stojanov, P., Voet, D., Zhang, H., Zou, L., Stewart, C., Bernard, B., Bressler, R., Eakin, A., Iype, L., Knijnenburg, T., Kramer, R., Kreisberg, R., Leinonen, K., Lin, J., Liu, Y., Miller, M., Reynolds, S. M., Rovira, H., Shmulevich, I., Thorsson, V., Yang, D., Zhang, W., Amin, S., Wu, C. J., Wu, C. C., Akbani, R., Aldape, K., Baggerly, K. A., Broom, B., Casasent, T. D., Cleland, J., Creighton, C., Dodda, D., Edgerton, M., Han, L., Herbrich, S. M., Ju, Z., Kim, H., Lerner, S., Li, J., Liang, H., Liu, W., Lorenzi, P. L., Lu, Y., Melott, J., Mills, G. B., Nguyen, L., Su, X., Verhaak, R., Wang, W., Weinstein, J. N., Wong, A., Yang, Y., Yao, J., Yao, R., Yoshihara, K., Yuan, Y., Yung, A. K., Zhang, N., Zheng, S., Ryan, M., Kane, D. W., Aksoy, B. A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Kahles, A., Ladanyi, M., Lee, W., Lehmann, K. V., Miller, M. L., Ramirez, R., Ratsch, G., Reva, B., Sander, C., Schultz, N., Senbabaoglu, Y., Shen, R., Sinha, R., Sumer, S. O., Sun, Y., Taylor, B. S., Weinhold, N., Fei,

S., Spellman, P., Benz, C., Carlin, D., Cline, M., Craft, B., Ellrott, K., Goldman, M., Haussler, D., Ma, S., Ng, S., Paull, E., Radenbaugh, A., Salama, S., Sokolov, A., Stuart, J. M., Swatloski, T., Uzunangelov, V., Waltman, P., Yau, C., Zhu, J., Hamilton, S. R., Getz, G., Sougnez, C., Abbott, S., Abbott, R., Dees, N. D., Delehaunty, K., Ding, L., Dooling, D. J., Eldred, J. M., Fronick, C. C., Fulton, R., Fulton, L. L., Kalicki-Veizer, J., Kanchi, K. L., Kandoth, C., Koboldt, D. C., Larson, D. E., Ley, T. J., Lin, L., Lu, C., Magrini, V. J., Mardis, E. R., McLellan, M. D., McMichael, J. F., Miller, C. A., O'Laughlin, M., Pohl, C., Schmidt, H., Smith, S. M., Walker, J., Wallis, J. W., Wendl, M. C., Wilson, R. K., Wylie, T., Zhang, Q., Burton, R., Jensen, M. A., Kahn, A., Pihl, T., Pot, D., Wan, Y., Levine, D. A., Black, A. D., Bowen, J., Frick, J., Gastier-Foster, J. M., Harper, H. A., Helsel, C., Leraas, K. M., Lichtenberg, T. M., McAllister, C., Ramirez, N. C., Sharpe, S., Wise, L., Zmuda, E., Chanock, S. J., Davidsen, T., Demchok, J. A., Eley, G., Felau, I., Ozenberger, B. A., Sheth, M., Sofia, H., Staudt, L., Tarnuzzer, R., Wang, Z., Yang, L., Zhang, J., Omberg, L., Margolin, A., Raphael, B. J., Vandin, F., Wu, H. T., Leiserson, M. D., Benz, S. C., Vaske, C. J., Noushmehr, H., Knijnenburg, T., Wolf, D., Van 't Veer, L., Collisson, E. A., Anastassiou, D., Ou Yang, T. H., Lopez-Bigas, N., Gonzalez-Perez, A., Tamborero, D., Xia, Z., Li, W., Cho, D. Y., Przytycka, T., Hamilton, M., McGuire, S., Nelander, S., Johansson, P., Jornsten, R., Kling, T., and Sanchez, J. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120. [PubMed Central:PMC3919969] [DOI:10.1038/ng.2764] [PubMed:24071849]. 18

Welcome-Trust (2017). Welcome Trust open access policy. `https://wellcome.ac.uk/funding/managing-grant/open-access-policy`. [Online; accessed 07-November-2017]. 3

West, V. L., Borland, D., and Hammond, W. E. (2015). Innovative information visualization of electronic health record data: a systematic review. *J Am Med Inform Assoc*, 22(2):330–339. [PubMed Central:PMC4394966] [DOI:10.1136/amiajnl-2014-002955] [PubMed:25336597]. 17, 159, 184

White, M. (2012). Digital workplaces. *Business Information Review*, 29(4):205–

214. 6

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, 13(9):R50. [PubMed Central:PMC3491394] [DOI:10.1186/gb-2012-13-9-r50] [PubMed:22951020]. 72

Wikipedia (2017). Title 21 CFR Part 11. `https://en.wikipedia.org/wiki/Title_21_CFR_Part_11`. [Online; accessed 01-November-2017]. 22

Wishart, D. S. (2008). DrugBank and its relevance to pharmacogenomics. *Pharmacogenomics*, 9(8):1155–1162. [DOI:10.2217/14622416.9.8.1155] [PubMed:18681788]. 3, 42, 121

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, 37(Database issue):D603–610. [PubMed Central:PMC2686599] [DOI:10.1093/nar/gkn810] [PubMed:18953024]. 3, 42, 121

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., Nieva de la Hidalga, A., Balcazar Vargas, M. P., Sufi, S., and Goble, C. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.*, 41(Web Server issue):W557–561. [PubMed Central:PMC3692062] [DOI:10.1093/nar/gkt328] [PubMed:23640334]. 29

Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C., and Snoep, J. L. (2011). The SEEK: a platform for sharing data and models

in systems biology. *Meth. Enzymol.*, 500:629–655. [DOI:10.1016/B978-0-12-385118-5.00029-3] [PubMed:21943917]. 12

Worman, H. J., Fong, L. G., Muchir, A., and Young, S. G. (2009). Laminopathies and the long strange trip from basic cell biology to therapy. *J. Clin. Invest.*, 119(7):1825–1836. [PubMed Central:PMC2701866] [DOI:10.1172/JCI37679] [PubMed:19587457]. 101

Wruck, W., Peuker, M., and Regenbrecht, C. R. (2014). Data management strategies for multinational large-scale systems biology projects. *Brief. Bioinformatics*, 15(1):65–78. [PubMed Central:PMC3896927] [DOI:10.1093/bib/bbs064] [PubMed:23047157]. 2

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Honigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N., and Rost, B. (2014). PredictProtein–an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.*, 42(Web Server issue):W337–343. [PubMed Central:PMC4086098] [DOI:10.1093/nar/gku366] [PubMed:24799431]. 3, 8

Yamanaka, T., Wong, H. K., Tosaki, A., Bauer, P. O., Wada, K., Kurosawa, M., Shimogori, T., Hattori, N., and Nukina, N. (2014). Large-scale RNA interference screening in mammalian cells identifies novel regulators of mutant huntingtin aggregation. *PLoS ONE*, 9(4):e93891. [PubMed Central:PMC3976342] [DOI:10.1371/journal.pone.0093891] [PubMed:24705917]. 126

Yang, S. H., Meta, M., Qiao, X., Frost, D., Bauch, J., Coffinier, C., Majumdar, S., Bergo, M. O., Young, S. G., and Fong, L. G. (2006). A farnesyltransferase inhibitor improves disease phenotypes in mice with a Hutchinson-Gilford progeria syndrome mutation. *J. Clin. Invest.*, 116(8):2115–2121. [PubMed Central:PMC1513052] [DOI:10.1172/JCI28968] [PubMed:16862216]. 102, 112

Yun, H., Lee, D. Y., Jeong, J., Lee, S., and Lee, S. Y. (2005). MFAML: a standard data structure for representing and exchanging metabolic flux mod-

els. *Bioinformatics*, 21(15):3329–3330. [DOI:10.1093/bioinformatics/bti502] [PubMed:15905275]. 6

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., and Ide, N. C. (2011). The ClinicalTrials.gov results database–update and key issues. *N. Engl. J. Med.*, 364(9):852–860. [PubMed Central:PMC3066456] [DOI:10.1056/NEJMsa1012065] [PubMed:21366476]. 43

Zhang, H., Das, S., Li, Q. Z., Dragatsis, I., Repa, J., Zeitlin, S., Hajnoczky, G., and Bezprozvanny, I. (2008). Elucidating a normal function of huntingtin by functional and microarray analysis of huntingtin-null mouse embryonic fibroblasts. *BMC Neurosci*, 9:38. [PubMed Central:PMC2377268] [DOI:10.1186/1471-2202-9-38] [PubMed:18412970]. 123

Zhang, Z., Sang, J., Ma, L., Wu, G., Wu, H., Huang, D., Zou, D., Liu, S., Li, A., Hao, L., Tian, M., Xu, C., Wang, X., Wu, J., Xiao, J., Dai, L., Chen, L. L., Hu, S., and Yu, J. (2014). RiceWiki: a wiki-based database for community curation of rice genes. *Nucleic Acids Res.*, 42(Database issue):D1222–1228. [PubMed Central:PMC3964990] [DOI:10.1093/nar/gkt926] [PubMed:24136999]. 10