

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Development of Methods for the Analysis of Chemical Genetic Screens

Xueping Liu

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Dieter Langosch

Prüfer der Dissertation: 1. Univ.-Prof. Dr. Hans-Werner Mewes
2. Univ.-Prof. Dr. Iris Antes

Die Dissertation wurde am 19.05.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 25.09.2014 angenommen.

Abstract

Chemical genetics has emerged in recent years as a study to screen small molecules from large chemical libraries that can be used to explore protein targets and phenotypes. It can be divided into forward and reverse strategies that are termed phenotype-based and target-based screening, respectively. Both experimental approaches can provide valuable tools for the dissection of biological processes, understanding gene function and molecular mechanism of action of small molecules and further speed up the discovery of novel chemical probes or drugs.

With the significantly increased experimental capabilities of performing extensive chemical genetic screens in the past decades, vast amounts of information have been generated. However, the development of methods for the analysis of such huge amounts of high-throughput screening (HTS) data lags far behind the fast rate of chemical screening generation. Many HTS laboratories apply sub-optimal solutions that are either too slow or suffer from a limited scope of analysis due to methodological challenges, such as development of high quality chemical hit identification methods, the identification of protein targets of hits of phenotype-based screens and the detection of promiscuous compounds.

In order to tackle these challenges, in this thesis, I applied a systems biology approach to develop a series of versatile and powerful methods to facilitate the analysis of chemical genetic screens and applied them to the collection of assays stored in ChemBank public repository. Briefly, I determined a hit identification approach that optimally retrieves chemical hits from ChemBank, developed a method to predict protein targets for the hits and introduced an efficient protocol to discard the promiscuous compounds. Some of these methods are available for public usage in a user-friendly web server – HitPick (<http://mips.helmholtz-muenchen.de/proj/hitpick>). In addition, I applied mentioned methods to interrogate the public chemical genetic screens in order to extract novel biological

information.

In the first place, I asked whether chemical screening assay pairs that share selective hits are biologically related. The analysis of the biological activities measured in assays sharing selective hits and the predicted targets of those hits confirmed this hypothesis. I showed that this approach can reveal novel relationships between biological activities as well as uncover novel molecular associations between drug targets and multiple biological activities.

Secondly, I devised a computational strategy to predict the protein targets that upon chemical modulation alter three phenotypes measured in three phenotypic screens. I proved that 88% of the drug targets predicted to affect the three phenotypes are confirmed by literature reports, evidencing the validity of the approach to detect targets related to phenotypes. This novel approach allows to obtain an overview of the druggable molecular repertoire behind the phenotype and to propose novel associations between targets and biological activities.

Zusammenfassung

Die chemische Genetik beschreibt ein in den letzten Jahren etabliertes Forschungsfeld das sich mit der Interaktion kleiner biologisch aktiver Moleküle zur Untersuchung von Zielproteinen und – als Ergebnis dieser Intervention – von Phänotypen befasst. Man unterscheidet zwischen vorwärts gerichteten oder Phänotyp-basierten und rückwärts oder target-basierten Strategien. Beide Ansätze sind nützliche Werkzeuge zur Untersuchung biologischer Prozesse, von Genfunktionen sowie der molekularen Mechanismen biologisch aktiver Moleküle. Sie sind geeignet, neuartige chemische Sonden zu entdecken und die Entwicklung von medizinischen Wirkstoffen zu beschleunigen.

Dank des enormen Zuwachses experimenteller Kapazitäten zur Durchführung umfassender chemisch-genetischer Tests konnten während der letzten Dekaden erhebliche Datenmengen generiert werden. Allerdings ist die Entwicklung adäquater Methoden zur Analyse solcher Datenmengen aus High-Throughput Screenings (HTS) weit hinter dem Durchsatz zurückgefallen, mit der chemische Screenings durchgeführt werden. Viele HTS Labore verwenden sub-optimale Verfahren die entweder zu langsam sind oder bedingt durch die methodischen Herausforderungen nur einen eingeschränkten Analyseumfang haben. Zu diesen Herausforderungen zählen die Entwicklung qualitativ hochwertiger Hit-Identifikationsmethoden, die Identifizierung von Zielproteinen für Hits aus Phänotyp-basierten Screens sowie die Erkennung promiskuitiver Wirkstoffe.

Um diese Herausforderungen anzugehen habe ich im Rahmen dieser Dissertation einen systembiologischen Ansatz angewandt, um eine Reihe von vielseitigen, leistungsfähigen Methoden zu entwickeln die die Analyse chemisch-genetischer Screens erleichtern. Diese habe ich anschließend zur Analyse der in der öffentlichen Datenbank ChemBank vorhandenen Assays verwendet. Ich konnte eine optimale Methode bestimmen die chemische Hits in ChemBank Daten identifiziert und entwickelte eine Methode zur Vorhersage von Zielproteinen für diese Hits.

Ebenso habe ich ein effizientes Protokoll entworfen um promiskuitive Wirkstoffe herauszufiltern. Einige dieser Methoden stehen über einen anwenderfreundlichen Webservice öffentlich der wissenschaftlichen Gemeinde zur Verfügung – HitPick (<http://mips.helmholtz-muenchen.de/proj/hitpick>). Darüber hinaus habe ich diese Methoden auf öffentlich zugängliche chemisch-genetische Screens angewandt um neue biologische Informationen zu gewinnen.

In einem ersten Schritt stellte ich die Frage, ob Paare von chemisch-genetischen Screeningassays, die selektive Hits teilen, biologisch zueinander in Beziehung stehen. Die Analyse der in diesen Assays gemessenen biologischen Aktivitäten sowie die Analyse der vorhergesagten Zielproteine bestätigte diese Hypothese. Ich konnte zeigen, dass der vorgestellte Ansatz sowohl bisher unbekannte Beziehungen zwischen biologischen Aktivitäten als auch molekulare Assoziationen zwischen Wirkstoffzielen und multiplen biologischen Aktivitäten aufdecken kann.

Im zweiten Schritt habe ich eine Strategie entwickelt, um Zielproteine vorherzusagen deren chemischen Modulationen drei Phänotypen beeinflussen, welche in drei verschiedenen Assays untersucht wurden. Ich konnte zeigen, dass die Phänotyp-spezifische Wirkung von 88% dieser vorhergesagten Wirkstoffziele bereits in der wissenschaftlichen Literatur belegt wurde. Dies beweist die Validität dieses Ansatzes zur Identifikation von Wirkstoffzielen, die mit spezifischen Phänotypen in Zusammenhang stehen. Dieser neuartige Ansatz erlaubt es, sich einen Überblick über das wirkstoff-zugängliche molekulare Repertoire bzgl. eines Phänotyps zu verschaffen und neue Assoziationen zwischen Wirkstoffzielen und biologischen Aktivitäten vorherzusagen.

List of Publications

1. X. Liu*, I. Vogt*, T. Haque, and M. Campillos, (2013) “HitPick: a web server for hit identification and target prediction of chemical screenings”, *Bioinformatics* **29**, 1910–1912 (* equal contributions)
2. X. Liu and M. Campillos, “Chemical screening assay pairs that share selective hits are biologically related”, *submitted*
3. X. Liu et al., “Target identification in high-throughput phenotypic screens”, *to be submitted*

Acknowledgements

At this point of my thesis, before the actual scientific content, I would like to express my acknowledgement to all the people who made it possible.

There are no proper words to fully convey my deepest gratitude to my direct supervisor – Dr. Monica Campillos, the head of Systems Biology of Small Molecules Research Group at the Helmholtz Zentrum München. Foremost, I sincerely thank Dr. Campillos for accepting me as a PhD student and for giving me the opportunity to conduct my PhD research in her group. Our regular scientific discussions were the source of inspiration for this work. Her suggestions, ideas and motivation helped me to successfully conclude my research. Without her commitment, this thesis would not have taken its present form. I am also very grateful to Dr. Campillos for the academic freedom that promoted development of my own, independent and scientific thinking.

I would like to thank Prof. Dr. Hans-Werner Mewes, the chair of the Genome-Oriented Bioinformatics department at the Technische Universität München, for being my supervisor at the University, participating in the thesis committee meetings from the Helmholtz Graduate School (HELENA), and all our scientific discussions as well as his insightful suggestions that resulted in improvements of this thesis.

Furthermore, I also thank the other two thesis committee members from the HELENA – Prof. Dr. Michael Sattler and Dr. Kamyar Hadian, for attending the annual meetings and providing clear plan with constructive suggestions toward my work, which allowed me to make improvements of the project, so that I was able to finish this thesis on planned time.

There are a number of my present and former colleagues who I would like to thank for their help, discussion and advice. Especially, Dr. Ingo Vogt was always of great help with preparation of the target prediction models. Together with Tanzeem Haque, Jonathan Hoser and Jeanette Prinz, they helped me a lot

with the technical support for HitPick. Verena Friedl, Sabrina Hecht and Johannes Höffler helped me with the comparison between SEA and HitPick target prediction method. I express my profound gratitude to Julia-Sophie Heier who proposed the excellent computational method and did preliminary analysis on the “Modulators of Wnt signaling” assay project. Also, I acknowledge Dr. Benedikt Wachinger and Dr. Volker Stümpflen from Clueda AG for the support of providing the EXCERBT text-mining tool. Besides, I would like to thank Dr. Corinna Montrone, Dr. Gisela Fobo, Dr. Barbara Brauner and Dr. Andreas Ruepp for the great amount of help in annotating the relationship between targets and phenotypes, as well as their helpful suggestions and discussions.

Moreover, I owe many thanks to the secretaries – Ms. Elisabeth Noheimer, Ms. Petra Fuhrmann, Ms. Maria Singer and Ms. Alessia Dell’Acqua, who greatly and thoroughly supported me with a lot of administrative matters related to my graduation.

Finally, I am extremely grateful to my parents and sisters for their unconditional support and constant encouragement. Last but definitely not least, I would like to sincerely thank Tomasz for always being there for me.

Munich, in May 2014

Xueping Liu

To my Family

*One never notices what has been done;
one can only see what remains to be done.*

– Marie Curie

Contents

Contents	1
1 Introduction	7
1.1 Genetics and Pharmacology	7
1.1.1 Genetics	7
1.1.2 Pharmacology	9
1.2 Chemical genetics	10
1.3 Chemical genetic process	12
1.3.1 Chemical library	12
1.3.2 Bioassay	14
1.3.3 Compound signal analysis	17
1.4 Application of chemical genetics	18
1.5 From HTS to ultimate drug discovery	18
1.6 Challenges of chemical genetics	19
1.6.1 Chemical hit identification	21
1.6.2 Target prediction	21
1.6.3 Promiscuity	23
1.7 Computational systems biology in chemical genetics	23
1.7.1 Systems biology	24
1.8 The goal and significance of the project	24
1.8.1 Goal	24
1.8.2 Significance	26
2 Materials and Methods	27
2.1 ChemBank	27
2.1.1 ChemBank assay data structure	27

2.2	Chauvenet's criterion	28
2.3	Composite Z-score and Reproducibility	29
2.4	Median polish procedure	29
2.5	Median absolute deviation	31
2.6	B-Score	31
2.7	Receiver operating characteristic space	31
2.8	Target prediction in HitPick	32
2.8.1	Database	33
2.8.2	Fingerprints	33
2.8.3	1NN similarity searching	33
2.8.4	Laplacian-modified naïve Bayesian target models	34
2.8.5	Combination of 1NN similarity searching and Laplacian-modified naïve Bayesian target models	36
2.8.6	MaxMinAlgorithm	37
2.9	Calculation of hit similarity	38
2.10	Similarity of assay project by applying EXCERBT	38
2.11	Promiscuity filters	39
2.12	Identification of significantly over-represented enriched targets	39
3	Chemical Hit Identification	43
3.1	Results and discussion	45
3.1.1	The ChemBank method to identify hits	45
3.1.2	The B-Score method to identify hits	47
3.1.3	The B-Score_A method to identify hits	49
3.1.4	The Well-Correction method to identify hits	49
3.1.5	Modifications of the above four methods	50
3.1.6	Performance comparison among the eight methods	50
3.2	Conclusions	53
4	HitPick	55
4.1	Results	56
4.1.1	Performance of target prediction	56
4.1.2	Implementation	58
4.1.3	Processing time	59
4.1.4	Privacy	61

4.2	Discussion	61
4.3	Conclusions	62
5	Chemical Screening Assay Pairs that Share Selective Hits Are Biologically Related	65
5.1	Results	66
5.1.1	ChemBank structure and chemical hit identification . . .	66
5.1.2	Promiscuity filters and similarity in biological activity . . .	69
5.1.3	Assay interaction network	71
5.2	Discussion	76
5.3	Conclusions	78
6	Target Identification in High-Throughput Phenotypic Screens	81
6.1	Results	82
6.1.1	Molecular space explained by enriched targets	84
6.1.2	Validation of the approach based on literature	84
6.1.3	Validation of the approach based on known activity of hits	90
6.2	Discussion	92
6.3	Conclusions	94
7	Summary and Outlook	97
7.1	Scientific achievements	97
7.2	Final conclusions	100
7.3	Extensions and future directions	100
	Bibliography	103
A	Additional data	127

Structure of the Thesis

Below I summarize the chapters of this thesis:

The first chapter is an introductory section intended to explain the evolution, definition, and function of chemical genetics in more details to provide the right context to readers not familiar with the topic of this thesis. I describe each essential element of the chemical genetics process, discuss common challenges that the field is facing, such as hit identification, and critically review various biochemical, genetic and computational approaches recently developed for target identification. Based on these challenges, I also give an insight into various methods in the field as well as provide a series of analysis tools to facilitate the analysis of chemical genetic studies.

The second chapter introduces the data and provides details and explanations of the methodology followed to process the chemical genetics data, to tackle the challenges and to infer novel biological information.

In the third chapter of my thesis, I intend to cope with the challenge of hit identification. First, I introduce eight different hit identification methods that I applied on ChemBank assays and discuss the weakness of each method when addressing the systematic variation of signals. Then, I explain how I determined the best method for chemical hit identification by comparing their performance when discriminating positive and negative controls in the assays.

In the fourth chapter, I focus on addressing the challenge of target prediction of small molecules. I explain the creation of a novel drug target prediction method based on a combination of two 2D molecular similarity based methods, namely, 1-nearest-neighbor (1NN) similarity searching and Laplacian-modified naïve Bayesian target models. This method along with B-Score, a well-known chemical hit detection method, were implemented for public usage in a web server – HitPick. In the end of this chapter, I also explain how to use the hit identification

and target prediction functions implemented in HitPick.

In the fifth chapter, I first introduce an efficient protocol to remove promiscuous compounds. I describe the application of this filter for detecting selective hits and use HitPick to predict their drug targets. Afterwards, I test the hypothesis of whether the biological activities of pairs of chemical screening assays sharing selective hits are related. The analysis of biological activities measured in the assays confirmed this hypothesis. This finding was reinforced by the biological role of the predicted targets of shared hits as they evidenced known associations between targets and the two biological processes measured in the assays, such as the enrichment of known anticancer targets in the growth inhibition screens. It allowed me to propose novel associations between them, like the potential growth inhibitory effect of ATP2A1, etc.

In the sixth chapter, I incorporate the methods that I developed, that is, hit identification and HitPick target prediction, in a computational strategy to detect drug targets statistically associated to biological processes. To demonstrate the powerfulness of my approach, I applied this computational approach to three case studies, and validated the drug target-phenotype relationships by literature reports. Further, I was able to propose novel targets involved in the observed phenotype as well.

The seventh chapter outlines, concludes and places in a wider picture all results of the computational studies on the chemical genetic data. Finally, this chapter offers an outlook on the future studies motivated by the findings from the present thesis.

Chapter 1

Introduction

1.1 Genetics and Pharmacology

Genetics and pharmacology are the two available principal approaches aiming at discovering the protein function in cells of an organism. In the following sections I briefly introduce and explain them.

1.1.1 Genetics

Since the seminal study of pea genetics by Mendel in 1865, genetics has been widely used to study biology by manipulating the biological system at the level of the gene. The function of gene products - proteins - is what researchers ultimately desire to understand, and the perturbation of gene function is one of the most direct ways to identify the protein function [1]. Genetically, gene function can be modulated through a mutation, such as DNA substitution, deletion, insertion, etc. [2], which can result from the action of physical and chemical mutagens [3].

Once a series of gene mutants in a biological model, including cells lines or organisms, have been yielded, generally one needs to check out thousands of individuals to locate the altered phenotype of interest, like a modified behavior, appearance, etc. Such a search in a mutagenized population is called as genetic screen [4]. These mutants are then used to find and study the genes that regulate the biological processes or pathways. This strategy is defined as “forward genetics”, that is, from phenotype to gene, involving the random mutagenesis in collaboration with screening with the aim to identify a gene that particularly

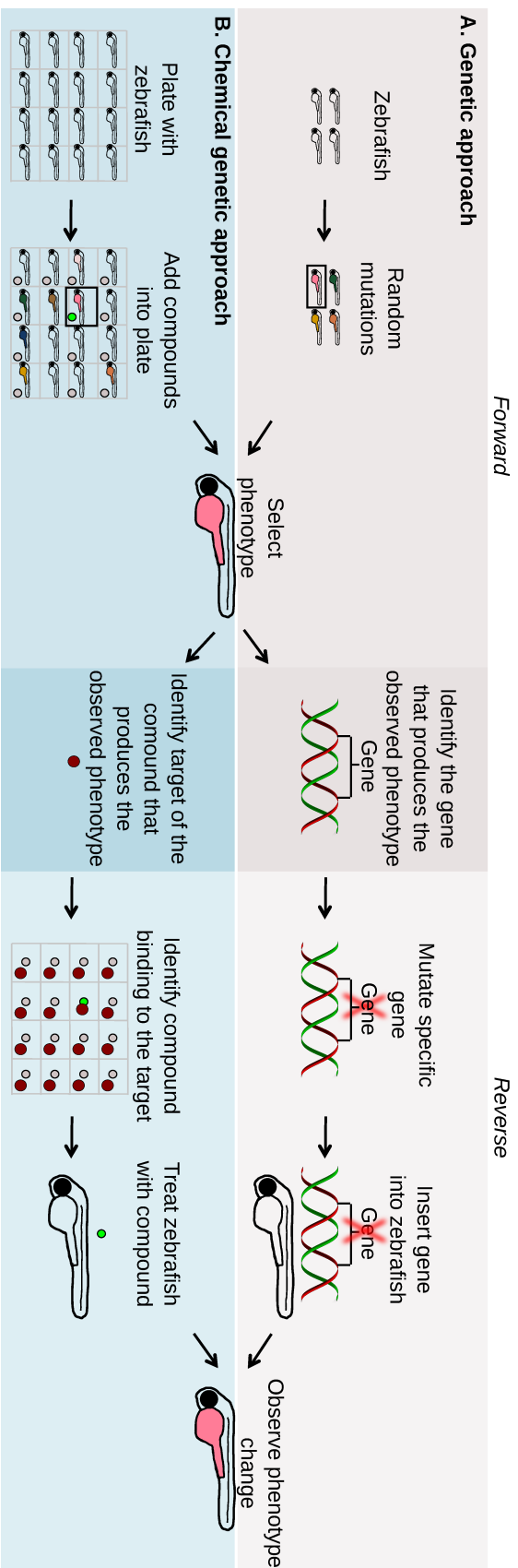


Figure 1.1.1: Forward and reverse techniques for genetics and chemical genetics to identify genes and proteins that are responsible for a particular target or biological process. (A) Genetics. In the forward direction, it can be divided into three discrete steps. Firstly, the random genetic mutations are introduced. After selecting the desired phenotype, the gene for the desired phenotype is then identified. In the reverse direction, the starting point is to mutate a particular gene, and then the phenotypic change by the gene is searched. (B) Chemical genetics. Similar to the genetic approach, in the forward direction there are three steps included. A collection of small molecules is firstly added into the cells or organisms of interest, compounds that can produce the phenotype are then selected, and finally the cellular molecular targets (such as proteins, lipids, etc.) of the bioactive compounds are identified. In the reverse direction, bioactive ligands of the specific protein are searched from the large pool of small molecules.

produces the phenotype (Fig. 1.1.1A) [1].

In addition to forward genetics, reverse genetic methods are also available to ascertain the gene functions (Fig. 1.1.1A). The mutation of a particular known gene by DNA engineering methods makes the gene become a permanent part of the genome [5] and the resulting phenotype helps to determine role of the gene in the cell or organism. Such approach is called as “reverse genetics” that is from gene to phenotype [1].

1.1.2 Pharmacology

Pharmacology, a discipline of biomedical science, is the combination of biology and medicine with the aim to provide an understanding of the effect of drugs on human [6]. The basis of classic drugs is formed by small molecules, which are described as those carbon based compounds whose molecular weight is usually less than 500 Daltons and always smaller than that of macromolecules, such as DNA, RNA and proteins [7]. The tiny size in structure and chemical composition of small molecules often help them easily pass through cell membranes and thus, if the drugs are more effective and less toxic than previous generations, then they are normally processed into ingestible tablets or capsules to reach the desired destination in the body and further cure the diseases. For example, in 1929, Alexander Fleming [8] discovered the small molecule penicillin as the most compelling case for antibiotics. Penicillin exerts its cytotoxic effect through the inhibition of the cross-linking of small peptide chains in peptidoglycan, the main cell wall polymer of bacteria that was formed via binding of the four-membered β -lactam ring of penicillin to the enzyme D-Ala-D-Ala carboxypeptidase/transpeptidase (DD-peptidases) [9]. The existing bacterial cells will not be influenced by the treatment of penicillin; however all the newly produced cells will grow abnormally due to the impairment of cell walls, and thus they are prone to osmotic lysis.

Although both genetics and pharmacology can be used to study the function of proteins, the two techniques sometimes can evoke notably different phenotypes even when they target the same protein [10]. For example, there is a paradox phenomenon when antidiabetic thiazolidinediones (TZDs) and genetic manipulation are used to modulate the target PPAR- γ , a nuclear hormone receptor involved in adipogenesis [11]. More detailed, from the pharmacological point of view, TZDs are the marketed drugs for treating type 2 diabetes; however, this treatment results

in the direct activation of PPAR- γ [12], which is a transcription factor to promote adipogenesis. Quite unexpectedly, from the genetic point of view, heterozygous deletion of PPAR- γ gene actually prevents insulin resistance in mice [13], pointing that only PPAR- γ inhibitors, instead of activators, can be developed as antidiabetic drugs. Eventually, Yamauchi et al. [14] explained the paradox and showed that pharmacological agonists and genetic antagonists of PPAR- γ can both improve glucose metabolism through different mechanisms. They have shown that TZD drugs clinically increase insulin sensitivity in muscle and liver by elevating number of small adipocytes and weight gain; by contrast, genetic antagonists reduce insulin resistance by potentiating leptin's effect, increasing fatty acid burning and energy dissipation [14]. This example revealed TZDs as the therapeutic potential of PPAR- γ agonists that could not be predicted from genetic analysis. In addition to this, there are many other well described examples shown by Knight and Shokat [10] who illustrated that due to the different ways to perturb the activity of a protein by a small molecule and a genetic mutation, different phenotypes can emerge.

The most important strength of genetic approach is that it is possible to unconditionally detect the mutation that is responsible for the observed phenotype. However, this approach is impossible to effectively identify the cellular targets of the small molecules. This ambiguity ignores direct comparisons between genetic and pharmacological phenotypes because it is possible that any of the differences can reflect off-targets of the drug [10]. Furthermore, because the mutations in some essential proteins often lead to lethality at early stage making subsequent study impossible, it is difficult to control the protein function using genetic knock-out/knockin experiments [15].

One way to overcome these difficulties is to perturb cellular function by small molecules targeting a protein (the 'mutations'), an approach referred to as "chemical genetics" which is rising markedly worldwide.

1.2 Chemical genetics

Chemical genetics has emerged over the last 10 years [16, 17] as a study to distinguish small molecules from large chemical libraries that can be used to explore protein targets and signal transduction pathways [7]. It is a discipline where genetics and pharmacology meet [10]. The booming number of publications on

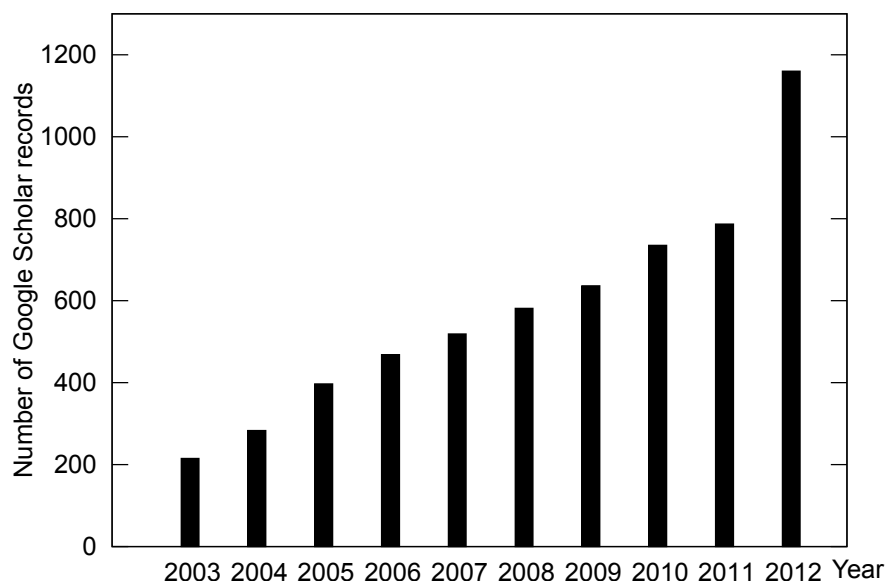


Figure 1.2.1: Increased tendency of applications of chemical genetics. The number of literature reports is retrieved by the search term “chemical genetics” from the year of 2003 to 2012.

chemical genetics within the last decade strongly shows the growing interest in this field (Fig. 1.2.1).

By analogy to genetics, chemical genetics can as well be divided into forward and reverse strategies (Fig. 1.1.1B). Forward chemical genetics involves directly screening small molecules against one or a few desired phenotypic effects in a cellular, or even whole organism-based context in order to identify active chemicals. Afterwards, identification of the protein target(s) that induce the observed phenotype is required [17]. Reverse chemical genetics uses small molecules targeting directly a protein of interest (e.g. enzymatic assay or protein-DNA interaction studies) in a cell-free context. Once the active compound targeting a given protein is identified, then the challenge is to check if the changed phenotype can be observed by including the active compound in a cellular context [17].

In both forward and reverse directions, the identification of selective small molecules followed by detailed biological investigation is required [18, 19].

In addition to having the potential of deriving new drugs, chemical genetics still has many advantages over classical genetic techniques due to the perturbation of protein functions by selective small molecules in the biological system.

In the first place, small molecules are easy to apply on different cell types

of interest and they work rapidly and often reversibly [20]. Hence, they offer excellent temporal tools to switch the processes on and off by adding or removing the compound. For example, brefeldin A has been reported to block the process of protein transport from the endoplasmic reticulum (ER) to Golgi in different organisms, such as yeasts, plants and mammalian cells [21].

Secondly, instead of simply turning protein activity up or down, small molecules can also alter protein translation and transcription in more subtle ways, such as modulating one of its several functions [22]. For example, histone deacetylase 6 (HDAC6) has two distinct and active catalytic domains, but only one of them possesses α -tubulin deacetylase activity that can be selectively inhibited by small molecule tubacin [23].

Last but not least, small molecules can be used for chemical combination interventions, making them especially advantageous for integrating systems and chemical biology [24, 25]. Furthermore, synergistic and potentiative drug combinations have been explored to achieve one or more favorable outcomes, such as enhanced safety and efficacy and decreased drug resistance, etc. [26], which are the adorable effects for the treatment of complex diseases, like cancer and cardiovascular disease.

1.3 Chemical genetic process

Chemical library, bioassay and compound signal analysis are the three foundations involved in the outline of a chemical genetic process (Fig. 1.3.1). In the following sections I describe them more in detail.

1.3.1 Chemical library

As the application of small molecules to living organisms can mimic the effect of mutated strains or organisms that are essential for genetic studies, the preparation of a chemical library is crucial as well for every chemical genetic screening. The first chemical libraries were assembled over the past century by pharmaceutical companies whose aim was to find novel drugs [27]. The library can be composed of low-molecular-weight organic molecules, and can also include peptide aptamers [28, 29] (Fig. 1.3.1). Historically, the majority of the agents were synthesized based on the already existing biologically active small molecules - either from Food and Drug Administration (FDA) approved drugs or

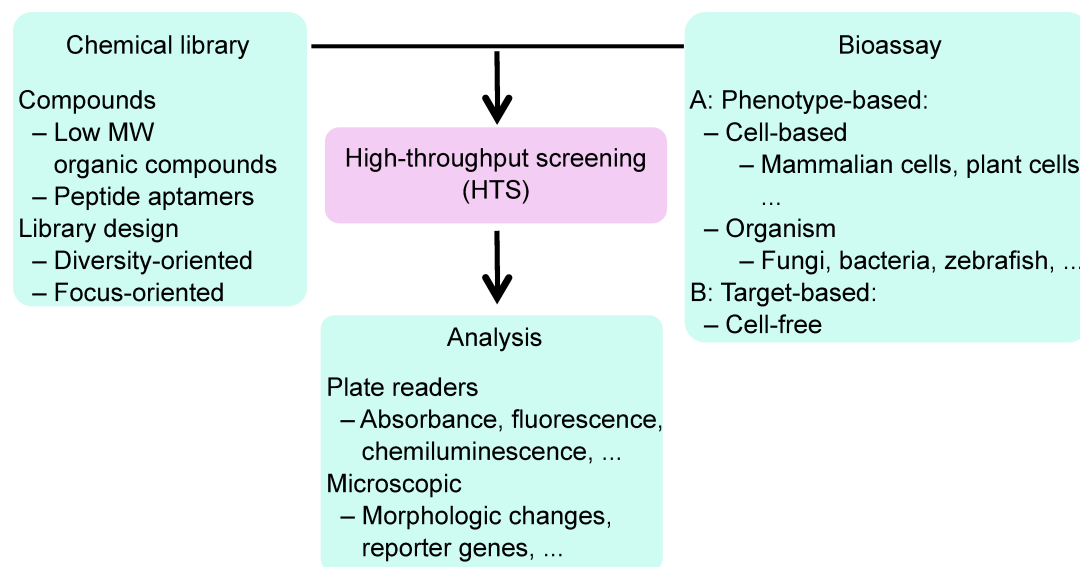


Figure 1.3.1: Chemical genetic process.

natural products [30]. However, chemical genetic screening approach was largely unavailable for academic researchers until the expansion of interest in chemical genetics along with the proliferation of commercial chemical library suppliers, such as Enamine (<http://www.enamine.net/>, 1.8 million compounds), ChemDiv (<http://eu.chemdiv.com/>, 1.5 million compounds), BioFocus (<http://www.biofocus.com/>, 0.9 million compounds), etc. In addition to these commercial library vendors, non-profit research organizations that offer small-molecule libraries are also available, such as the “Diversity Set IV” of the National Cancer Institute that can provide around 1,600 compounds. Although each of the available library displays a high degree of diversity in structures, the individual compounds in these libraries typically fulfill the following criteria: (i) they are easy to penetrate cell membranes; (ii) they possess well solubility in organic solvents, such as dimethyl sulfoxide (DMSO); (iii) they are metabolically stable in liver, plasma, etc.; (iv) they contain substructures resembling known bioactive molecules; (v) they do not contain “functional groups” (e.g. highly reactive groups) that are possible to produce cytotoxic effects [20, 31].

In practice, there are basically two types of chemical libraries that are synthesized today, “diversity-oriented libraries” and “focused libraries” [32,33]. “Diversity-oriented libraries” contain diverse collections of small molecules that can target any protein classes. They are normally used in broad screens in which the targets

are unknown and thus offer the opportunity to discover new classes of targets [34]. In contrast, “focused libraries” include compounds which have been designed or assembled with a specific protein target or protein family, such as proteases [35], kinases [36], phosphatases [37] and G-protein coupled receptors [38]. The advantage of screening a focused library is that compared to “diverse-oriented library”, fewer compounds are needed to be screened to get chemical hits for a particular target, and it generates higher hit rates [39]. Based on the goal of the experiments, the screeners should decide the nature of the library. For example, if chemical biology researchers attempt to create and assess the vast potential chemical space for an unexplored bioactivity, then the diverse-oriented library should be chosen [7].

Regardless of the source of chemical library, it is of the utmost importance that the provider and users of libraries must adopt quality and reporting standards to advance the impact of small-molecule high-throughput screening (HTS) [40]. Compound purity, stability, accuracy of compound concentration, sufficient solubility and lack of notoriously toxic or promiscuous molecules are all factors of library quality and should be carefully considered before the screening starts [41].

1.3.2 Bioassay

A biological screen is also called as “assay”, “HTS assay”, “bioassay” or “chemical genetic assay”. Considering the challenge that typically thousands of compounds have to be analyzed to find the desired bioactive small molecules in a bioassay, it is evident that HTS should be set up in a robust way in a model system. “High throughput” is a relative term as generally it is defined as the screening of 10,000 to 100,000 compounds per day [42], which is a major technological break-through in biological experimentation [43]. Designing a suitable bioassay is vital to the success of the drug discovery. Generally, bioassays can run by multi-well assay plates (96-, 384-, 1536-, 3456- and even can extend to 6144-well [44–46] in a parallel fashion. Assays that are run in 1536-, 3456- and 6144-well plates are referred as ultra HTS (uHTS).

The possible model systems of the screening can vary from cell-free to cell-based or even the whole organisms and chemical genetic approaches can be applied in forward or reverse directions. Forward chemical genetics is also called as “phenotype-based” or “cell-based” screening (Fig. 1.3.1), which detects the small molecules that can induce a specific phenotype in a cell (including mammalian

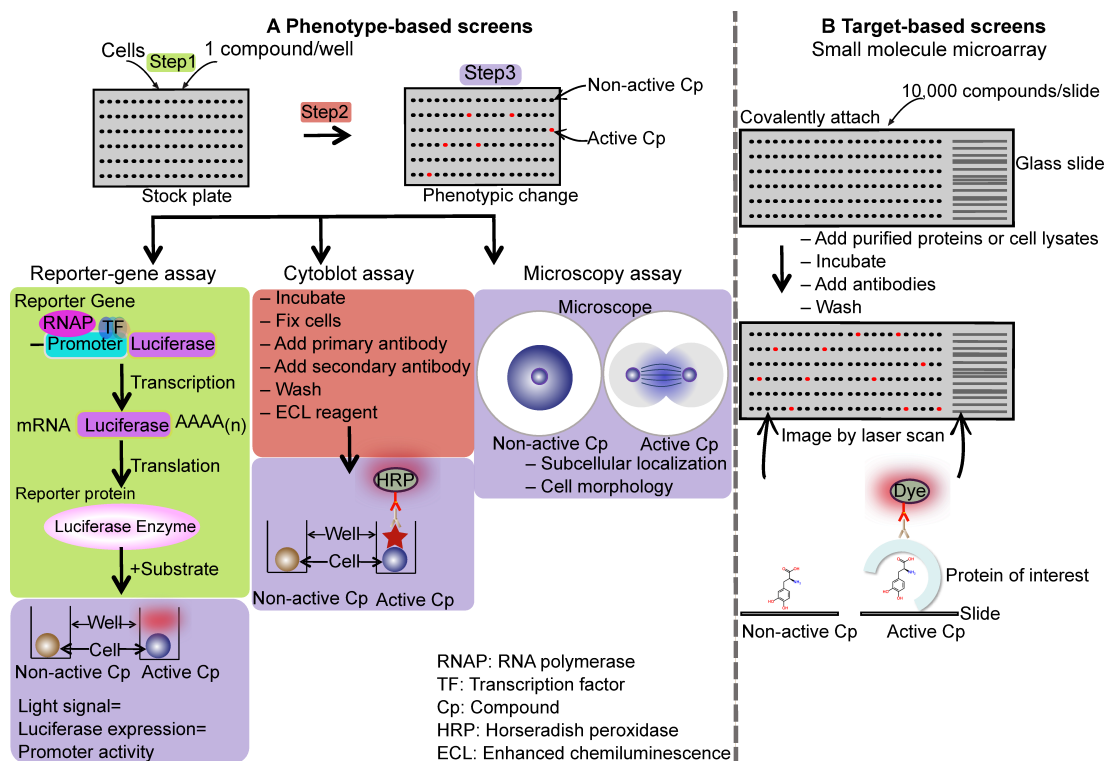


Figure 1.3.2: Schematic representation of four different small-molecule screening technologies. A. Phenotype-based screening. B. Target-based screening. This figure was adapted from Ref. [7, 17].

and plant cells, etc.) or whole organism (including single-cell organism, such as bacteria and fungi, and multicellular organism, such as zebrafish). Reverse chemical genetics is also called as “target-based” or “cell-free” screening (Fig. 1.3.1) and is done on “pure protein”. This approach identifies ligands for some specific protein of interest *in vitro*, and afterwards, these active ligands are tested *in vivo* to investigate their activity on physiological conditions.

Approaches applied on phenotype-based screening

There are three types of biological assays commonly used in phenotype-based screens, namely reporter gene, cytotblot and microscopy assays [7, 17, 22] (Fig. 1.3.2A).

Reporter gene assay

Reporter gene assay is a common cell-based approach in which the accumulation of an easily detectable enzymatic activity such as luciferase, depends on the activity of a gene promoter [20]. This method is used to measure the gene

promoter activity, however it does not fully account for the complete regulation of the gene due to the lack of distant promoter sites in the construct. For rapid and quantitative assessment of gene promoter activity, the reporter gene (often luciferase, used as a light-based technology) is cloned downstream of the promoter fragment (Fig. 1.3.2A). Luciferase-based reporter assays are quite powerful to detect the changes of gene expression within the cells at a specific promoter due to their ultrasensitive detection capacity and wide dynamic range. These assays involve transferring the resulting reporter construct into cells or whole organism via transfection, transformation or injection. The self-multiplication or mating between wild type and homozygous transgenic reporter animals generates progeny of 100% heterozygous reporter cells or embryos [47]. The expression of the luciferase reporter gene can be quantified by measuring the released light.

Cytoblot assay

The cytoblot approach makes use of functional readouts (e.g. cell viability) or of “whole-cell immunoassays” (Fig. 1.3.2A), a luminescence-based method that uses a suitable antibody to detect an epitope in cells whose occurrence or disappearance can be used as readout for a specific process of interest [20,46]. Growing cells are seeded onto the bottom of wells and a single compound is added to each well. After incubation of cells and compounds, cells are fixed and then a primary antibody of desired specificity is added. Later, a secondary antibody covalently linked to horseradish peroxidase is added, and finally, the enhanced chemiluminescence reagent is used to detect the antibody complex [7]. Because antibodies can directly recognize proteins and specific protein modifications, cytoblot cell-based assay is able to screen for biosynthetic processes, such as DNA synthesis as well as post-translational protein modifications, such as acetylation and phosphorylation [46].

Microscopy assay

The aforementioned two cell-based methods are not biased towards one specific protein but they are conceptually broader, offering the potential to find compounds regulating poorly characterized or even unknown targets [20]. Unlike the above two methods, microscope-based approach represents a new and exciting trend in cell-based screening towards acquiring more complex data [20]. This approach is capable to detect even subtle morphological changes (Fig. 1.3.2A).

Moreover, it allows the collection of data on the particular process under investigation and data on effects indirectly influencing the process of interest [20]. Recently, this type of approach, which obviously demands the use of automated microscopes and image-analysis software [48], has been successfully applied to identify small molecules that interfere with a wide range of biological processes, such as embryonic development [49], cell differentiation [50] and the transport of intracellular vehicles [51]. For example, if a screen searches for inhibitors of cell migration, the total cell count in a counting chamber can then be used as the parameter to identify the cytotoxic compounds, whose inhibitory activity on cell migration is indirectly caused by the fact that the cells are dying.

The experimental procedure of microscope-based assay is as follows: after incubation with cells and compounds in the plate, morphological or subcellular localization changes in the cells can be visualized by microscopy. It includes “nuclear foci formation assay”, “cell morphology assay” and “protein translocation assay”.

Approaches applied on target-based screening

A powerful method for target-based assay is the small molecule microarrays (SMM) (Fig. 1.3.2B). Maximum 10,000 small molecules are firstly covalently attached onto a glass slide in high density [7, 22]. Subsequently, the microarray is incubated with purified proteins or cell lysates, which are called as “purified protein binding assay” and “cell lysate binding assay”, respectively; afterwards, a primary antibody against a protein of interest and a secondary antibody conjugated with a fluorescent dye are added [7]. Finally, a laser scan of the entire slide can be used to detect the binding of proteins to selective small molecules [7].

1.3.3 Compound signal analysis

After the HTS, setting up a screening platform with mechanization that ranges from manually operated workstations to fully automatic robotic systems is required to analyze the effect of the compounds on the model system [42]. Depending on the output needed for the results, the detection method includes “fluorescence”, “absorbance” and “luminescence” (Fig. 1.3.1).

1.4 Application of chemical genetics

There is no doubt that small molecules are invaluable tools for the dissection of complex biological processes, understanding gene function and molecular mechanisms of action of selective small molecules via the perturbations of protein functions. For example, taxol, a compound known to disturb the microtubule dynamics has been used to understand cell cycle [52]. In addition, because small molecules can be used to develop drugs or chemical probes, the chemical genetic approaches have become the cornerstone technology in the majority of pharmaceutical companies [53]. The final goal of the HTS is to accelerate modern drug discovery and development process by screening large small molecules libraries [54].

1.5 From HTS to ultimate drug discovery

Target-based screening is basically the chemical approach pharmaceutical industry follows, and has allowed many major advances in the development of drug discovery [55], such as the discovery of the antiretroviral drug maraviroc by Pfizer and approved by the FDA in 2007 (trade name: Selzentry or Celsentri). For the discovery of maraviroc, in 1997 a screening based on a CC-chemokine receptor 5 (CCR5) protein binding assay was employed [56].

A prerequisite for a target-based screen is a reasonably well-characterized target, such as the CCR5 for maraviroc, which is somewhat limiting when exploring new molecular mechanisms of phenotypes and new fields of biology. In contrast, phenotype-based screening allows to explore poorly molecularly characterized phenotypes and thus, has become a renewed approach for drug discovery [57, 58]. In fact, Swinney and Anthony [59] analyzed the mechanisms and methods of discovery for first-in-class FDA approved drugs and showed that between 1999 and 2008, 37% of them were discovered in phenotypic assays. For example, one of the FDA approved drugs in 2008 derived from a cell-based luciferase reporter assay run by GlaxoSmithKline in 1997 is eltrombopag (trade name: Promacta or Revolade) [60].

In addition to the selection of either target-based or phenotype-based screening strategies for the discovery of drugs or chemical probes, an effective HTS strategy considers both primary and subsequent secondary assay designs carefully

(Fig. 1.5.1). In the primary assay, the nature of the response to be measured should be clearly defined. The screeners should clearly know which signal response is what they are interested in, that is, whether the signal should increase, decrease, change in nature of the location, or belong to part of a more complex response. The active compounds are labeled “hits” and can be used for follow-up assays.

The follow-up assays are also referred as secondary screens or counter screens. Ideally, the only difference between the primary and follow-up assay is the protein target or the countering property such as cytotoxicity, whereas other reagents and parameters such as concentration of the compounds, should be the same. Secondary screen evaluates the involvement of compounds in the intended biological interaction, and consequently assists in the recognition of compounds that generate the positive signal through other mechanisms, such as showing activity on other related targets or inhibiting a cellular pathway from a cytotoxic response [42]. If the activity is observed in both primary and secondary assays, then the small molecule is likely to be a false positive hit. Furthermore, the testing of primary active hits on secondary screens is also expected to remove artefacts caused by aggregate formation and small molecule precipitation [62]. In the secondary assays, only few compounds (around 1% of the most active compounds from the primary screens [61]) are generated and at least duplicates are typically used. The active compounds are termed “confirmed hits”. If these hits have an established biological activity according to a structure-activity relationship (SAR) series and medical chemistry, then they are termed “leads”, which can be used to develop for drug candidates in clinical trials [61]. On average one lead compound is identified for every 120,000 compounds screened [63].

1.6 Challenges of chemical genetics

Chemical genetic screening is definitely a useful tool for drug candidate generation; nevertheless, the development of analysis methods that are able to handle and process the chemical screening data is lagging behind. The most crucial issues that need to be tackled are chemical hit identification, target prediction of small molecules in phenotypic screens and elimination of unspecific compounds that show activity in multiple assays. In the light of these challenges, efforts from the computational systems biology field is required to generate novel and suitable

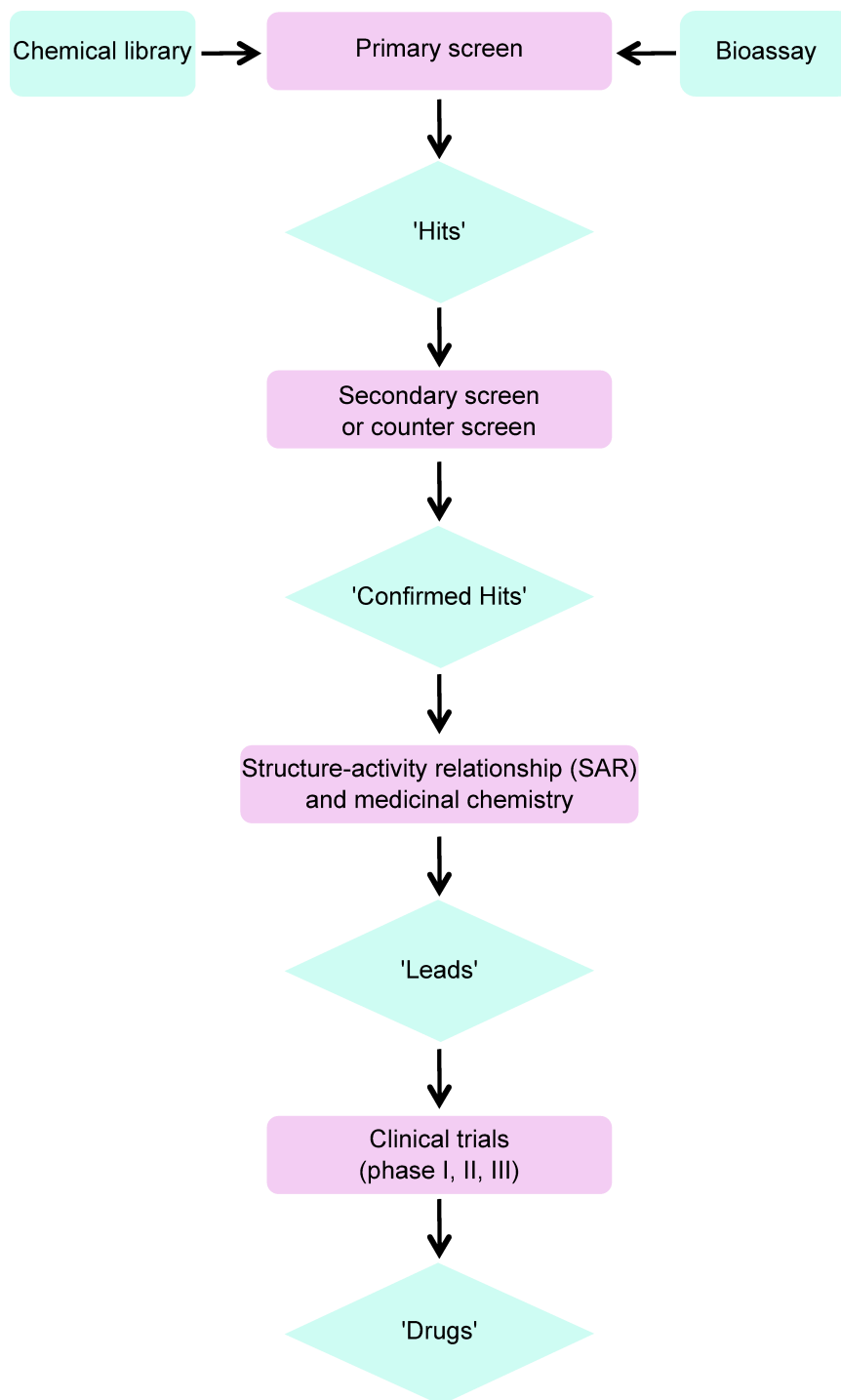


Figure 1.5.1: From HTS process to ultimate drug development. This figure was adapted from Ref. [61].

approaches to analyze the chemical genetic data and finally maybe with the cooperation of biologists to validate the active compounds and their molecular targets *in vivo*.

1.6.1 Chemical hit identification

A great number of chemical biology researchers employ electronic spreadsheets as the essential data analysis tool, resulting that their data exploration capabilities are extremely limited [64]. Based on the activity values of compounds, the experimental researchers usually sort the compounds and choose a cut-off manually while searching for a trade-off between the number of hits and the risk of missing important molecules. Their behavior obviously has several disadvantages: (i) they leave room for several types of mistakes such as subjective judgment and tiredness; (ii) the systematic errors in the assay plates are difficult to detect; (iii) they also hinder the possibility to effectively explore the chemical space and to associate the activity levels of compounds with structural features [64]. To overcome these problems some computational methods have been developed including the ChemBank [65], the B-Score [61] and the Well-Correction [66], although their performance in terms of discrimination between positive and negative controls in the assays is poor (see Chapter 3 Chemical Hit Identification). Therefore, there is an urgent need for novel automatic methods capable to correct systematic errors in an accurate and fast manner.

1.6.2 Target prediction

There is no guarantee that the chemical hits of target-based screening will be cell-permeable or affect the protein in a way that results in a functional phenotype in cell-based or organism-based context, and indeed, many drugs that are identified in these screens ultimately fail [59, 67]. As a consequence, phenotype-based screening is emerging as a cost-efficient and translational small-molecule discovery technology to identify efficacious therapeutics. For example, an innovative drug discovery strategy has been reported recently in larval zebrafish to identify metabolically active drugs with potential therapeutic function [47].

In the phenotype-based screens, however, the crucial challenge is to identify the target altered by the small molecule for the observed phenotype [55]. Several strategies have been proposed in the identification of drugs targets through affinity

chromatography, genetic interaction and computational approaches (reviewed in Ref. [68]), which significantly help to alleviate this problem.

Affinity chromatography approaches

Affinity chromatography techniques use traceable (radioactive or otherwise tagged - for example - biotin [20]) compound derivatives or compounds bound to solid-phase matrices to detect protein targets of compounds. They have successfully identified the protein targets of acetylcholine, steroids and natural products such as cyclosporin and rapamycin [1]. However, this strategy requires a stringent criteria - the abundance of the target protein(s) and the strong affinity between protein and small molecule partner - that rarely meet [69, 70].

Genetic interaction approaches

The advent of whole-genome sequence information has allowed the application of the new gene-based approaches for drug target identification. Relying on the idea of genetic modifiers (activators or inhibitors of the gene), genetic methods use the principle of genetic interaction to stimulate hypothesis of targets [68]. As both genetic methods (such as gene knock down and RNA interference) and chemical interference approaches are biological tools to alter the protein functions in organism (reviewed in [68]), the resulting phenotypes of the the two types of techniques can be combined to generate hypothesis on the target relevant to the phenotype. In fact, there are a number of very promising candidate drug targets that have been discovered for new cancer therapeutics by using genetic interaction methods [71, 72].

Computational approaches

Purely computational approaches, especially those relying only on ligand structure information are less explored. Nevertheless, they have recently been shown to powerfully predict previously unknown targets for drugs, with the goal of drug repositioning and explaining off-target effects. The most well-known ligand-based method is the Similarity Ensemble Approach (SEA) [73] that has been successfully employed to predict new molecular targets for known drugs, chemical hits and probes [74–76]. However, although the method is implemented in a web server (<http://sea.bkslab.org>), the results of the method cannot be downloaded from their web site. Besides, as SEA treats multiple compounds as a single set,

the predicted targets of a set of query compounds are displayed collectively. As a result, the discrimination of the predicted targets that correspond to a single compound is a difficult task. Therefore, this method cannot be practically applied for large-scale compounds.

Hence, faster and better methods of target prediction are required to accelerate the final phases of the chemical genetic process and provide valuable new tools for the dissection of various diseases.

1.6.3 Promiscuity

Another critical challenge in HTS procedure is the occurrence of “frequent hitters” or “promiscuous hits”, that is, many compounds are active in multiple assays via irrelevant mechanisms. Small molecules may interfere with the assay signal [77], act as oxidants [78] or chemically react with targets [77–79]. A common mechanism underlying this phenomenon is the formation of particles of 30-400nm diameter composed of small molecules [80] that inhibit the targets non-specifically. Several methods have been developed to predict the likely false positives produced via these mechanisms. For example, Baell et al. [81] described a number of sub-structural filters to selectively identify compounds that appear as frequent hitters in many biochemical HTS. Also, Gamo and co-authors [82] calculated an “inhibition frequency index” to exclude the promiscuous and non-specific compounds from the analysis. Jacob and collaborators, in turn, calculated HTS “promiscuous index” to filter 136 out of 2,999 compounds. However, the promiscuity threshold should not be set neither too high nor too low, and people should adjust the value according to the goal of chemical genetic study.

1.7 Computational systems biology in chemical genetics

Due to the steady growth of the amount of data emerging from chemical genetic studies, automatic methods for processing and handling the resulting assay data are necessary in order to extract the maximum amount of information from such screens. Computational systems biology with two major advantages – high speed and low cost – is making an increasing contribution and plays a crucial role in the analysis of chemical genetic assay data [83].

1.7.1 Systems biology

Systems biology is a scientific discipline that endeavors to quantify all of the molecular elements of a biological system, to assess their interactions and to integrate complex information such as that originated from development of high-throughput platforms for genomics, transcriptomics, proteomics and metabolomics, into graphic network models that serve as predictive hypotheses to explain emergent behaviors [84–87]. The goal of modern systems biology is to offer exciting new prospects for determining the causes of human diseases from the level of molecular pathways, regulatory networks, cells, tissues, organs and ultimately the whole organism and find possible cures [85].

Systems biology approaches in combination with computational methods can be extremely helpful for drug discovery and can aid in the optimization of medical treatment regimes for individual patients [84]. It has been predicted that computational systems biology and bioinformatics approaches could help cut the cost of creating a new drug in half and save 2 ~ 3 years of the development [88].

1.8 The goal and significance of the project

1.8.1 Goal

In this project, I aimed to systematically analyze the chemical screening data from public repositories, such as ChemBank [65], with the goal to develop analysis tools, like chemical hit identification, target prediction of hits and promiscuity filtering protocols for chemical genetic screens.

Ultimately, I applied these methods to answer two biological questions. The first one was whether novel relationships between biological activities can be extracted by comparing the profiles of compounds for different readouts of HTS. The concept is analogous to chemical profiling methods where relationships between compounds were established based on the similarity of their biological profiles. For example, chemical genetic experiments, where a chemical was screened in a panel of genetically different cell lines have been proved useful to reveal new mechanism of action of compounds as well as new gene functions [89]. As in profiling methods, I defined a fingerprint for every assay. This fingerprint was formed by the activity of a collection of chemicals on a biological assay. I hypothesized that

the bioactivities measured in two assays sharing selective hits are likely to be related. Later on, to understand better the novel associations between phenotypic assays I determined the molecular targets of specific hits associated with the biological process. Thus, by analyzing the fingerprints of each assay readout together with the drug targets of hits I aimed to find unexpected relationships between the biological activities measured in the assays.

As the second application of these methods, I proposed a novel computational approach to identify targets of specific hits and applied it to three phenotypic screens in ChemBank repository. By analyzing the targets involved in the phenotypes of high-throughput phenotypic screens, I aimed to find novel relationships between drug targets and biological processes.

The thesis has been organized in the following different sections and the workflow of the work is shown in Fig. 1.8.1.

(i) Collection of HTS data from the ChemBank database and analysis of the individual data sets of the repository.

(ii) Hit identification and data standardization of results of different screening assays.

(iii) Target prediction of tested compounds.

(iv) Description of a filtering protocol to detect promiscuous chemicals. Compounds showing positive activity in multiple assays were removed in order to discard unspecific activities.

(v) Computation of assay similarity and application of statistical methods to assess similarity between biological assays.

(vi) The associations between biological assays were integrated with predicted targets of hits in order to understand better the molecular basis of the similarity of assays.

(vii) Description of a method to systematically identify the targets involved in phenotypic screens. In the first place, I selected well-studied phenotypic assays. Then, I divided the tested compounds into specific hits and inactive compounds sets, and predicted human drug targets for these two sets. Using a statistical approach, I identified novel over-represented targets in the specific hits set of the assays.

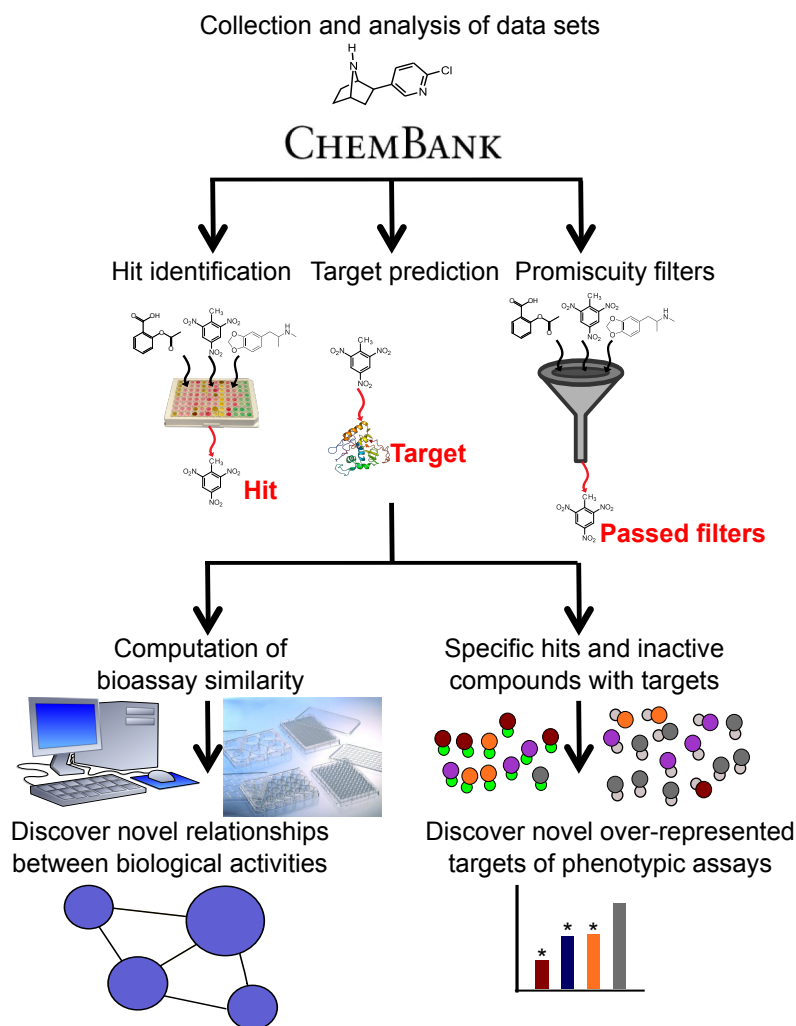


Figure 1.8.1: Workflow of the project. Asterisks denote the over-represented targets in the specific hits set.

1.8.2 Significance

By using the large chemical compound libraries, chemical screening is typically performed for early-stage drug development in academic institutions and pharmaceutical companies. The creation of a series of powerful tools is quite helpful and flexible for analyzing chemical genetic screening data, since these tools not only track and analyze chemical screening data, but they can also be used find novel biological connections. I believe that this work will facilitate sophisticated chemical genetic screening analysis in a wide variety of academic and industrial laboratories and will have a rapidly growing impact on life science research.

Chapter 2

Materials and Methods

2.1 ChemBank

ChemBank was created by the Broad Institute’s Chemical Biology Program and funded mainly by the National Cancer Institute’s Initiative for Chemical Genetics (ICG) [22]. This repository archives information on hundreds of thousands of small molecules as well as thousands of assays that have been performed at the ICG in collaboration with worldwide biomedical researchers [22].

2.1.1 ChemBank assay data structure

ChemBank [65] data was downloaded in May 2011 and comprised 193 projects with loaded screening plates, including 3,852 assays and 228,887 tested compounds. I also extracted information about assay names and description, project names, description and motivation of the projects. Three projects containing 18 assays were discarded because they lacked information about compound IDs, resulting in 190 projects. If a project comprises assays containing in the assay name an annotation of “raw” and “user”, such as the project of “Pseudomonas Cell Wall Synthesis”, I only kept the assay annotated as “user” as I observed that this type often reports the specific activity of the compounds. This step retained 3,617 assays. Then I combined the assays performed with the same annotated “experimental protocol” indicated by identical title and description, such as assay ID 1133.0005, ID 1133.0006 and ID 1133.0007 of the project “Glioblastoma Modulators”, into the same “assay type”. In total, 3,617 assays were grouped into 1,640 assay types. The analysis presented here was based on the assay type, which for

simplicity I named “assay”. I assigned the activity of a compound both on an assay level and a project level. A compound is active in a project when it is active in at least one of their assays.

I classified the assays into “cell-free”, “cell-based” or “microorganism” assays according to the assay description provided by ChemBank. If the assay was performed in a cell line (e.g. all the assays in the “Glioblastoma Modulators” project were done in U251 human glioma cells), this assay was classified as “cell-based”; if the assay was performed in a microorganism (e.g. the “SigB Inhibition” project that identified small-molecule inhibitors of *Listeria* SigB transcription factor to reduce Listeriosis was performed in *Vibrio* sp. S1063), this assay was classified as “microorganism”; the remaining biochemical or biophysical assays were classified as “cell-free”.

2.2 Chauvenet’s criterion

Chauvenet’s criterion (named for William Chauvenet [90]) provides a statistical approach to assess whether one experimental sample of a set of observations is likely to be a suspicious outlier and should be removed from the set.

To apply Chauvenet’s criterion, first the mean value and standard deviation using the set of “ n ” data points was calculated. Then the normal distribution function was used to calculate the probability of a given data point being the suspicious data point. Subsequently, this probability was multiplied by the number of data points (n). If the result was less than 0.5, the suspicious data point may have been eliminated, that is, a sample may have been rejected if the probability of obtaining the particular deviation from the mean was less than $1/(2n)$. After removing the outliers, the new mean value and standard deviation was calculated.

In the ChemBank hit identification method, in order to apply Chauvenet’s criterion, first all mock-treatment wells for each plate (number = n) were collected, next normal distribution function was used to calculate p-value for each mock-treatment, if $p - value * n < 0.5$, then this mock-treatment was discarded. The remaining mock-treatment wells that passed the outlier trimming were used to calculate mean and standard deviation.

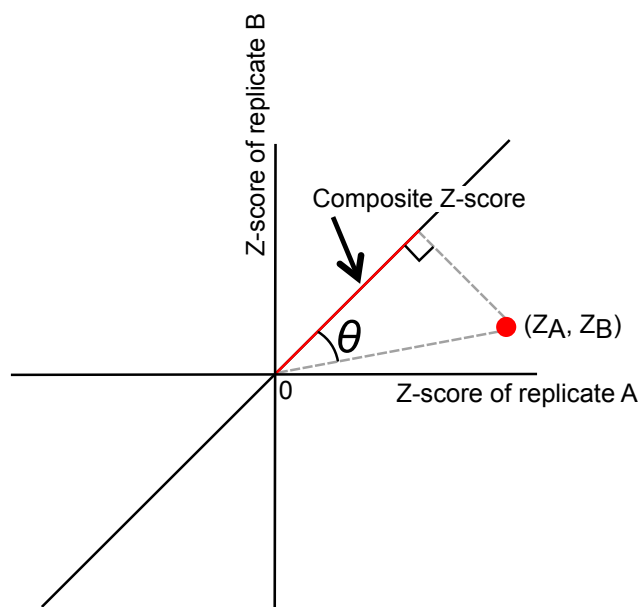


Figure 2.3.1: From assay duplicates to yield Composite Z-score.

2.3 Composite Z-score and Reproducibility

Due to the fact that a compound was normally tested in two, three even four replicates, after combining the information from replicates, Composite Z-score (CompositeZ) and Reproducibility for each compound were calculated to obtain the values of Z-score and Reproducibility in the ChemBank hit identification method. To illustrate the calculation of both parameters, I used two replicates as an example. CompositeZ was obtained by projecting vector (Z_A, Z_B) to “perfect reproducibility” (that is, equal Z-scores in both replicates, see the diagonal in Fig. 2.3.1) using cosine correlation (see equation 2.3.1), while Reproducibility is the cosine value (see equation 2.3.2).

$$\text{CompositeZ} = \sqrt{Z_A^2 + Z_B^2} \cdot \cos\theta \quad (2.3.1)$$

$$\text{Reproducibility} = \cos\theta = \frac{Z_A + Z_B}{\sqrt{Z_A^2 + Z_B^2} \cdot \sqrt{1 + 1}} \quad (2.3.2)$$

2.4 Median polish procedure

As the name implies, median polish is a technique using medians rather than

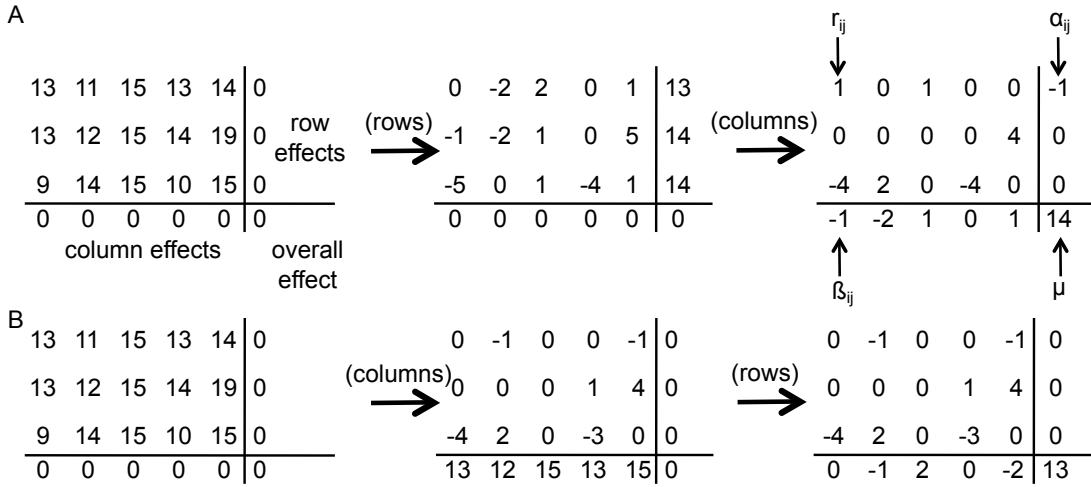


Figure 2.4.1: Illustration to get the residual activity. (A) Procedure operating on the rows first. (B) Procedure operating on the columns first.

arithmetic means for extracting/polishing row and column effects in a two-way data layout [91]. In every row (including the row of column effect), the median of all entries was taken to subtract from the row of all entries and next operated similarly on columns instead of rows, then returned to operate on rows, then columns,..., etc. The procedure was terminated when the two-way layout of residuals had zero medians in every row and column, and where the row and column effects each had median zero (see example in Fig. 2.4.1A).

Thus, if x_{ij} is the entry of row i and column j , if r_{ij} is the corresponding residual, μ is the overall effect, α_i is the i th row effect and β_j is the j th column effect, then

$$x_{ij} = \mu + \alpha_i + \beta_j + r_{ij}, \text{ with}$$

$$\text{median}_i(\alpha_i) = \text{median}_j(\beta_j) = \text{median}_i(r_{ij}) = \text{median}_j(r_{ij}) = 0 \quad (2.4.1)$$

Since the median of column and row effects will not influence r_{ij} , to simplify the algorithm, when the medians of both row and column are 0, the remaining value is the residual activity r_{ij} of each well. Alternatively, the polishing method can be operated on columns first rather than rows. This may lead to different, but qualitatively similar results (Fig. 2.4.1B). Median polish approach is completely analogous to that in 2-way ANalysis Of Variance (ANOVA), a procedure based

on fitting row, column and plate means. However, mean values are sensitive to outliers, and median polish algorithm possesses good robustness properties.

2.5 Median absolute deviation

Median absolute deviation (MAD) for each plate is a robust estimate of spread of the residual activity values (r_{ij}) (see equation 2.5.1).

$$MAD_p = \text{median} \{|r_{ijp} - \text{median}(r_{ijp})|\} \quad (2.5.1)$$

2.6 B-Score

The B-Score is calculated as follows (equation 2.6.1):

$$B - \text{Score} = \frac{r_{ijp}}{MAD_p} \quad (2.6.1)$$

2.7 Receiver operating characteristic space

The receiver operating characteristic (ROC) graph [92] is a widely used approach to evaluate the performance of a binary classification method. For example, it has been used as a quality metric in microarray transcriptomics [93, 94]. Given positive and negative controls, ROC graph offers a fast and intuitive understanding of dynamic ranges in data [95]. A ROC space plots sensitivity versus (1-specificity). Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP). TP: true positive, TN: true negative, FN: false negative, FP: false positive.

A completely random guess provides a point on a diagonal line (also called as no-discrimination line) that is from the left bottom to the top right corners. The diagonal divides the ROC space into two parts. Points above the diagonal show good results (better than random, such as point A and B points in Fig. 2.7.1), points below the line represent poor results (worse than random, such as point C in Fig. 2.7.1) and points on the diagonal show neutral results (as bad as random, such as point D in Fig. 2.7.1). The distance of a point to the diagonal of the ROC space can be used as a quality metric of the method. For example, a distance of $1/\sqrt{2}$ shows a perfect classification, yielding a point in the upper left corner (point A in Fig. 2.7.1) of the ROC space, representing 100% sensitivity (no false

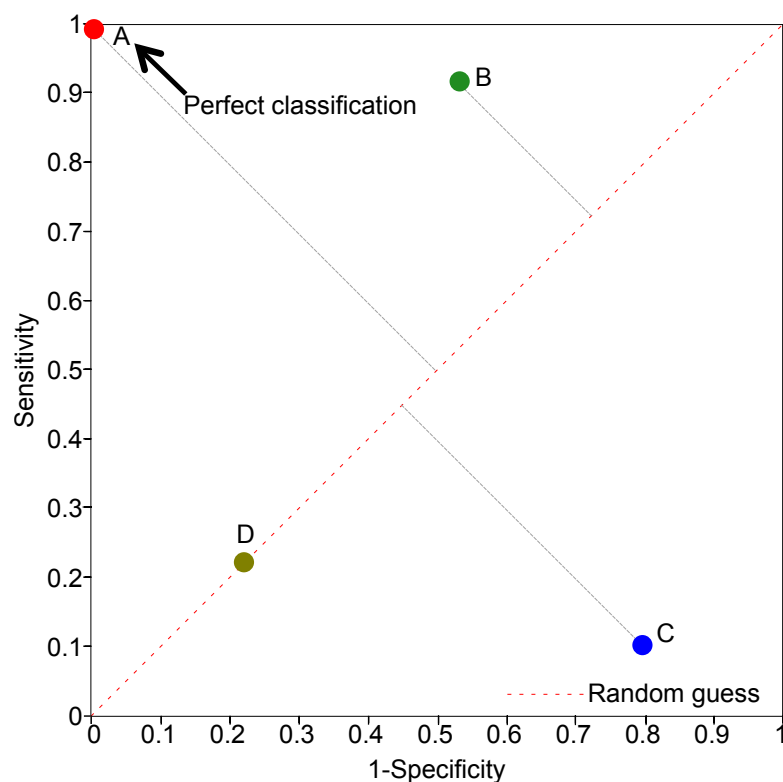


Figure 2.7.1: The ROC space and plots of the four prediction examples.

negatives) and 100% specificity (no false positives). However, a distance of 0 shows that the performance of the method is as bad as random chance.

One advantage of using ROC curves is that multiple thresholds for defining positives and the resulting trade-offs between sensitivity and specificity can easily be investigated by plotting multiple ROC curves. For that reason, I used ROC curves for the evaluation of hit retrieval in the chemical screens.

2.8 Target prediction in HitPick

HitPick is a novel web server for hit identification and target prediction of chemical screens [96].

To predict the protein targets of small molecules, HitPick uses a newly developed approach that combines two ligand-based methods based on two dimensional (2D) molecular fingerprints. The two methods are the 1-Nearest-Neighbor (1NN) similarity searching [97] and Laplacian-modified naïve Bayesian target models [98].

2.8.1 Database

The Search Tool for Interactions of Chemicals (STITCH) version 3 [99] that can be accessed through <http://stitch.embl.de>, is a database containing known and predicted interactions of chemical and proteins. STITCH includes interactions from 1,133 organisms for between 300,000 chemicals and 2.6 million proteins [99]. In this study, I restricted the target prediction to human proteins, as it is currently the species with the largest number of known drug targets, enabling thus more accurate predictions. For human species, I selected targets with at least three known ligands due to the later model validation. In total, there are 1,375 targets interacting with 99,572 unique compounds indicated by SMILES strings. STITCH compounds may have more than one target, so each ligand - target pair was considered during the model training. In total, a set of 145,549 human chemical - protein physical interactions extracted from the STITCH database.

2.8.2 Fingerprints

Two dimensional (2D) fingerprint is a binary vector denoting the presence or absence (1 or 0) in a molecule of some fragment substructures [100]. The fingerprint designs can vary dramatically in length ranging from ~ 100 to millions of bits [101]. If the fingerprints of two molecules have many bits in common, then they are considered to be similar in structure [102,103] (The evaluation of similarity is described in section 2.8.3). Since the circular fingerprints are well-established for building models to predict the biological activities of small molecules, the in-house 2D circular fingerprints were generated for STITCH compounds based on the Morgan algorithm [104] for radius up to 3 bonds and maximum length bit string (9,192) using RDKit (<http://rdkit.org>).

2.8.3 1NN similarity searching

Recently, chemical similarity searching methodology based on 2D fingerprints has been shown the simplest but efficient tool for ligand-based virtual screening [105–107] due to following characteristics: (i) only structural information (provided by chemical fingerprints) of the compound is required to formulate the query; (ii) the implementation of chemical similarity methods are computationally inexpensive (only a few seconds), (iii) due to the generally valid “Similar Property Principle” - Structurally similar molecules tend to exhibit similar biological activ-

ities [108–110], computational chemists frequently retrieve from chemical library compounds that are similar to the active lead compound; (iv) chemical similarity searching method can also contribute the design of the diverse chemical libraries which require the compounds should be as dissimilar as each other [111].

K Nearest Neighbor (KNN) search is a similarity method that searches k most similar chemicals from the dataset. 1NN similarity searching is a particular case of KNN with $k = 1$ in which the similarity between reference set R with N molecules and a query compound x is defined as similarity between x and its nearest neighbor (measured by similarity) in R [97].

Many approaches are available to measure the chemical similarity [112–115]. So far the most commonly used method for the quantitative comparison of binary molecular fingerprints is the Tanimoto coefficient (Tc). The Tc between the fingerprints of two molecules is calculated using the equation (2.8.1).

$$Tc = \frac{c}{a + b - c} \quad (2.8.1)$$

where c is the number of bits shared between the two fingerprints of molecules, and a is the number of bits that are set on in the first fingerprint, and b is the number of bits that are set on in the second fingerprint. The higher Tc value is, the more similar between the compared fingerprints are.

2.8.4 Laplacian-modified naïve Bayesian target models

Bayesian theory was proposed several decades ago to calculate posterior probabilities. It assumes that on a given class, the effect of the presence or absence of a feature will not be influenced by the presence or absence of other features, namely these features are totally independent. Bayesian theory is called as “Bayesian classifier” as generally it is used to calculate the probability whether a given sample can be classified into a particular class.

Bayesian based statistics in combination of chemical descriptors have been proposed recently and have been shown to be a powerful classification tool in recent studies to predict protein targets for compounds [98, 116–120]. Unfortunately, in chemoinformatic data analysis, Bayesian theory is not realistic because the descriptors/features (interpreted by fingerprint bits) of the compounds are not independent. For simplification purposes, when it is supposed that all the

molecular features are independent, Bayesian theory is “naïve”, and the resulting models are called as naïve Bayesian target models.

According to the Bayesian statistical analysis, naïve Bayesian model first calculates the conditional probability of a given sample belongs to a particular class (equation 2.8.2).

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (2.8.2)$$

where B is the event for compound composed of a group of n features; and A is the target class for which a compound is a ligand (active). $P(A)$ and $P(B)$ are the probabilities of events A and B occurring; and $P(A|B)$ is, given events A and B , the probability of A occurring under the condition of given B .

Since each compound is characterized by the 9,192 chemical fingerprint bits, $P(B)$ can be calculated by multiplying all the probability of each individual fingerprint bit according to naïve Bayesian theory (equation 2.8.3).

$$P(B) = \prod_{i=1}^{9192} P(f_i) \quad (2.8.3)$$

where f_i is the i th feature of compound B . f_i is represented by binary value 0 or 1 (0 inactive, 1 active).

Some fingerprint bits are 0. To avoid the case of 0 probability values, Laplacian corrected estimator [98, 116, 121] is applied to calculate the probability of target A occurring by given feature f_i (equation 2.8.4).

$$P(A|f_i) = \frac{A_{f_i} + 1}{T_{f_i} \cdot \frac{A}{T} + 1} \quad (2.8.4)$$

The above formulation was adapted from Xia et al. [116] and Nidhl et al. [98], where A_{f_i} is the number of active fingerprint bits among compounds binding to target A , T_{f_i} is the active fingerprint bits among total compounds. A/T is the number of compounds binding to target divided by number of total compounds for all the targets.

Then, given the presence of 9,192 features, the probability of the compound being active toward the target should be given by equation (2.8.5).

$$P(\text{active}) = \prod_{i=1}^{9192} P(A|f_i) \quad (2.8.5)$$

In order to avoid potential numerical problems of the resulting small values ($\ll 1$) and to interpret better the results, the above equation is typically implemented by logarithms to yield a combined value, it is normally called score S (see equation 2.8.6).

$$S = \log P(\text{active}) = \sum_{i=1}^{9192} \log P(A|f_i) \quad (2.8.6)$$

The model yields a score for any test compound by equation (2.8.7).

$$S_{test} = \sum_{i=1}^{9192} x_i \cdot \log P(A|f_i) \quad (2.8.7)$$

where S_{test} is the generated score for the compound, and x_i is the bit value of the fingerprint of the compound.

A model was built for each target based on the fingerprints of small molecules in the STITCH dataset where the activity of compounds on multiple targets (1,375) is stored. Such models are called as "multiple-class Laplacian-modified naïve Bayesian target models" [98].

2.8.5 Combination of 1NN similarity searching and Laplacian-modified naïve Bayesian target models

In this work, 1NN similarity searching method was applied first to search for the most similar compound (measured by Tc similarity) of the query compound. Then, the Laplacian-modified naïve Bayesian target models generated a score for all known targets of the most similar compound, resulting in a list of ranked target predictions. Each score of the target is indicative of the likelihood of the prediction. That is, if the target A 's Bayesian score is 90, and target B 's score is 80, then it means the target A is more likely than target B to be the target of the query compound. The target with the highest score is the most likely target for the test compound. Similarly, the second highest score of a target is supposed to be the second most probable target, etc.

Before the models were applied to predict the targets for compounds, the validity of the models was evaluated. For benchmarking, I randomly assigned 85% of the known ligands of each target to the training set and the remaining 15% to the validation set (Fig. 2.8.1). Due to the separation of the ligands, I required

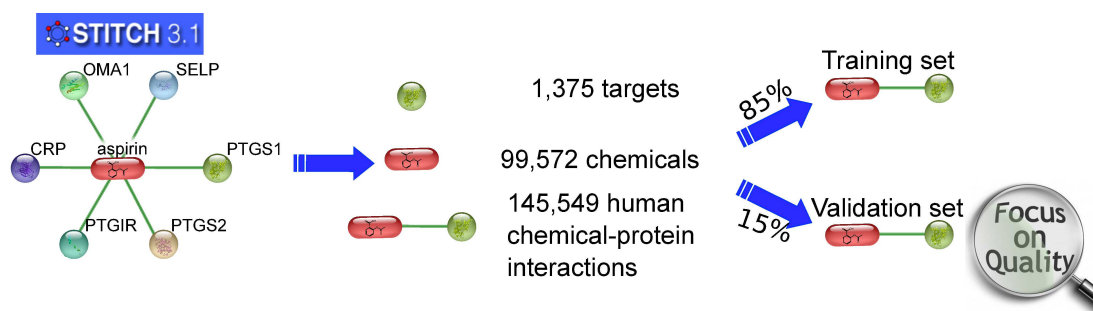


Figure 2.8.1: Benchmark of target prediction.

that every target should have at least three ligands. In total, the validation set contained 22,868 positive and 20,779,507 negative compound-target relationships, respectively. For each validation compound, the model generated a score for all possible targets through each Laplacian-modified naïve Bayesian model of each target class.

Two types multiple-class models were built. The first type of model contained 85% of the ligands as the training set, so that the remaining 15% of the ligands could be used as the model validation. The second one contained 100% the ligands for their application to predict any test compound.

HitPick target prediction method reports all targets of the most similar database compound along with the precision. The reported precision refers to the probability of a target being true regardless of whether the target prediction for higher ranked targets are true or false (see Chapter 4 HitPick).

The fingerprint creation for the STITCH compounds, building and application of Bayesian target-specific fingerprint models were implemented in a KNIME (<http://www.knime.org>) workflow making use of the chemoinformatic functionality provided by KNIME itself as well as by RDKit.

2.8.6 MaxMinAlgorithm

MaxMinAlgorithm [122] is provided by RDKit (<http://www.rdkit.org/>), which follows a simple yet efficient approach. It is initialized with a random seed compound and subsequently adds compounds iteratively from outside the subset that are maximally dissimilar to the current subset until the desired number of compounds is selected.

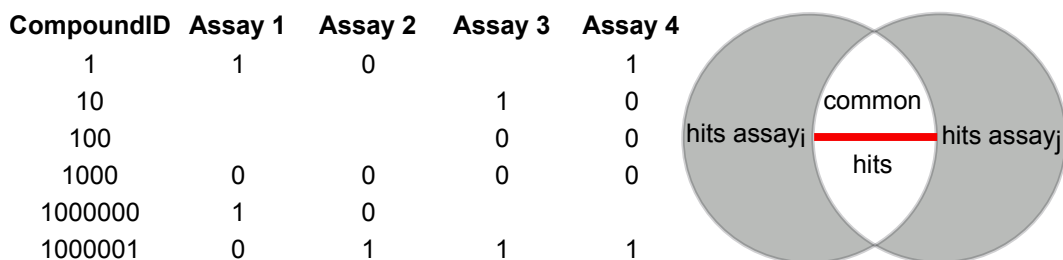


Figure 2.9.1: Hit comparison between different assays.

2.9 Calculation of hit similarity

I represented the hit profile of an assay using a binary fingerprint of the chemical activity of compounds (“1” indicates that the compound is a hit; “0” indicates that compound has been tested, but inactive; missing value signifies that the compound has not been tested) (Fig. 2.9.1).

To calculate the hit profile similarity based on the shared and non-shared hits between two assays (Fig. 2.9.1), I used the continuous Tc equation (equation 2.9.1),

$$Tc = \frac{\sum_{i=1}^n (w_i x_i)(w_i y_i)}{\sum_{i=1}^n (w_i x_i)^2 + \sum_{i=1}^n (w_i y_i)^2 - \sum_{i=1}^n (w_i x_i)(w_i y_i)} \quad (2.9.1)$$

where n is the total number of compounds tested in both assays, i iterates over all compounds, w_i is the promiscuity (ratio of the number of assays where the compound is active and the number of assays where the compound was tested) of the compound at ChemBank database level including 1,640 assays. x_i , y_i are the activity values (1 or 0) of the compound.

2.10 Similarity of assay project by applying EXCERBT

EXtraction of Classified Entities and Relations from Biomedical Texts (EXCERBT) [123] is a free-to-use biomedical text-mining system based on all available abstracts from PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>), full-text articles from PubMed Central (<http://www.ncbi.nlm.nih.gov/pmc/>) and articles in OMIM (<http://www.ncbi.nlm.nih.gov/omim/>). As EXCERBT is case-sensitive I searched for different combinations of lower and upper case of keywords avail-

able in EXCERBT. For example, for the keyword “Wnt”, EXCERBT provides different forms such as “wnt”, “WNT” and “Wnt”. For “histone deacetylase”, the terms “HDAC”, “HISTONE DEACETYLASE”, “Histone deacetylase”, and “Histone Deacetylase” are available. Furthermore, the search result depends on the order of two keywords. For instance, there are four evidences linking “HDAC” and “Wnt” and only two for the search of “Wnt” and “HDAC”. Thus, I searched every combination of keywords of two projects and selected the search that retrieved the highest number of occurrences.

2.11 Promiscuity filters

In order to increase computational efficiency, I applied F1 to keep compounds from the initial ChemBank dataset showing activity in more than one project. The removal of the compounds active in only one project or inactive in all the projects does not have an effect on the hit similarity (continuous Tc) between assays (Fig. 2.11.1). Then, I applied two additional filters to keep selective compounds at project level (F2) and assay level (F3), respectively. F3 was applied to projects with at least nine assays, which was determined by averaging the number of assays per project in the ChemBank screening repository.

2.12 Identification of significantly over-represented enriched targets

All the compounds tested in both experimental and control assays are separated into positive and negative groups, respectively. A modification of B-Score_A method that mainly uses the median polish procedure to remove the row/column biases in a plate [61] (see Chapter 3 Chemical Hit Identification, with p-value < 0.05 as the threshold), was applied to identify the hits of both “experiment” and “control” assays. I define “experimental” assay as the assay that measures the intended biological activity of a project and a “control” assay as an assay that controls for the specificity of the biological signal in the “experiment” assay. The positive group contained specific hits, i.e. compounds are only active in experimental assay of the project; the negative group contained all the remaining inactive compounds. After predicting the targets for both groups by applying Hit-Pick target prediction method [96], hypergeometric tests were used to assess the

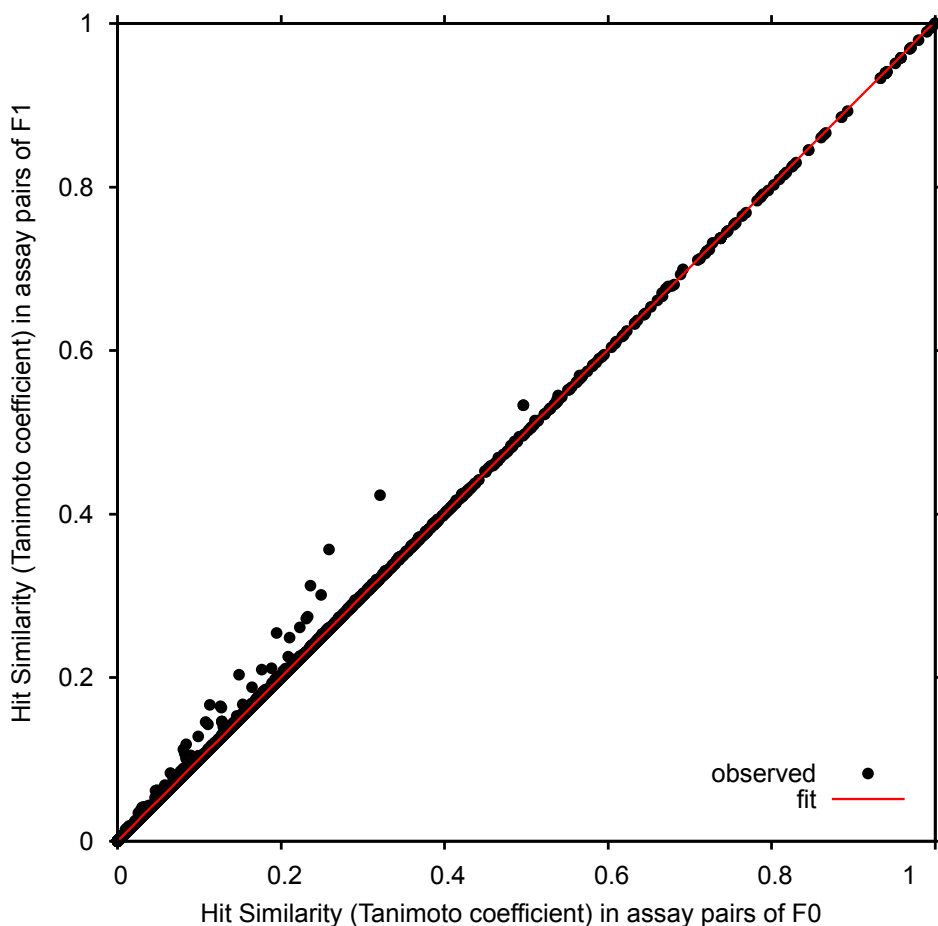


Figure 2.11.1: Chemical hit similarity comparison of the assay pairs after application of filter F0 and F1. I randomly chose 10,000 assay pairs and compared the hit similarity (Tc) after the filter F0 and F1. F0 contains all the initial compounds; F1 includes the compounds that are active in more than one project.

statistical significance of over-representation of certain predicted target of specific hits compared to all distinct proteins of the compounds with predictions.

The data can be represented in a 2×2 contingency table (Table 2.12.1). q is the number of compounds with targeting A in positive group; $(k - q)$ is the number of compounds targeting A in negative group; $m - q$ is the number of compounds not targeting A in positive group; and $n - (k - q)$ is the number of compounds not targeting A in negative group. The over-represented calculations were done in R [124] by command: `phyper(q - 1, m, n, k, lower.tail = FALSE, log.p = FALSE)`. All reported p-values were adjusted with a false discovery rate (FDR) [125] correction.

Table 2.12.1: 2 X 2 contingency table of hypergeometric test

	Positive group	Negative group	Total
Number of compounds targeting A	q	$k - q$	k
Number of compounds not targeting A	$m - q$	$n - (k - q)$	$m + n - k$
Total	m	n	$m + n$

Chapter 3

Chemical Hit Identification

In a chemical screening assay, those tested compounds with activity levels, which are reactive in the experiment, are termed “hits”. Chemical hit identification process determines which activity values of the compounds differ meaningfully from those of the negative controls in the assay [95]. This process is the starting point for discovering and developing successful new biologically active compounds. Typically, the experimental researchers organize the screening results in a sorted list according to compounds’ activity and choose a threshold value above which the compounds are considered as hits [64]. However, this comes at a price that the systematic errors in HTS data sets will not be detected. Systematic errors can be caused by inconsistent plate replication, agent evaporation, variation in incubation time, pipette malfunction and temperature differences [54, 126, 127] that lead to plate-to-plate variance, and also can be due to the positional effects of wells within plates that cause the variance within the plate [128]. For example, throughout the entire screening campaign of more than 1,000 plates, Brideau et al. [128] detected that activity values of the compounds located in the row A were on average 14% lower than those located in the row P of wells. To compensate for systematic effects, to remove errors from HTS data sets and improve the quality of raw screening data, visualization techniques [129] and rapid data-mining procedures [130] have been developed.

In this chapter I evaluate the performance of existing and novel computational chemical hit identification methods that account for systematic experimental errors on the chemical screens in ChemBank repository.

In ChemBank repository, 228,887 compounds were tested in 190 diverse projects,

3. CHEMICAL HIT IDENTIFICATION

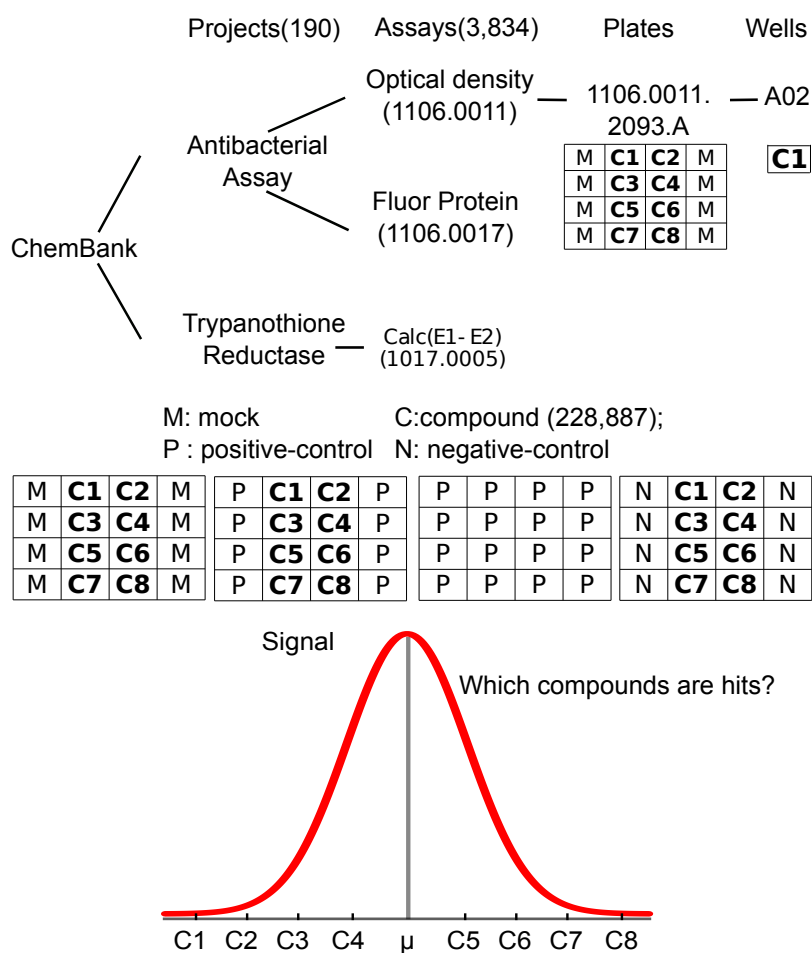


Figure 3.0.1: ChemBank assay structure. P: positive-control well; N: negative-control well; C: compound-treatment well; M: Mock-treatment well.

consisting of 3,834 assays. For convenience, all the controls are at an edge in an assay plate (see Fig. 3.0.1). Given such assay data structure of ChemBank and the resulting signals of compounds, I asked which compounds can be selected as “hits” or “screening positives”? Such particular and big amount of data requires urgent development of automatically statistical scoring procedures that can justly and better extract the hits from chemical genetic screens.

The results presented in this chapter have been submitted in X. Liu and M. Campillos, ‘Chemical screening assay pairs that share selective hits are biologically related’.

3.1 Results and discussion

The availability of raw bioassay values in forms of cell counts, absorbance, etc. as well as the information about the activity of the positive and negative controls in the assays in ChemBank repository make it possible the application and comparison of different hit identification methods. To identify the chemical hits in the ChemBank data set, I applied three published methods, namely, the ChemBank [65], the B-Score [61], the Well-Correction [66] methods and five modifications of them to adapt the methods to the ChemBank data structure that I summarize as follows:

3.1.1 The ChemBank method to identify hits

I named it as ChemBank method, because the authors introduced this approach when ChemBank database was published in 2008 to identify hits aiming to normalize the activity in the assay based on mock signals.

The calculation in the ChemBank method is based on Z-score to adjust for plate-to-plate changes in the assay noise or variability of sample values (Fig. 3.1.1).

Considering plate-to-plate differences in signal [128, 131], the median of raw value of mock-treatment wells on a given plate of the assay was firstly calculated, afterwards, the median was subtracted from each mock-treatment value on the same plate, providing a zero-centered distribution of mock-treatment wells for each plate in one assay [65]. Next, all the values from all mock-treatment wells in the assay were collected together, and those well values that fail to pass Chauvenet's criterion [90] (the calculation is described in Chapter 2 Materials and Methods) were eliminated to protect against the edge effects and other systematic artifacts, which are known technical problem in the chemical assays [132, 133]. The mock-treatment values that passed the outlier trimming were used later to normalize the compound-treatment wells. First, the mean value of remaining mock-treatment wells of each plate was calculated and it was subtracted from compound-treatment wells on the same plate to obtain background-subtracted values (*BSubValue*). Then, standard deviation (*stdev*) of mock-treatment wells of the assay was calculated, and a dimensionless Z-score of each compound-treatment well of the assay was obtained by $BSubValue/(2 * stdev)$. Since most of the chemical screening assays deposited in ChemBank were performed in two, three or even four replicates (A, B, C and/or D), these replicates were finally combined into a Composite

3. CHEMICAL HIT IDENTIFICATION

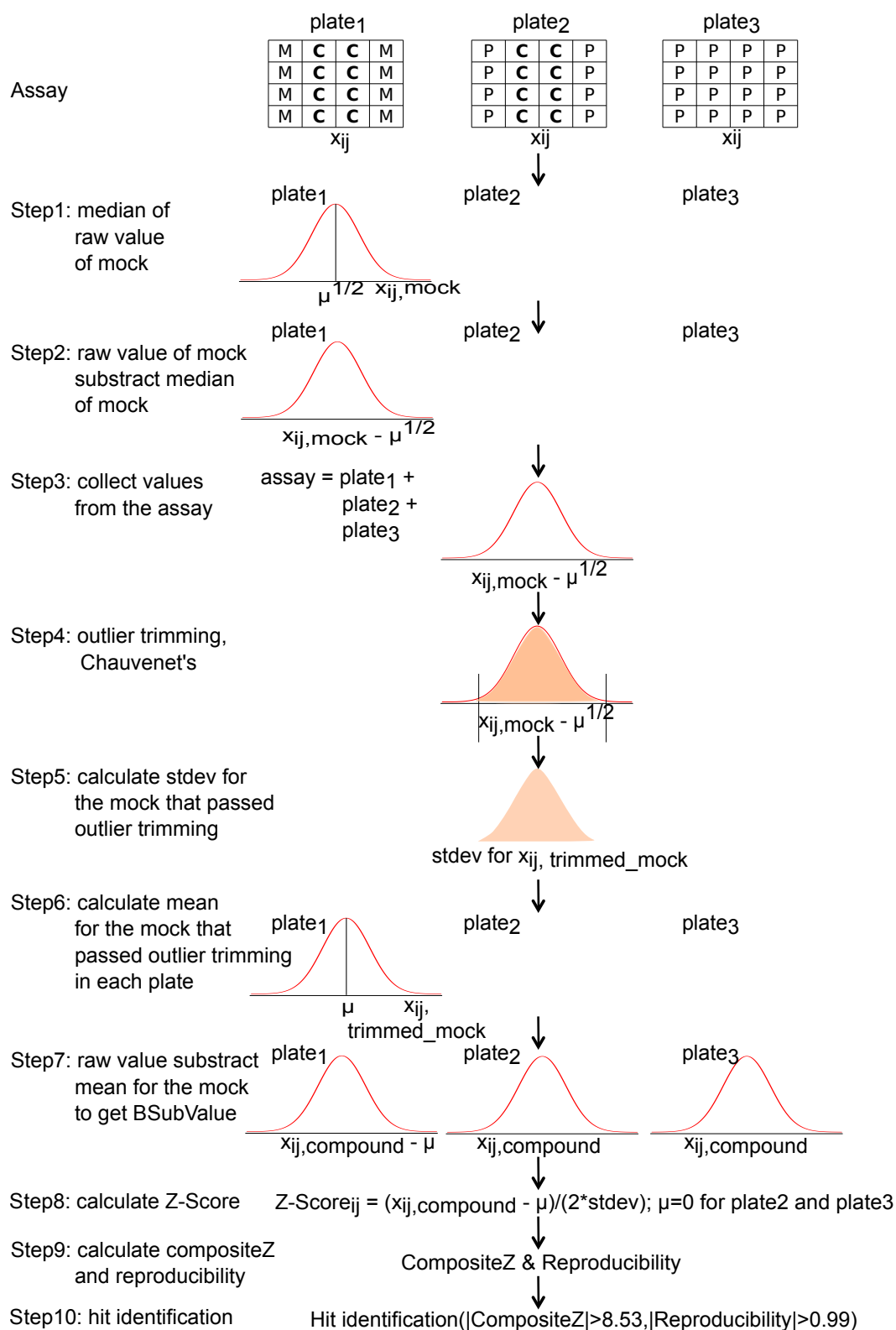


Figure 3.1.1: Scheme of the ChemBank method to identify hits. M: mock-treatment; C: compound-treatment; P: positive-control.

Z-score (CompositeZ) by projecting the vector (Z_A , Z_B , Z_C , and/or Z_D ,) to “perfect reproducibility” (that is, in all the replicates, each compound-treatment well had equal Z-scores) [65]. The calculated CompositeZ was the overall measure of whether a compound scored as active in an assay. If the CompositeZ and Reproducibility (the calculation of these two parameters are described in Chapter 2 Materials and Methods) of each compound followed the objective criteria: $|\text{CompositeZ}| > 8.53$ AND $|\text{Reproducibility}| > 0.99$, then this compound was referred as “hit”.

In the ChemBank method, the compound signals are normalized in the assay based on mock-treatment signals. However, the obvious drawback of this method is that if the plate does not contain any mock-treatment wells (see Fig. 3.1.1 plate2 and plate3), then the plate values are not normalized. For this reason, this method does not fully correct for the systematic effects within the plates (The performance of the ChemBank method in distinguishing positive and negative controls is described in section 3.1.6).

3.1.2 The B-Score method to identify hits

B-Score (for “Better” score) normalization may be applied to the row/column biases within a single plate via a procedure known as 2-way “median polish” (see Chapter 2 Materials and Methods). As the name implies, the procedure is based on the use of medians rather than means. Medians hold the advantage that they are not influenced by the statistical outliers, thus, the median of data which contains a few “wild” values (known as outliers) is almost the same as the same data without them [131]. This confers relative robustness to outliers of the B-Score method. In addition, the B-Score method has two other advantages [134]: (i) it is nonparametric, (ii) it minimizes measurement bias due to positional effects.

To account for row and column effects of the plate, a two-way median polish was first applied to compute residual activity (Fig. 3.1.2). To standardize for plate-to-plate variability, the resulting residuals within each plate were then divided by their median absolute deviation (MAD) (the calculation is described in Chapter 2 Materials and Methods) to calculate the B-Score. Finally, all the plates of the same assay were collected together to obtain p-value by normal distribution. Hits were determined using MAD or p-value statistics, i.e. (a) compounds with a residual larger than $2 \times \text{MAD}$ (“2MAD”), (b) $p < 0.01$, (c) $p < 0.05$, were defined

3. CHEMICAL HIT IDENTIFICATION

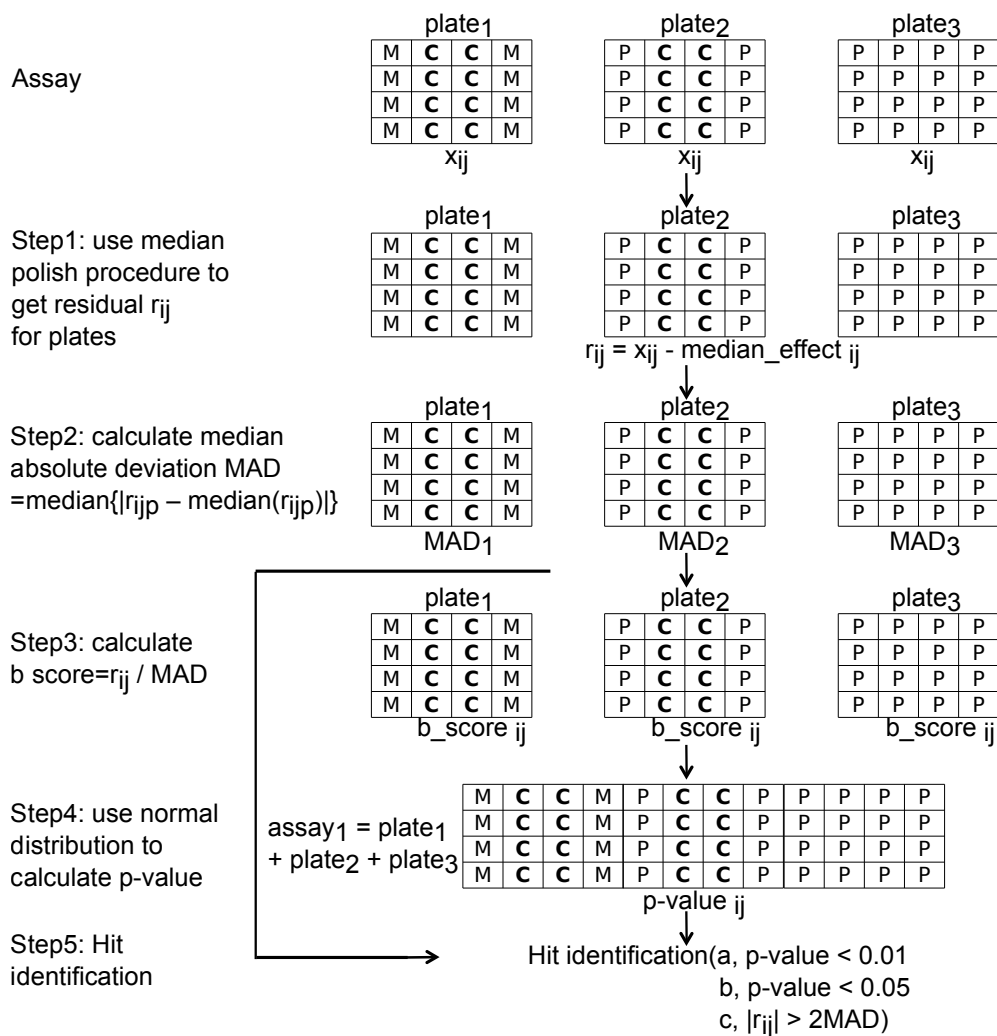


Figure 3.1.2: Scheme of the B-Score method to identify hits. M: mock-treatment; C: compound-treatment; P: positive-control.

as hits.

Considering the two systematic variations, the B-Score method uses median polish procedure to remove both column and row effects. However, again when the plate exclusively includes positive-control wells whose signals are normally higher than other plates, the signals will be cancelled during polishing step (The performance of B-Score method see section 3.1.6). To overcome this problem, I developed a modification of B-Score method called as B-Score_A method that treats such plates specially to obtain the residual activity.

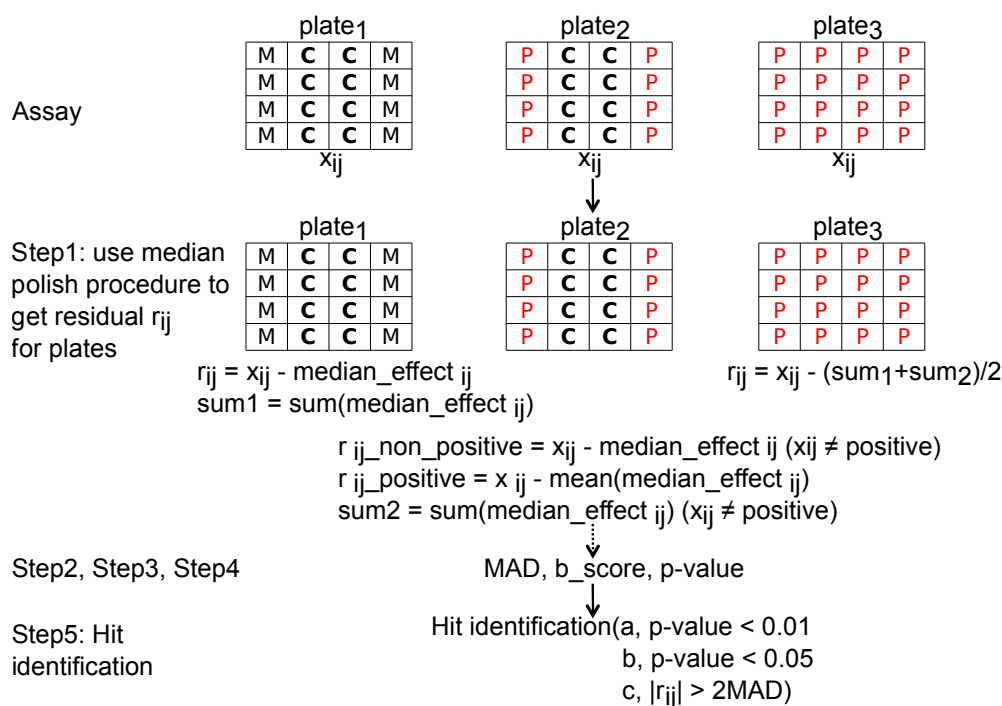


Figure 3.1.3: Scheme of the B-Score_A method to identify hits. M: mock-treatment; C: compound-treatment; P: positive-control.

3.1.3 The B-Score_A method to identify hits

As B-Score required ideally the controls to be located randomly among the wells of each plate, or at most localized in the first and last columns, I modified the method to adapt it to the ChemBank dataset structure where some plates only contained positive-control wells (e.g. plate ID 1031.0004.Pos.A and B). For this, positive controls were not involved in the median polish procedure and their residual activity was computed by subtracting the mean median effects of non-positive controls from their raw values. The next steps, including hit detection thresholds, were the same as in the B-Score method and I named this modification B-Score_A (see Fig. 3.1.3) (The performance of B-Score_A method see section 3.1.6).

3.1.4 The Well-Correction method to identify hits

The Well-Correction method rectifies the distribution of assay measurements by normalizing data within each well across all assay plates [66,135]. Firstly it nor-

malized the all plate signals using Z-score standardization so that each plate had a mean of zero and standard deviation of one. Once the data were plate-normalized, the linear regression ($y = ax + b$, where x indicates the plate number, and y represents the well value after plate normalization) was applied for each well. The obtained trend was then subtracted from the original value of each well, bringing the mean of this well across plates to zero. Afterwards, the Z-score normalization of each well was carried out, and p-value was computed using normal distribution in the assay. In the end, threshold of $p < 0.01$ or 0.05 was applied to capture the hits (Fig. 3.1.4) (The performance of Well-Correction method see section 3.1.6).

3.1.5 Modifications of the above four methods

The Well-Correction method analyzed well values measured across all assay plates. This method required that wells across plates should not systematically contain compound samples belonging to the same family. In the ChemBank dataset, many wells across different plates contained high number of positive controls (e.g. well A24 of assay ID 1017.0030) and therefore, the Well-Correction method could not be applied directly. To correct for this, I discarded wells with higher number of positive controls (i.e. number of positive controls \geq number of non-positive controls). In order to keep all the methods comparable, I applied this modification for the above four methods (Fig. 3.1.5) (The performance of these four methods see section 3.1.6).

If the assay contains replicates of compounds, I required all replicates to be identified as hits to consider them as hits. All the eight methods described above that rely on the application of existing methodology are statistical and can be applicable to any HTS assay.

3.1.6 Performance comparison among the eight methods

I determined the performance of the eight hit identification methods using the receiver operating characteristic (ROC) graph [92] (see Chapter 2 Materials and Methods) and the positive and negative controls (including mock treatments) of the assays were used as a benchmark set (Fig. 3.1.6). The total number of positives is 96 and the number of negatives is 7,590,042 and 7,620,521 for non-modified and modified versions of methods, respectively. The modification of the B-Score_A method with two different thresholds, namely, “2MAD” and “ $p < 0.05$ ”, showed the

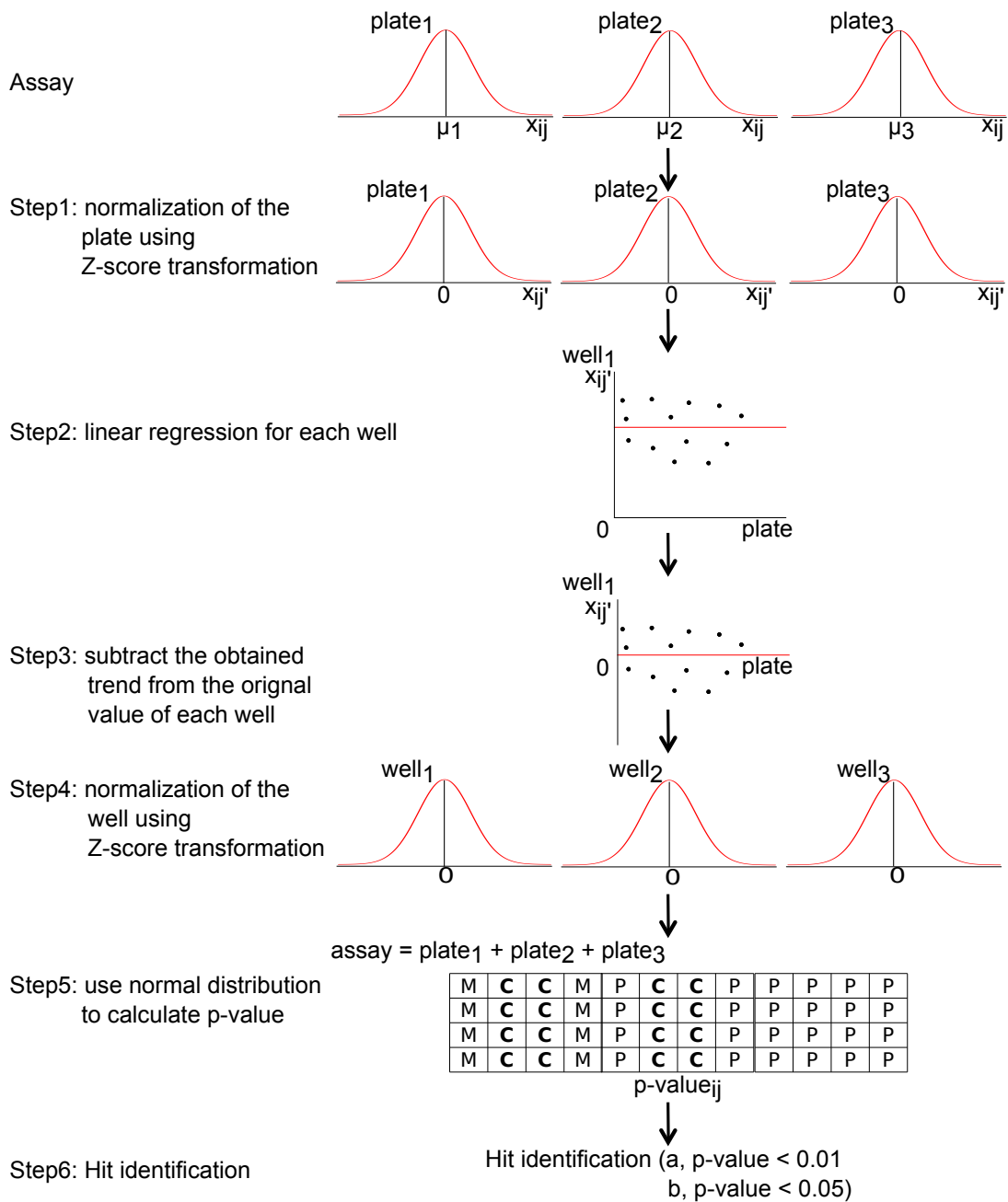


Figure 3.1.4: Scheme of the Well-Correction method to identify hits. M: mock-treatment; C: compound-treatment; P: positive-control.

3. CHEMICAL HIT IDENTIFICATION

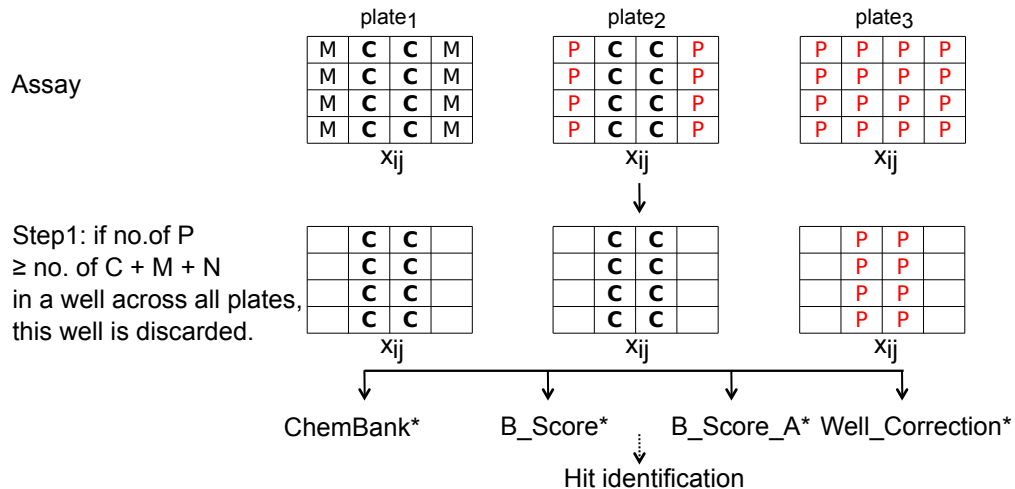


Figure 3.1.5: Modification of the four methods to identify hits. M: mock-treatment; C: compound-treatment; P: positive-control; N: negative-control.

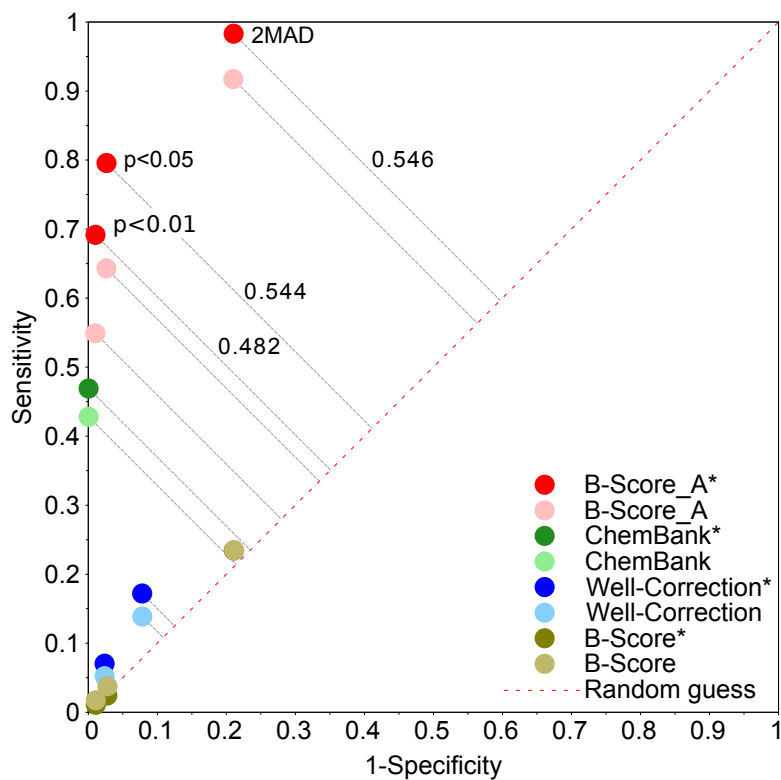


Figure 3.1.6: ROC plot for eight hit identification methods.

best performances. Due to its higher specificity (97.4%) with 79.6% of sensitivity, the latter one was determined to identify hits for chemical screens.

3.2 Conclusions

HTS is a large-scale approach that screens many thousands of small molecules in order to identify potential lead and drug candidates rapidly and precisely. All HTS campaigns are prone to systematic errors due to plate-to-plate and within-plate variance, resulting in decreasing the validity of results by either over- or under-estimating true values [61]. Normalization of raw data based on the two variations helps to remove systematic errors, making all the measurements comparable.

The availability of the raw bioassay data in the ChemBank dataset allowed me to assure a high quality in the detection of hits by testing and determining, among eight different hit identification methods, the method that best discriminated between the positive and negative controls within the assays. The modification of the B-Score_A method showed the best performance by achieving a sensitivity of 79.6% and a specificity of 97.4%.

Chapter 4

HitPick

Chemical biology experiments are increasingly used to search for chemical modulators of biological processes in cell-based and even whole-organism assays as illustrated by the thousands of phenotypic screens stored in public repositories [65, 136]. In these assays, the identification of the molecular targets of hits is essential to understand the molecular basis of the chemical activities in the bioassay. Recently, drug target prediction methods have been applied to the hits of cells [137] and zebrafish [76] phenotypic screens showing that computational approaches are suitable tools that facilitate the interpretation of the biological activity of chemicals.

Although diverse *in silico* methods have been proposed to identify hits [61, 66] and predict targets for chemicals (reviewed in Ref. [138]), only few of them are available as easy-to-use online tools [73, 139]. To overcome this situation and assist in the analysis and interpretation of chemical phenotypic screens, I introduce HitPick, the first web server for hit identification and target prediction of chemical screens. HitPick provides the functionality to detect bioassay hits using the B-Score method [61] (the calculation is described in Chapter 3 Chemical Hit Identification) and predicts targets of a chemical of interest using a newly developed approach combining 1-nearest-neighbor (1NN) similarity searching [97] and a machine-learning method [98] (see Chapter 2 Materials and Methods).

The results presented in this chapter have been published in X. Liu*, I. Vogt*, T. Haque, and M. Campillos. HitPick: a web server for hit identification and target prediction of chemical screenings. **Bioinformatics** 2013, 29, 1910–1912. (* These two authors contributed equally. Ref. [96]).

Table 4.1.1: Performance of the three target prediction methods

	HitPick	1NN	Bayesian
Precision (%)	92.11	84.72	80.03
Sensitivity (%)	60.94	NA	52.95
Specificity (%)	99.99	NA	99.98

Note: NA, “not available”.

Table 4.1.2: Comparison of the performance of HitPick and SEA target prediction methods.

	HitPick	SEA
Precision (%)	84.8	82.8
Sensitivity (%)	56.9	55.3
Specificity (%)	99.99	99.99

4.1 Results

4.1.1 Performance of target prediction

For the implementation of the target prediction approach, I used a set of 145,549 human chemical-protein physical interactions extracted from the STITCH 3 database [99] (preparation of the data is mentioned in Chapter 2 Materials and Methods). I assessed the performance of the method in HitPick using as validation set 15% of all ligands that were not part of the training set. When evaluating the highest scoring target prediction for each compound, HitPick achieved a sensitivity of 60.94% (with 66.16% being the maximum possible sensitivity), a specificity of 99.99% and a precision of 92.11%, an improvement over naïve Bayesian models and 1NN similarity searching (see Table 4.1.1).

I used the same validation data and compared the performance of HitPick to the Similarity Ensemble Approach (SEA) [73], a well-known target fishing application that relates proteins based on the chemical similarity of their ligands. For each method, I selected the best-predicted target (i.e. highest precision for HitPick and lowest E-value lowest for SEA) for each validation compound and then calculated precision, sensitivity and specificity over all predictions. The performance of HitPick is comparable with the target prediction quality achieved by the SEA (see Table 4.1.2).

I also evaluated the performance of the HitPick target prediction method at dif-

Table 4.1.3: Precision (%) for the first five predicted targets in relation to the Tc similarity of a validation compound to the most similar molecule in the training set

Ranked prediction	[0.2~0.3)	[0.3~0.4)	[0.4~0.5)	[0.5~0.6)	[0.6~0.7)	[0.7~0.8)	[0.8~0.9)	[0.9~1.0)	1.0
1st	15.7	26.4	53.3	77.0	89.8	94.8	96.7	97.7	97.3
2nd	15.4	14.5	43.5	54.3	64.1	68.9	88.2	88.2	83.0
3rd	NA	NA	24.6	39.1	48.3	63.2	77.9	77.9	66.7
4th	NA	NA	15.4	33.3	36.1	62.0	77.6	77.6	56.5
5th	NA	NA	NA	NA	29.6	46.1	NA	NA	NA

Note: The precision in the Tc bins of 0~0.1 and 0.1~0.2 are not available due to the low number of compounds (0 and three compounds, respectively). The precision for cells marked as ‘NA’ could not be determined because of the low number of compound–target predictions (< 30). When I evaluated the 2nd (3rd, 4th and 5th) prediction, I required that there should be at least 2 (3, 4 and 5) targets respectively of the most similar compound so that there is no bias for the calculation between the precision of 1st and 2nd (3rd, 4th and 5th) prediction. Therefore, the values in each column of Table 4.1.3 do not sum up to 100%. Due to the design of the widely and successfully used fingerprint scheme, Tc of 1 does not mean that two molecules are necessarily identical (see 4.2 Discussion). For the compounds that are in the STITCH database, I assigned their known targets with 100% target prediction precision.

ferent ranges of chemical similarity of the query compound to the closest training compound and for up to five top scoring known targets of this training molecule independently. To obtain robust precision estimates, I required a minimum of 30 compound-target predictions for each target rank in a given Tc interval (Table 4.1.3). The first step of the validation process consists of finding the most similar compound from the training set to a compound from the validation set (calculated by Tc). For instance, in the evaluation of the targets that ranked on the 3rd position there are only 5 compounds in the validation set whose highest Tc maps to the bin 0.2~0.3. In this case “NA” is assigned for the prediction precision for the 3rd target because of the threshold of 30 compound-target predictions. As a consequence, I did not report targets ranked on third position if the Tc for the most similar database compound maps to 0.2~0.3.

I observed that the precision increases with increasing Tc . For compounds with a Tc of ≥ 0.7 to the training set, the first predicted target was nearly always correct (almost 100%, see Table 4.1.3). Furthermore, the precision reached at

least 53% for a Tc in the range of 0.4~0.5. Thus, I chose 50% as default precision threshold for the predicted targets on the web server.

4.1.2 Implementation

HitPick web server offers two independent functions, namely, hit identification and target prediction of chemical screens. Below I described these two functions in detail.

Hit identification

The first function identifies bioassay hits based on the B-Score method and predicts targets for up to 100 hits (Fig. 4.1.1). As input, it requires the data from a bioassay, including plate names, compound identifiers, well positions, activity values and SMILES strings. The output is a table listing the hits and their chemical structures. Hits are determined by a p-value cut-off of 0.05. If the assay contains replicates of compounds, I require all replicates to be identified as hits. This table is used as input source for the target prediction method. The output of the target prediction is a list of target predictions for the input compounds ranked by decreasing precision.

Whenever the hit identification routine returns more than 100 compounds, target prediction is carried out for a structurally diverse (meaning as dissimilar as possible) subset consisting of 100 compounds by applying the MaxMinAlgorithm [122] (see Chapter 2 Materials and Methods) implemented in RDKit. This procedure is intended to facilitate the analysis of molecular targets putatively involved in the measured biological processes by focusing on a representative subset of hits.

Target prediction

In addition, HitPick allows the prediction of targets for up to 100 compounds independently from bioassay data (Fig. 4.1.2). For this second function, only SMILES strings are required as input. To ensure reliability of reported precision values I require a minimum of 30 compound-target predictions. The precision depends on the similarity to the most similar compound in the set of known interactions as well as on the rank of the target's score.

The results are displayed sorted by precision with a threshold of 50% by default. Users can select different precision thresholds for the target prediction

Hit Identification

Please upload your bioassay data: ?

Input example: -- | ⌵ ?

Input example:

PlateName	CompoundID	PlatePosition	RawValue	SMILES
A	1398603	I16	0.8658	Nc1nonc1N2CCCCC2
B	NA	H23	0.9019	NA

NA, "Not Available".

?

Number of tested compounds: 2092

Number of hits by B-Score method (p-value<0.05): 27 ?

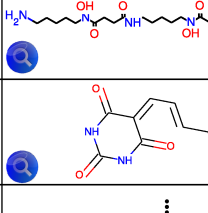
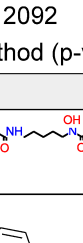
ID	Structure
745	
3062722	
⋮	⋮

Figure 4.1.1: Input and output scheme of hit identification.

results as desired. Under a lower threshold, more chemicals will have predictions at the cost of a lower precision. The targets are reported as gene symbols and more information can be found at STITCH (<http://stitch.embl.de/>) or GeneCards (<http://www.genecards.org/>). In addition, an overview of the predicted targets is given in form of pie chart.

4.1.3 Processing time

The processing time for hit identification depends on the size of the assay data. For bioassays containing less than 5,000, 10,000 and 100,000 compounds, the web server returns the results in less than 1, 2 and 30 minutes, respectively. The target prediction takes around 2 minutes per batch of query.

Target Prediction

Please upload your target prediction data:

... or paste the SMILES or, IDs (or names) and SMILES:

Input examples:
 1098 C[C@@]12C[C@H](O)C3C(CCC4=CC(=O)CC[C@]34C)C2CC[C@]1(O)C(=O)CO
 aspirin CC(=O)Oc1ccccc1C(=O)O

1525 *iraitga*

Type the two words

[Privacy & Terms](#)

Please select a precision higher than

ID	Structure	Target	Precision(%)	Tc similarity
3214824		HDAC3	97.7	0.94
1954149		NR3C1	89.8	0.6
⋮	⋮	⋮	⋮	⋮

Figure 4.1.2: Input and output scheme of target prediction.

However, as calculations are carried out on a shared cluster environment, actual processing time depends on the cluster workload.

4.1.4 Privacy

To preserve the privacy of the user data, only users are able to access to their uploaded data and results (from the same IP address as on query submission).

In addition, all data will be deleted automatically in seven days after they are submitted.

4.2 Discussion

HitPick is a novel web server for predicting the targets of small molecules with high quality. The drug target prediction methods implemented in HitPick, that is, 1-Nearest-Neighbour (1NN) similarity searching and Laplacian-modified naïve Bayesian target models exist. However, HitPick does not simply use these methods but integrate them in a single method that shows better performance than the two methods independently. An additional novelty of HitPick is that it is the first webserver available that identifies the hits for bioassays and predict their molecular targets. Regarding the quality of the server, I found that the performance of HitPick and Similarity Emsemble Approach (SEA) is comparable.

During the validation of HitPick target prediction approach, I observed that the precision of the drug-target prediction increases with increasing structural similarity (Tc) between the query compound and compounds of the database. However, for drug-target prediction with a structural similarity of 1 ($Tc=1$) I noticed a decreased precision (see Table 4.1.3). This decrease is due to the fact that Tc of 1 does not necessarily mean that two molecules are identical. On the one hand, the binary fingerprints capture the presence of molecular features, but not the frequency. For instance, they are unable to distinguish between a molecule and its dimer. On the other hand, these fingerprints do not rely on pre-defined moieties but can detect all possible combinations of atom environments. To lessen the computational complexity these fingerprints usually map the found features to a fixed-length bit string by means of a hashing function, so that in the end a single bit could potentially account for more than one feature. In order to reduce the chance of two non-identical compounds being Tc of 1, I selected the maximum fingerprint length determined by the utilized software. However,

this procedure did not remove all dissimilar compounds from the bin $Tc=1$. The manual inspection of compounds in the bins of [0.9-1.0) and of 1 and their most similar compounds revealed that with the exception of those actually identical compounds, the compounds in the bin of [0.9-1.0) are in fact more similar than the pair of compounds in the bin of $Tc=1$. The latter bin was enriched in compounds containing long aliphatic chains whose fingerprints were identical to functionally distant compounds with very short aliphatic chains and this explains the decrease in the drug target precision in this bin. To avoid the underestimation of target prediction precision for query compounds identical to compounds in the STITCH database, I introduced a feature in HitPick that automatically recognized if the query compound is identical to compounds in the STITCH database by comparing their SMILES strings prior to the generating of fingerprint bits and assigned their known targets with 100% target prediction precision.

HitPick is the first web server publicly available to facilitate the analysis of chemical screens by identifying hits and predicting their molecular targets. These two functionalities of HitPick can still be extended in the future. Currently HitPick allows to apply only the widely used B-Score method [61] for hit identification. I have shown in this thesis that other existing and novel methods have an optimal and even better performance to detect hits of high-throughput chemical assays (see Chapter 3 Chemical Hit Identification). All these methods can be easily implemented into HitPick hit identification functionality. For target prediction, as human is currently the species that contains largest number of known drug targets, HitPick focuses on human drug targets. With the increasing number of chemical-protein interactions in other species (STITCH 3 contains information of 1,133 organisms [99]), HitPick can also be extended to other species.

4.3 Conclusions

High-throughput phenotypic assays reveal information about the molecules that modulate biological processes, such as a disease phenotype and a signaling pathway. In these assays, the identification of hits along with their molecular targets is critical to understand the chemical activities modulating the biological system. Here, I present HitPick, the first web server for identification of hits in high-throughput chemical screens and prediction of their molecular targets. HitPick applies the B-Score method for hit identification and a newly developed

approach combining 1-nearest-neighbor (1NN) similarity searching and Laplacian-modified naïve Bayesian target models to predict targets of identified hits.

When evaluating of the highest scoring prediction for each compound, HitPick target prediction method achieved a sensitivity of 60.94% (with 66.16% being the maximum possible sensitivity), a specificity of 99.99% and a precision of 92.11%, which performs better than two individual target prediction methods, namely Bayesian models (sensitivity of 52.95%, specificity of 99.98% and precision of 80.03%) and 1NN similarity searching (precision of 84.72%). I believe that the application of HitPick to identify hits and predict targets of chemical screens in a systematic and comprehensive manner may help to unravel hidden molecular targets of chemicals, contributing to understand side effects of drugs and propose new drug therapeutic indications.

The server can be accessed at <http://mips.helmholtz-muenchen.de/proj/hitpick>.

Chapter 5

Chemical Screening Assay Pairs that Share Selective Hits Are Biologically Related

The screening of a library of compounds in a biological assay is a common first step in drug discovery to find chemical hits for the drug leads. A single chemical screening experiment provides information about the activity of compounds on a target or biological process. However, to select a chemical hit as chemical probe or drug lead, it is important to know additional properties of the compound such as its specificity and toxicity. An inexpensive and efficient manner to obtain information about these properties is to learn about the activity of this compound across multiple chemical screens. This approach is followed routinely in chemical screening programs such as the NCI60 project run by “US National Cancer Institute (NCI)” where the activity of a compound across 60 different cancer cell lines is measured to detect selective chemical hits for a particular cancer and avoid general toxicity [140].

In the last decade several initiatives including the NIH Molecular Libraries Program [141] and ChemBank [65] have compiled chemical biology experiments performed by different laboratories using diverse experimental set-ups ranging from cell-free to cell-based and even whole organism-based assays. The analysis of these heterogeneous datasets is challenging yet offers the possibility to obtain a global view of the chemical and biological activities of chemicals. In this regard, the integration and analysis of the collection of assays stored in the PubChem

5. CHEMICAL SCREENING ASSAY PAIRS THAT SHARE SELECTIVE HITS ARE BIOLOGICALLY RELATED

BioAssay [136] repository has proven to be useful to determine chemical properties of promiscuous compounds [142–144] and to predict adverse drug reactions [145].

The results of these studies suggest that a plethora of hidden molecular and biological information in these repositories can be uncovered using integrative computational methods. This is particularly relevant for the hits of phenotypic assays, for which the underlying molecular targets responsible for their activity is unknown. To determine the protein targets of the chemical hits of these assays, *in silico* target prediction methods [73,96,139] are arising as an efficient approach to obtain insights into the compound mode of action. For instance, Young et al. [137] have shown recently that the predicted molecular targets of hits are able to explain complex readouts of high-content screening assays.

Here, I exploited the vast amount of publicly available chemical screening assays present in the ChemBank database to evaluate in a systematic manner if a pair of biological activities modulated by common chemicals is related. I tested and confirmed this hypothesis by the systematic analysis of the molecular activities and biological processes measured in pairs of assays sharing non-promiscuous compounds in this repository. Subsequently, to understand the molecular mechanism linking pairs of phenotypic assays sharing chemical hits, I annotated the molecular targets of the shared hits. To that aim, I used HitPick [96], a recently developed *in silico* target prediction method to predict the molecular targets of compounds (see Chapter 4 HitPick). I found that the known biological role of the predicted targets of common chemical hits confirms the biological relationships between the assay pairs and provides mechanistic understanding of the relationships. This approach allows me to find relationships between biological activities and to understand better the molecular basis of the shared biological activities.

The results presented in this chapter have been submitted in X. Liu and M. Campillos, ‘Chemical screening assay pairs that share selective hits are biologically related’.

5.1 Results

5.1.1 ChemBank structure and chemical hit identification

I chose the ChemBank repository of chemical screens to test the hypothesis of whether a pair of biological processes modulated by the same chemicals is

related. In the ChemBank repository, the raw activity of a total number of 228,887 compounds in 3,834 assays (representing experimental batches) of 190 diverse projects is available.

In a first step, I identified the chemical hits of the individual assays. Since out of the eight methods applied (see Chapter 2 Materials and Methods), the B-Score_A method, a modification of the well known B-Score method [61] achieved the best performance with a sensitivity of 79.6% and a specificity of 97.4%. I thus, selected this method to determine the chemical hits of ChemBank assays. Then, I grouped chemical screen batches performed using identical experimental protocols into “assay types” (hereafter named “assays”) reducing the number of assays to 1,640 (see Chapter 2 Materials and Methods).

Next, in order to understand better the molecular or biological activity measured in the assays I analyzed and classified the assays part of ChemBank projects. I first classified the assays into “experiment” and “control”, according to whether the activity measured in the assay was the intended biological activity of the project or unspecific activities, respectively (Fig. 5.1.1A). In the second place, I classified the assays into cell-free, cell-based and microorganism based on the biological object of the experiments (Fig. 5.1.1A) (see Chapter 2 Materials and Methods). Lastly, I annotated the molecular activities and biological processes measured in the projects by assigning manually specific Gene Ontology (GO) [146] terms (biological process for phenotypic assays or molecular function for cell-free assays) to the projects (Fig. 5.1.1A). As an additional description of the activity tested in projects, I manually assigned suitable keywords representing protein/gene names or biological processes to the projects (Fig. 5.1.1A). I then propagated the GO terms and keywords of each project to its “experiment” assays.

I observed that the projects differ both in the number of assays (ranging from 1 to 113, Fig. 5.1.1B) and the percentage of “experiment” assays (Fig. 5.1.1C) they include. This observation underlines the heterogeneity of the composition of ChemBank dataset. The distribution of cell-free, cell-based and microorganism assays is also heterogeneous. More than 40% of the projects are composed of phenotypic assays (cell-based and microorganism), and the majority of them are cell-based assays (Fig. 5.1.1D, also see Appendix Fig. A.0.1). Interestingly, despite the inhomogeneity of the ChemBank dataset, I found that approximately 80% of the assays have more than 1,000 tested compounds (Fig. 5.1.1E) in common, indicating that the different assays can be compared based on the activity of a

large number of compounds.

5.1.2 Promiscuity filters and similarity in biological activity

Next, I tested the hypothesis of whether chemical screening assays belonging to different projects with a similar chemical hit profile are biologically related. To evaluate if two assays are related biologically, I applied the Lin measurement [147] that quantifies the semantic similarity between GO terms assigned to the assays. Additionally, I applied the biomedical text-mining tool “EXtraction of Classified Entities and Relations from Biomedical Texts (EXCERBT)” [123] that detects terms co-mentioned in abstracts of scientific literature to evaluate if the keywords linked to the assays of the pair are related (see Chapter 2 Materials and Methods).

Afterwards, for every assay with the set of compounds that show activity in at least two projects (Filter 1, F1) (Fig. 5.1.2A, F1) (see Chapter 2 Materials and Methods), I constructed a binary fingerprint vector representing the activity of the set of compounds in the assays (1 active chemical hit, 0 inactive). Next, for all possible pair wise fingerprint combinations of “experiment” type assays belonging to different projects, I calculated the chemical hit similarity using a weighted Tanimoto coefficient (Tc) [112] (see Chapter 2 Materials and Methods). Under these conditions, the assessment of the relationship between chemical hit similarity and the molecular and biological similarity of assay pairs did not reveal an association between hit and biological similarity (Fig. 5.1.2B and 5.1.2C, F1). I reasoned that promiscuous compounds might be responsible for the high chemical hit similarity in unrelated assays as the prevalence of nonspecific or promiscuous compounds is a well-known problem in High-Throughput Screening (HTS) assays commonly explained by their ability to form aggregates and act on unrelated targets [148]. Thus, their presence might be especially disturbing for the detection of biological connections between assay pairs.

Based on this assumption, I tested if the removal of promiscuous compounds increases the biological relatedness for assays sharing hits. To that aim, I applied two promiscuity filters. The first filter retained compounds with activity observed in less than 20% of the projects (Fig. 5.1.2A, F2) and the second filter (F3) kept compounds that are active in less than 20% of the assays within a project. To avoid discarding specific chemical hits in projects with low number of assays

5. CHEMICAL SCREENING ASSAY PAIRS THAT SHARE SELECTIVE HITS ARE BIOLOGICALLY RELATED

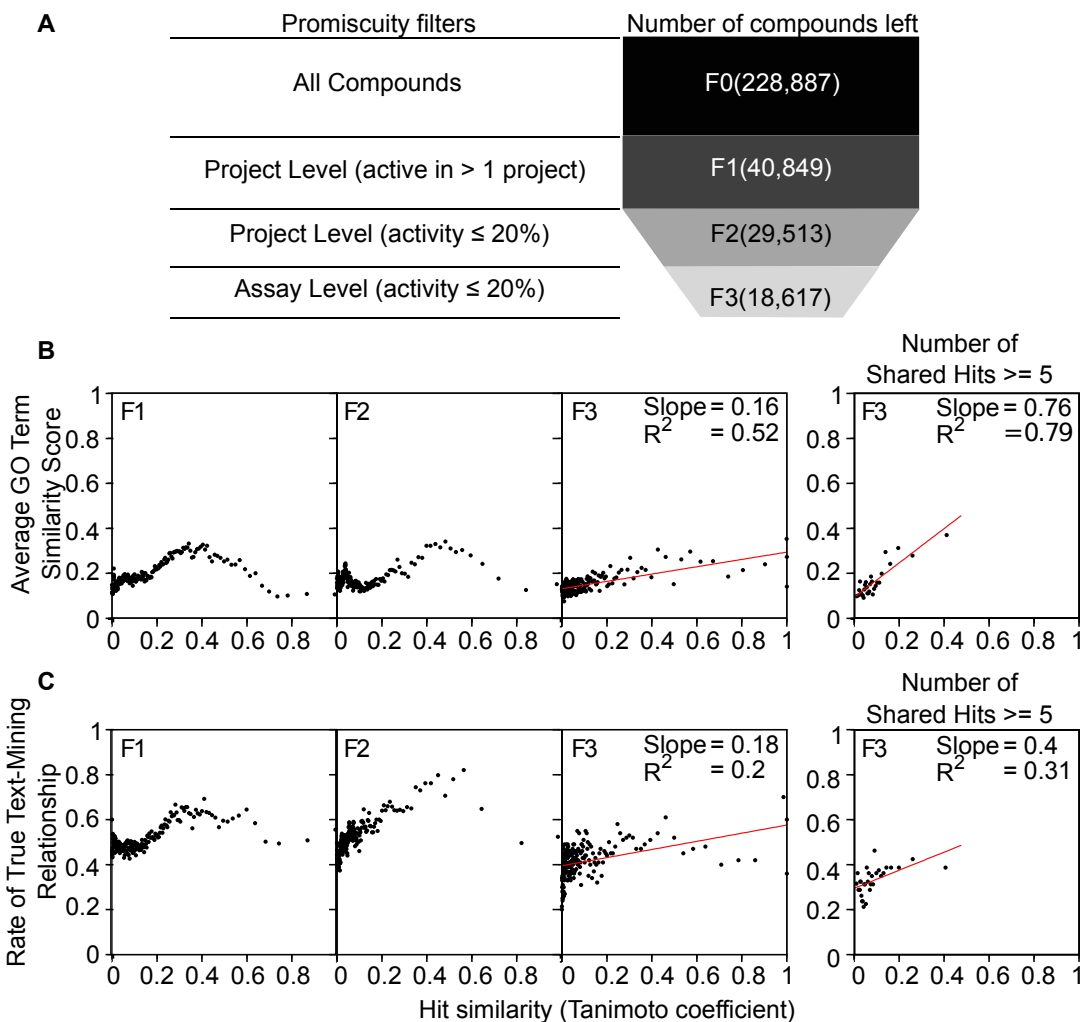


Figure 5.1.2: Promiscuity filters and correlation between hit similarity and known relationships of assay pairs. (A) Promiscuity filters. F0 contains all the compounds of the dataset. F1 keeps the compounds active in at least one project, and F2 retrieves the compounds active in $\leq 20\%$ of the projects. F3 retains compounds active in $\leq 20\%$ of the assays for the projects with higher than average number of assays (average number of assays per project is 9 for ChemBank). The number of remaining compounds after filtering is given in brackets. (B and C) Correlation between hit similarity and known relationships of ChemBank assay pairs. (B) Relationships indicated by GO terms and (C) relationships indicated by text-mining. Each point in the plot represents a bin of assay pairs according to the sorted Tc values. In F1, each bin contains 1,000 assay pairs. Bins in F2 and F3 contain 500 and 100 pairs, respectively. Separately, the performance of assay pairs in F3 sharing five or more hits is shown for (B) and (C).

where “experiment” assays represent more than 20% of all assays, the filter F3 was applied only to projects with at least nine assays (Fig. 5.1.2A, F3) (see Chapter 2 Materials and Methods). For example, the latter filter would discard all specific chemical hits in projects composed of one experiment and one control assay like the project “Glioblastoma Modulators” (Fig. 5.1.1A) that searched for PI3K and mTOR modifiers in glioblastoma cells. If applied to this project, this filter would remove all specific hits, that is, those compounds that are active in cells treated with rapamycin (“experiment”) and inactive in cells not treated with the mTOR inhibitor (“control”), since they are active on 50% (>20%) of the assays in this project.

As can be observed in Fig. 5.1.2B and 5.1.2C, only after the application of the most stringent promiscuity filter F3, a linear relationship between hit similarity and known biological relationships was observed. This trend became stronger when I discarded combinations of assays sharing low number of hits (Fig. 5.1.2B and 5.1.2C, number of shared hits ≥ 5 , also see Appendix Fig. A.0.2A and A.0.2B) indicating that the larger the number of common chemical hits is, the more likely it is to capture biological relationships between assays.

5.1.3 Assay interaction network

Next, I visualized and inspected manually the assay pairs showing high chemical hit similarity. For that, I constructed an assay interaction network with the assay pairs showing the highest hit similarity ($Tc > 0.4$) and sharing five or more chemical hits. This network contains 32 nodes and 26 edges (Fig. 5.1.3).

Interestingly, 92% of the edges in the network connect assays of the same experimental type. That is, phenotypic assays share hits with other phenotypic assays and cell-free assays tend to share hits with other assays of the same type. I found, for instance, a group of four interconnected assay pairs of the “microorganism” type (i.e. “Bacterial Viability”, “SigB Inhibition”, “Worm Anti-Infective” and “Anti-Bacterial” assays) where the same biological activity, that is, the antibacterial activity, was sought in all of them. An example of a connection of two clearly related cell-free assays is the link between “Kinesin Activity Eg5” and “Kinesin Activity MKLP1” comprised by two assays aiming to find inhibitors of proteins of the Kinesin family. These instances provide evidence that molecular and biological relationships between assays can be captured by our approach.

5. CHEMICAL SCREENING ASSAY PAIRS THAT SHARE SELECTIVE HITS ARE BIOLOGICALLY RELATED

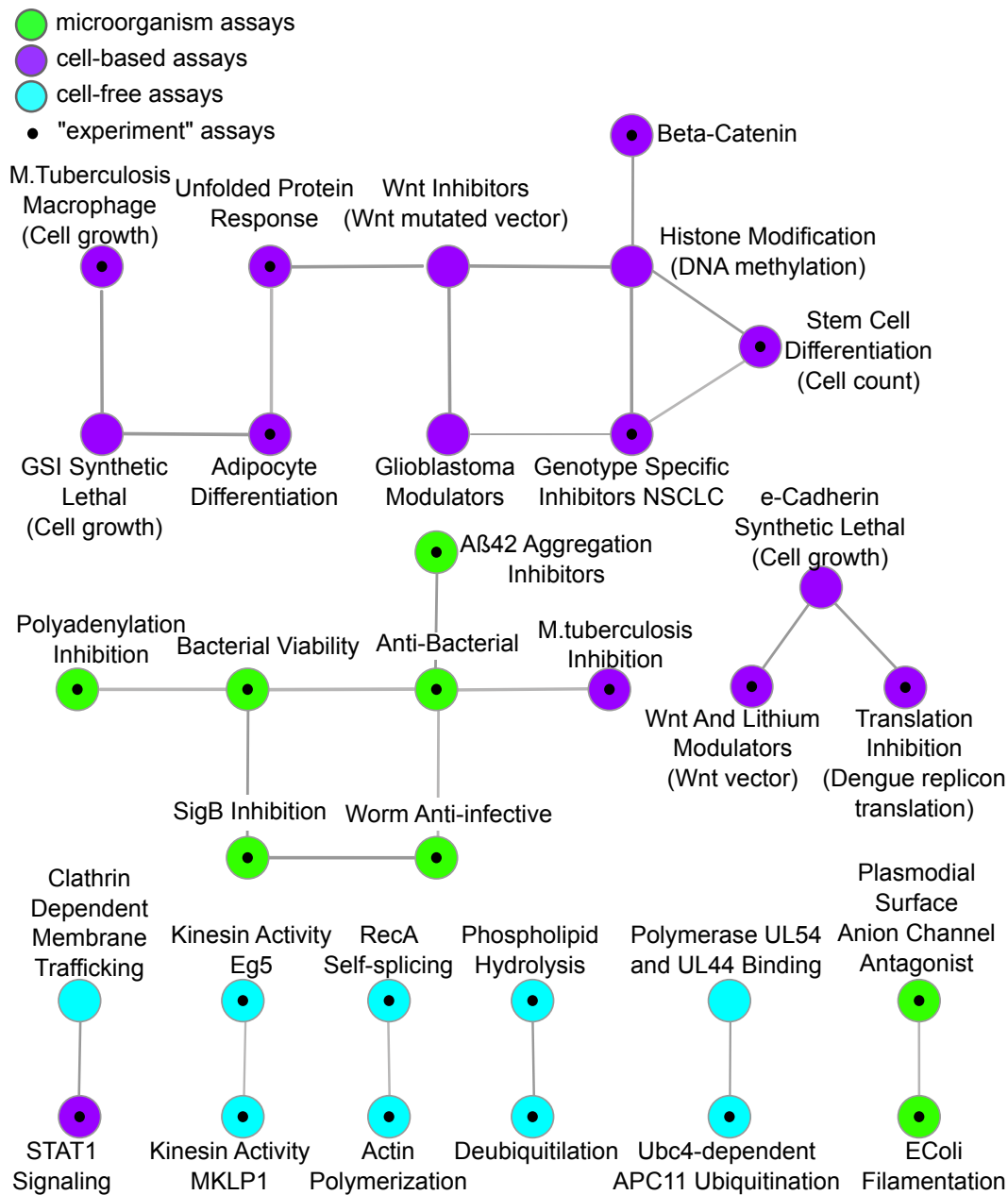


Figure 5.1.3: Network of assay pairs from ChemBank repository sharing selective hits.

Intriguingly, I found a high number of edges (11, representing 42% of the edges) connecting “control” assays to “experiment” assays, the majority of them (9) linking two cell-based assays. A closer inspection of the activities measured in these assays indicates that cell growth related processes such as differentiation or growth inhibition, were often measured in the assays as the sought activity, for example in assays seeking for chemicals with anticancer activity or in assays controlling the cytotoxicity of compounds. To gain deeper insights into the molecular basis of these assay combinations, I extracted molecular information of the chemical hits shared by these pairs by annotating predicted human drug targets of the compounds. For that, I applied the HitPick target prediction method [96] to predict the molecular targets of hits with high confidence (precision > 50%). Interestingly, I found the same predicted drug targets related to several assay pairs. For example, compounds specifically targeting the glucocorticoid receptor (NR3C1) are active in four consecutive assays in the network, namely “Mycobacterium tuberculosis (M.tuberculosis) Macrophage”, “Gamma Secretase Inhibitor (GSI) Synthetic Lethal (Cell growth)”, “Adipocyte Differentiation” and “Unfolded Protein Response (UPR)” (Fig. 5.1.4A). The role of NR3C1 in macrophages as the target of anti-inflammatory agents [149] and its anticancer activity [150] provide an explanation for the molecular basis of the relationship between the “M.tuberculosis Macrophage”, that screened for inhibitors of M.tuberculosis growth in macrophages and “GSI Synthetic Lethal (Cell growth)”, a “control” assay that tested the growth inhibitory activity of molecules in T-cells.

Moreover, the known ability of NR3C1 to induce adipocyte differentiation [151] explains the common link between the cell growth and differentiation activities measured in “GSI Synthetic Lethal (Cell growth)” and “Adipocyte Differentiation” assays, respectively. Interestingly, although the link between UPR and differentiation processes has been proposed in the literature [152], the molecular basis of this connection is not fully understood. Here, our result suggests the function of NR3C1 as intermediary between UPR induction and differentiation. However, this proposal should be taken with caution, as the specificity of the chemical hits on UPR process cannot be assessed due to the lack of controls assay in the project. In this context, the UPR assay is linked to a control assay of the “Wnt Inhibitors (Wnt mutated vector)” project, that measures the promoter activity of a mutated version of Wnt responsive construct (Fig. 5.1.4B). A closer look at this relationship reveals that ATP1A1 (ATPase, Na⁺/K⁺ transporting, alpha 1

5. CHEMICAL SCREENING ASSAY PAIRS THAT SHARE SELECTIVE HITS ARE BIOLOGICALLY RELATED

polypeptide), CYP1B1 (cytochrome P450, family 1, subfamily B, polypeptide 1) and ADORA2B (adenosine A2b receptor), are the predicted targets of the chemical hits of this pair. The role in cancer of ATP1A1 [153], CYP1B1 [154] and ADORA2B [155] indicate that the activity of compounds in the “Wnt Inhibitors (Wnt mutated vector)” assay is likely due to their cytotoxicity. Although the known role of UPR to induce cell cycle arrest [156] and the recently reported role of ouabain, specific inhibitor of ATP1A1, on the modulation of UPR [157], would suggest that the relationship between this assay pair is due to the UPR-dependent growth inhibitory activity, further research is needed to assess the specificity of the shared hits on the UPR assay.

The growth inhibition measured in the “Wnt Inhibitors (Wnt mutated vector)” assay is further confirmed by the association of this assay with the anti-cancer “Glioblastoma Modulators” and “Genotype Specific Inhibitors in Non-Small Cell Lung Cancer (NSCLC)” assays (Fig. 5.1.4C). Our target prediction approach revealed that, within this group of growth inhibitory assays, the cytotoxic activity is partly mediated through well-known anticancer targets, such as histone deacetylases (HDACs) [158], ATP1A1 [153], farnesyltransferase, CAAX box, alpha (FNTA) [159] and mouse double minute 2 homolog (MDM2) [160]. Furthermore, the modulation of these targets also explains the link between the chemical screens measuring stem cell differentiation [“Stem Cell Differentiation (Cell count)” assay], and DNA methylation [by 4,6-diamidino-2-phenylidole (DAPI) staining in “Histone Modification (DNA methylation)” assay]. Intriguingly, other predicted targets behind the growth inhibition activity in this group of cancer related assays include adenosine receptor A3 (ADORA3), cannabinoid receptor 2 (CNR2), cholesteryl ester transfer protein, plasma (CETP), 5-hydroxytryptamine receptor 6 (HTR6) and ATPase, Ca²⁺ transporting cardiac muscle, fast twitch 1 (ATP2A1). The modulation of these targets in anticancer screens suggests the possible role of these proteins in growth inhibition. In fact, the activity of ADORA3 as a potential target for tumor growth inhibition has been proposed before [161].

Another well-known biological connection is represented by the link between “Beta-Catenin” assay that measured the nuclear translocation of beta-catenin and “Histone Modification (DNA methylation)” assay (Fig. 5.1.4D). HDAC, the predicted target of the common hits, has been shown to inhibit Wnt signaling through disruption of the interaction between beta-catenin and T cell factor [162]. Thus, the biological relationship between these two assays is explained by the known

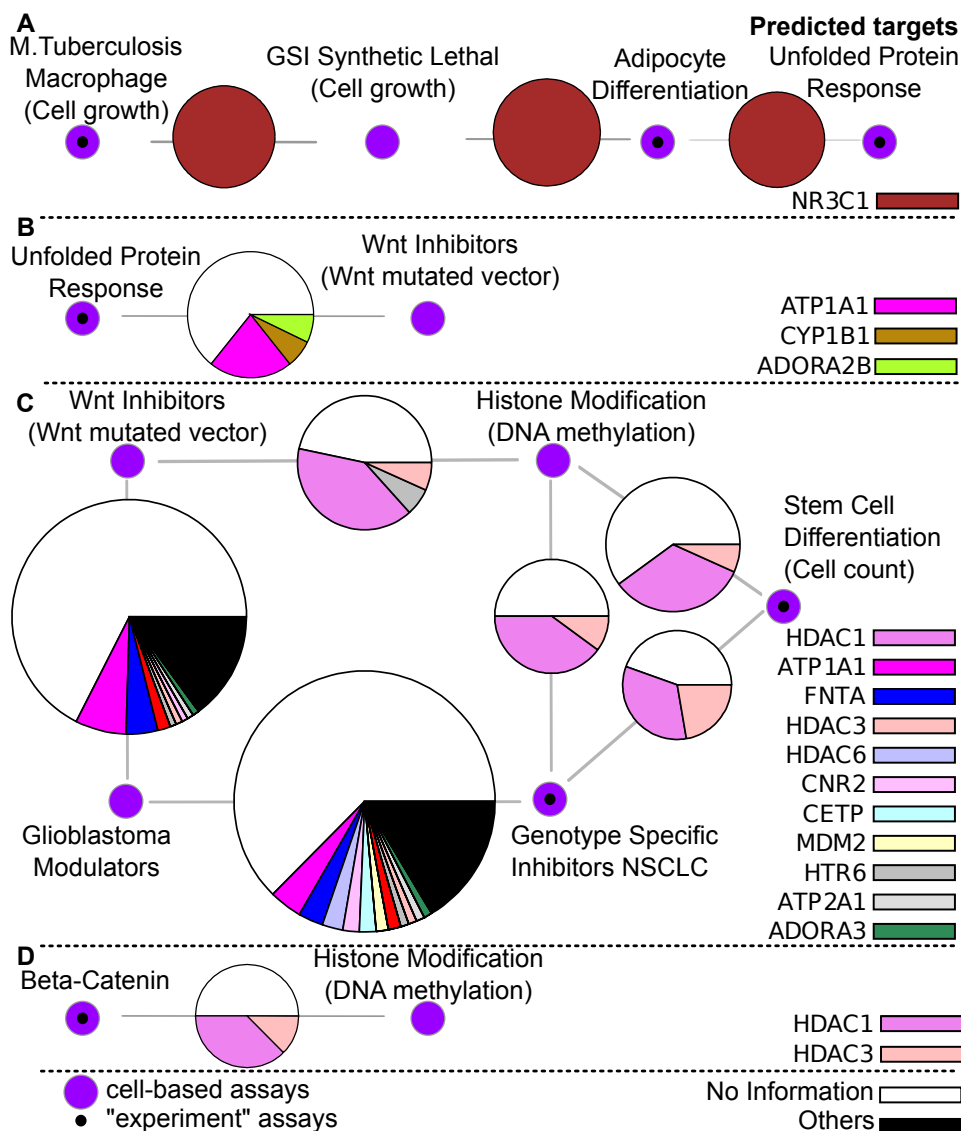


Figure 5.1.4: Enriched targets between assay pairs. (A, B, C, D) are the examples of assay connections (shown by assay name). The size of each pie chart is proportional to the logarithm of the number of shared hits. For simplicity, in the pie charts I show the most frequently predicted targets (with a precision higher than 50%) of the shared chemical hits (see Appendix Table A.0.1 for the full target list of each assay pair in Fig. 5.1.4). The fraction of the pie charts representing hits with no predicted targets is shown in white as “No Information”. In Fig. 5.1.4C, only those representative targets common to 3 hits for assays pairs in the group are shown, and the remaining targets common to ≤ 2 hits are shown in black as “Others”.

relationship of HDACs.

In summary, after retrieving the chemical hits from the ChemBank assays, I observed that biological activities measured in two assays sharing selective hits are related. The close inspection of the assay pairs sharing specific hits in the network is able to confirm the biological associations of assay pairs and reveal molecular information underling the shared activity.

5.2 Discussion

In this work, I have integrated and analyzed the information stored in ChemBank and demonstrated that biological activities of assay pairs sharing selective chemical hits are often related. The biological relationships between phenotypic assays are furthermore supported by the role of protein targets predicted for the shared hits.

Fingerprint-based approaches, where profiles of a collection of predefined features of an object such as a compound or protein is compared, have often been exploited in Chemistry and Biology fields to infer properties of compounds [100,112] and genes [96]. These approaches are based on the observation that similar fingerprint profiles correlate with similar properties [163]. For example, compounds with similar chemical fingerprint profiles tend to have similar biological activities [164]. Likewise, compounds with similar modes of action have also been observed to exhibit similar behavior across multiple assays [165]. In contrast, in this study I use chemical hit-based fingerprints constructed with selective compounds to infer bioactivity relationships. Interestingly, I show that the relationships between assays can only be captured when a stringent selectivity filter is applied to discard promiscuous compounds from the chemical hit profile. Currently, there is no consensus for the definition of compound promiscuity and different promiscuity filters have been proposed in the literature. Schürer et al. [144] and Jacob et al. [166] defined promiscuous compounds as those showing activity in more than 50% or 30% of the assays, respectively, while Gamo and colleagues [82] calculated an ‘inhibition frequency index’ for each compound and applied a variable threshold, ranging from 5 to 20% of screens, depending on the number of HTS screens a given compound had been through. Although these studies have revealed interesting chemical moieties associated to unspecific signals in chemicals screens, the question of what level of selectivity is necessary to capture hits carrying informa-

tion about specific biological signals has not been addressed yet. In this study, I have shown that a stringent promiscuity filter that first selects hits active in less than 20% of the projects (filter F2) and subsequently retains compounds with activity in less than 20% of the assays within a project (filter F3) is necessary to obtain hits with specific biological activities. I reason that the low number of projects performed in the same experimental backgrounds generating the same unspecific signals might be the cause for the lack of correlation between hit and biological similarity of two assays after the application of filter F2. Although this is partially overcome by discarding compounds active in several assays of the same project and consequently, performed in similar experimental backgrounds (filter F3), our approach also detects connections between cell-free assays that are apparently unrelated. For example, the “Phospholipid Hydrolysis” assay is associated to the “Deubiquitilation” assay (Fig. 5.1.3). A closer look at this connection reveals artifactual yet non-promiscuous hits, as the shared hits of the two connections appear active in the control assays of the project (termed “unspecific” chemical hits, see Fig. 5.1.1A). This indicates that the stringent promiscuity filters applied here might, for some experimental conditions, be insufficient to discard unspecific hits, and additional control assays might be necessary to remove non-selective chemical hits.

The presence of unspecific hits is also evidenced by the occurrence of edges that connect “control” and “experiment” assays. For example, the “e-Cadherin Synthetic Lethal (Cell growth)” “control” assay that controlled for the cytotoxicity of compounds in the human mammary epithelial HMLE cell line is connected to the “Wnt And Lithum Modulators (Wnt vector)” “experiment” assay (Fig. 5.1.4), suggesting that the shared hits of the pair are not specific of the Wnt signaling process. This hypothesis is further corroborated by the known or suspected anti-cancer activity of the predicted targets (HDAC1 [158], FNTA [159] and sigma non-opioid intracellular receptor 1 (SGIMAR1) [167], see Appendix Table A.0.1) of the shared hits and the modulation of these targets in a control assay of “Wnt Inhibitors (Wnt mutated vector)” (Fig. 5.1.4C, also see Appendix Table A.0.1). Similarly, the link between the cytotoxic “control” assay of the “e-Cadherin synthetic lethal (Cell growth)” project and the “Translation Inhibition (Dengue replicon translation)” assay that detected inhibitors of the translation of Dengue virus replicon (Fig. 5.1.3) points to the unspecificity of the chemical hits in the “Translation Inhibition” assay. These examples illustrate the need of

5. CHEMICAL SCREENING ASSAY PAIRS THAT SHARE SELECTIVE HITS ARE BIOLOGICALLY RELATED

additional control assays in these screening projects to assess the specificity of the compounds. Nonetheless, I show that this approach was able to capture meaningful biological connections even between different types of assays, such as the link between a microorganism assay with a cellular assay, which also able to inform about biological connections. For example, the microorganism “Anti-Bacterial” assay is connected with cellular “M.tuberculosis Inhibition” assay performed in BG1 ovarian cancer cells.

I observe that many relationships between different phenotypic assays are established based on the shared cytotoxicity of compounds in cell- or whole organism-based assays. Cytotoxicity appears thus as underlying biological effect common to phenotypic assays that accounts for the activity of many hits in these assays. Interestingly, the target prediction for those “non-promiscuous” but “cytotoxic” compounds reveals targets of drugs used as anticancer therapies, such as the HDACs [158] and ATP1A1 [153], or targets that have been proposed for cancer treatment such as FNTA [159] and MDM2 [160]. Hence, other predicted targets connecting these assays might represent potential targets for the treatment of cancers, such as CNR2, CETP, HTR6, ATP2A1 and ADORA3. Indeed, ADORA3 has been proposed as a potential therapeutic cancer target [161].

In summary, this work shows the potential of integrative approaches dealing with high-throughput chemical screening data to reveal novel biological connections. In the future, with the expected increase in HTS assay data available in public repositories, it is envisioned that many more biological relationships will be discovered with the application of this or similar computational approaches.

5.3 Conclusions

By integrating and analyzing the activity of small molecules across multiple chemical assays stored in ChemBank repository, I observe that assay pairs that share non-promiscuous chemical hits tend to be biologically related. A detailed analysis of a network containing assay pairs with the highest hit similarity confirms biological meaningful relationships. Furthermore, the biological roles of predicted molecular targets of the shared hits reinforce the biological associations between assay pairs, like the enrichment of known anticancer drug targets in growth inhibition assays. Thus, I show that the systematic comparison of the selective hits of chemical screening assays is a promising approach to uncover relation-

ships between biological activities, such as the potential growth inhibitory effect of ATP2A1, etc.

Chapter 6

Target Identification in High-Throughput Phenotypic Screens

The identification of pathways involved in human diseases forms the foundation for designing mechanism-based therapies. Chemical genetic approaches allow the detection of modulators of biological targets relevant for disease pathways by target-based screens as well as the discovery of small molecules with a desired outcome by phenotype-based screens. The latter biological genetic strategy is re-emerging as a valuable drug discovery approach due to the reduced success of target-based method to discovery of new medicines [59]. However, the crucial challenge of phenotypic assays is the identification of the targets of hits and subsequent validation of the relevant activity of the target on the phenotype.

Several target identification strategies have been followed to determine targets, including direct biochemical, genetic interaction, and computational inference methods (reviewed in Ref. [68]). Before identifying the targets of hits in phenotypic assays, one important aspect to be considered is that hits from high-throughput screening (HTS) have to be treated with caution, as they are not free of experimental artifacts or specific enough for the seek of biological outcome. Furthermore, since small molecules tend to interact with multiple targets [168] that might represent the false positives, even if the target of a compound is known or identified, it is necessary to prove the relationship between the protein targets and the phenotype. In this work, I hypothesized that the enrichment of multiple

compounds with the same molecular mechanism of action among the assay hits is likely to indicate unanticipated connections between targets and biological processes. Here, I explore this hypothesis by using a statistical method to determine the targets that are enriched among the specific hits of an assay when compared against targets of inactive compounds.

The results presented in this chapter are in preparation for submission in X. Liu et al., ‘Target identification in high-throughput phenotypic screens’.

6.1 Results

To test and validate my approach, I carefully selected phenotypic projects from ChemBank repository [65] where control assays accounting for non-specific hit effects (hit identification see Chapter 3 Chemical Hit Identification) were included. The “Modulators of lipid transfer”, “Modulators of PGC-1 α expression” and “Modulators of Wnt signaling” assays fulfilled this criterion as well as constituted interesting phenotypes to be analyzed. “Modulators of Lipid Transfer” project seeks regulators of the cholesterol transport mediated by scavenger receptor, class B, type I (SCARB1) transporter; “Modulators of PGC-1 α Expression” project searches modulators of the PGC-1 α expression and “Wnt Signaling Modulators” is a gene reporter assay to identify modulators of Wnt pathway. Then, for the three assay projects I defined the “Specific hits” and “Inactive compound” sets. “Specific hits” set includes those compounds which are active in the experimental assay measuring the phenotype of interest and inactive on corresponding control assays (Fig. 6.1.1A, number of specific hits for each project see Fig. 6.1.2). The “Inactive compounds” set contains all the remaining compounds that are inactive in the experimental assay (Fig. 6.1.1A)

Next, I predicted the molecular targets of compounds in the two sets for every project by applying HitPick [96], a ligand-based target prediction method that combines 1-Nearest-Neighbour (1NN) similarity searching and Laplacian-modified naïve Bayesian machine learning to predict direct human binding targets at a high confidence level (precision > 50%). On average, I predicted targets for 57% of the 1,300 specific hits (Fig. 6.1.2) and for 54% of the 39,353 inactive compounds.

In order to determine the targets of hits enriched in specific hits of the assays, and thus, more likely to be relevant to the phenotypic response, I subsequently applied the hypergeometric test to detect predicted target(s) that are

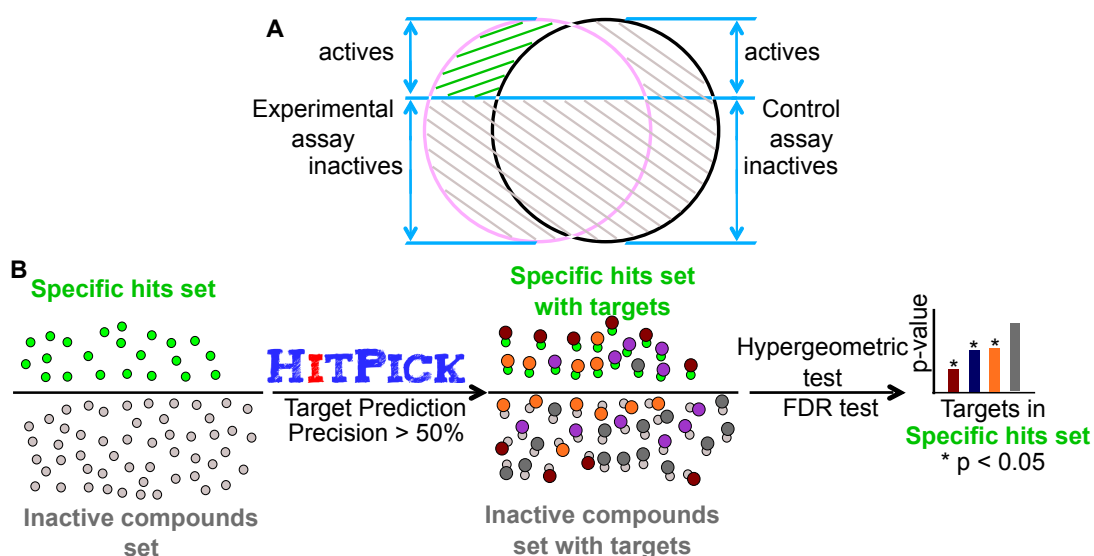


Figure 6.1.1: Specific hits and inactive compounds sets. (A) Classification of these two sets. (B) Scheme to get the significantly over-representative targets for specific hits set.

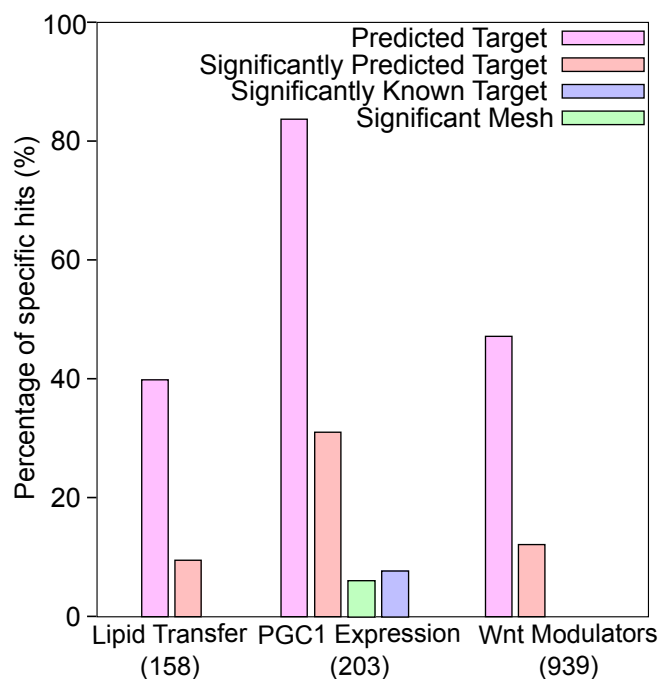


Figure 6.1.2: Percentage of specific hits in three analyzed assay projects. The number of specific hits in each assay project is displayed in brackets. The percentage of hits with predicted targets were compared with the percentage of hits with significantly predicted targets, significantly known targets and significant MeSH pharmacological action terms.

over-represented in the “Specific hits” set when compared to “Inactive compounds” set for each project. Targets with a resulting p-value lower than 0.05 after false discovery rate (FDR) multiple testing correction [125] were selected for further evaluation (Fig. 6.1.1B).

6.1.1 Molecular space explained by enriched targets

To evaluate the fraction of specific hits whose molecular action on the phenotype can be explained by the predicted targets obtained with my computational target prediction approach, I calculated the fraction of selective hits with predicted activity on the significant targets. In total, I detected 26 significantly over-represented targets for these three projects. I observed that on average, 18% of the selective hits were predicted to have activity on at least one of those significant targets (Fig. 6.1.2), indicating that I was able to explain the activity of up to 18% of the hits on the different phenotypic assays. In contrast, only 3% and 2% the specific hits can be mapped, on average, to known targets (from STITCH 3 database [99]) and pharmacological action terms (from Medical Subject Headings (MeSH) pharmacological action dictionary [169]), respectively. This implies that the molecular space of hits covered using target prediction information is of a 7-fold order magnitude higher than the space covered when using the information of the known molecular activity of compounds (Fig. 6.1.2), demonstrating that the use of target prediction information provides a better overview of the molecular space covered by hits of a phenotypic project than the known activity of the hits.

6.1.2 Validation of the approach based on literature

In order to demonstrate whether my approach reveals protein targets relevant to the measured phenotypes, I performed extensive literature searches for evidences supporting the relationships between the enriched molecular targets with the biological processes represented in the project. In total, I found well-supported literature confirmation for 23 out of the 26 predicted targets of all projects.

Below I explain the biological relationships found between significant targets and phenotype for the phenotypic assays.

Case studies illustrating targets enriched in phenotypic assays

Modulators of lipid transfer

The first project that I analyzed is the “Modulators of lipid transfer” project, that aimed to find selective modulators of the transfer of lipids mediated by the high-density lipoprotein (HDL) receptor, scavenger receptor, class B, type I (SCARB1), which functions both the selective uptake of HDL cholesterol esters, from HDL to cells and the efflux of cholesterol from cells to lipoproteins [170].

Five significant targets enriched among the selective hits of this assay appear, namely, the amyloid beta precursor protein (APP), the nuclear receptor coactivator 2 (NCOA2), the 1-acylglycerol-3-phosphate O-acyltransferase 2 (AGPAT2), the retinoid X receptor, alpha (RXRA) and the fms-related tyrosine kinase 3 (FLT3) related to lipid transfer (Fig. 6.1.3).

Out of the five proteins, three of them (AGPAT2, NCOA2 and RXRA) are known to be involved in the lipid transfer process. AGPAT2 increases the activation of peroxisome proliferator-activated receptor γ (PPARG) [171]. PPARG forms heterodimer with RXRA to control the expression of genes involved in adipogenesis, among other metabolic process [172]. NCOA2, a coactivator for steroid receptors, in turn, interacts with PPARG-RXRA complex [173] to alter the expression of key regulatory genes of energy metabolism, such as increasing the expression of SCARB1 [174]. Taken together, these evidences suggest that knockout any of the predicted targets will inhibit SCARB1 activity, repress lipid transferring process and further confirm the validity of my approach to detect targets involved in the lipid transfer.

Modulators of peroxisome proliferator-activated receptor- γ coactivator-1 α (PGC-1 α) expression

This project aimed to detect compounds modulating the expression of PGC-1 α , a transcriptional cofactor that plays a central role in the genetic regulation of pathways, such as glucose homeostasis and mitochondrial biogenesis [175]. Here, the targets predicted to be linked to the phenotype are the β adrenergic receptors (ADRB1, ADRB2 and ADRB3), the glucocorticoid receptor (NR3C1), the cytochrome P450, family 1, subfamily A and B, polypeptide 1 (CYP1A1 and CYP1B1), the mitochondrial NADH dehydrogenase subunit 4 (ND4), the serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6

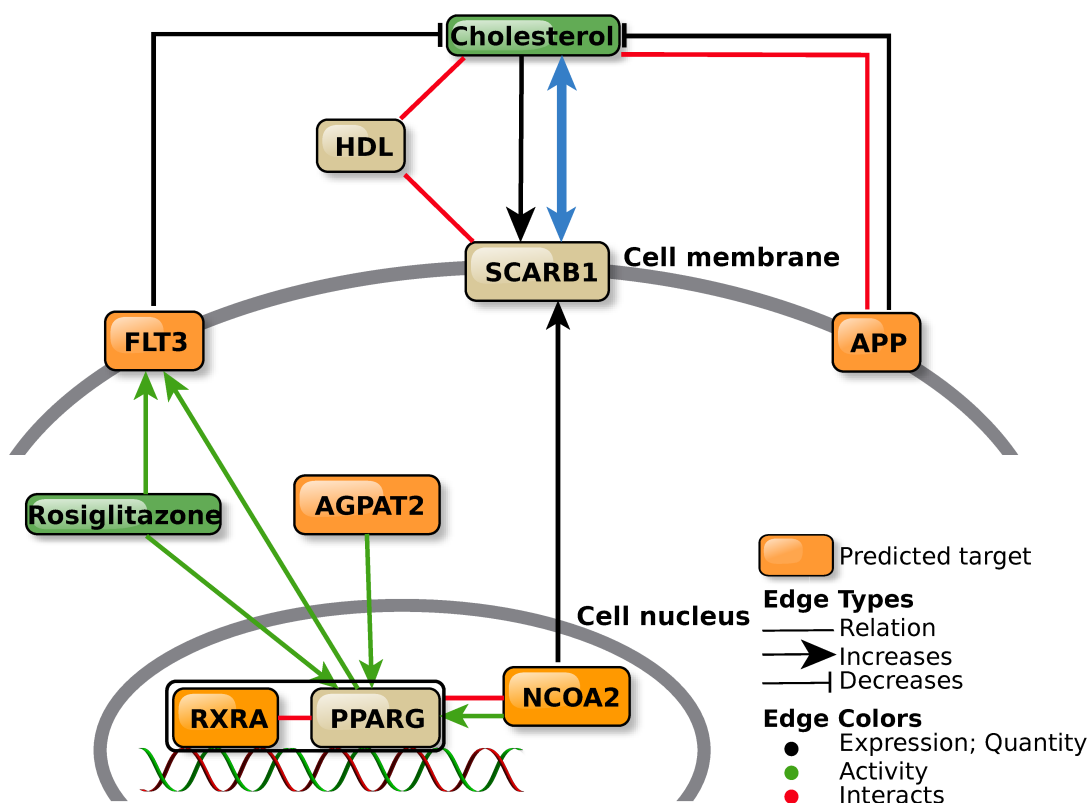


Figure 6.1.3: Assay project of “Modulators of lipid transfer”. The blue colored edge is the measured phenotype. In this graph, the phenotype is the transferring of cholesterol mediated by the HDL receptor, SCARB1. Brown colored rectangles are the background proteins and green colored rectangles are the chemical ligands of the predicted targets. These two colored objects facilitate explaining the molecular function of the predicted targets (orange).

(SERPINA6), the phospholipase A2, group IV A (PLA2G4A), Na⁺/K⁺ ATPase subunit alpha 1 (ATP1A1) and the NAD(P)H dehydrogenase, quinone 2 (NQO2) (Fig. 6.1.4).

Out of eleven targets, eight have been shown to related to modulate PGC-1 α expression. For example, glucocorticoids are transported in the blood by SERPINA6 [176]. They enter the cell and bind to their receptor (NR3C1, also known as GR) in the cytoplasm. Upon ligand binding, NR3C1 translocates to the nucleus [177], where it interacts with the glucocorticoid response elements (GRE) in the promoter region of the ADRB2 gene, resulting in the increased transcription ADRB2 [178]. The β adrenergic receptors (ADRB1, ADRB2 and ADRB3), in turn, promote the expression of PGC-1 α [179].

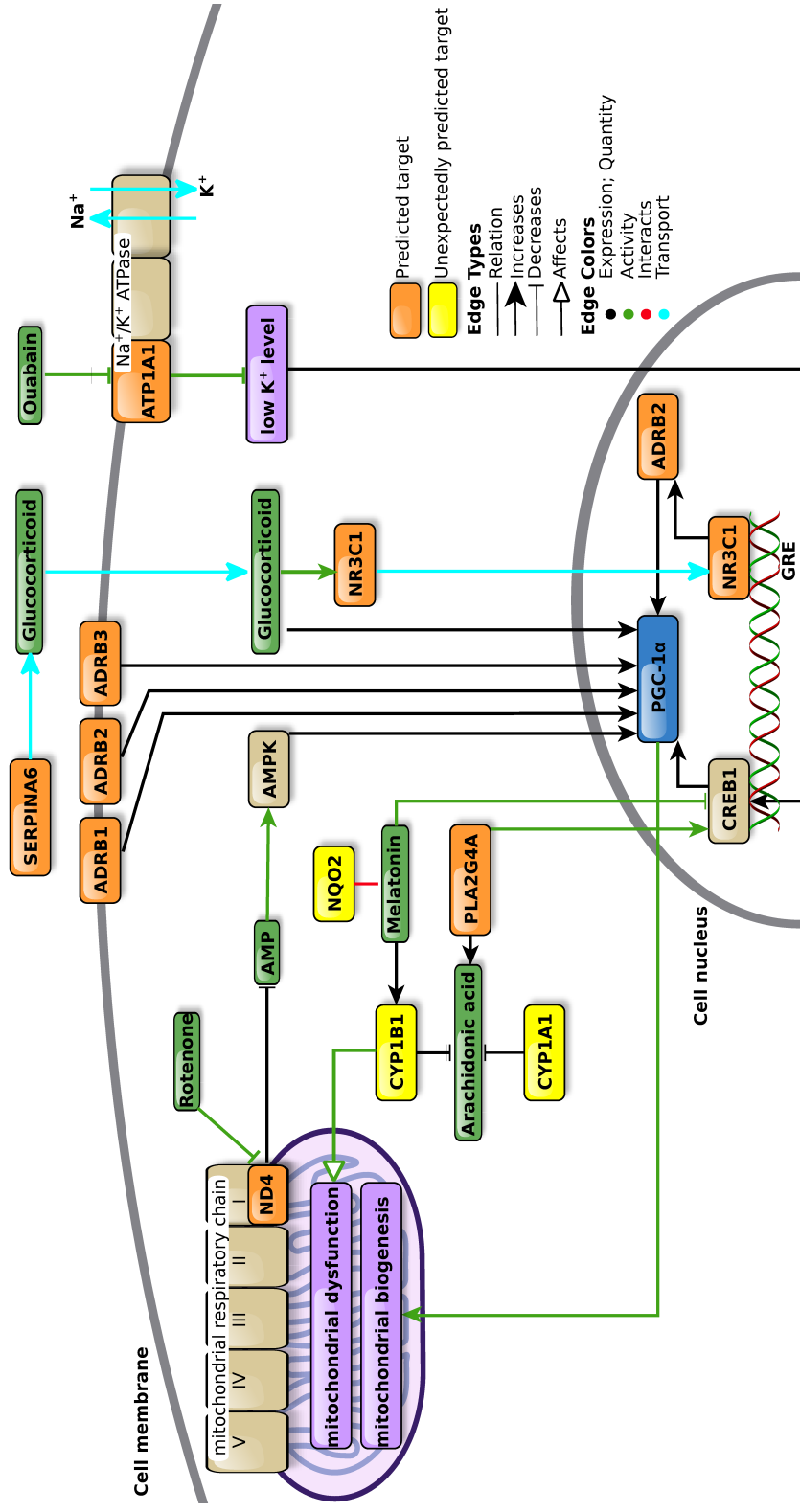


Figure 6.1.4: Assay project of “Modulators of PGC-1 α expression”. The blue colored rectangle is the measured phenotype. In this graph, the phenotype is the PGC-1 α expression. Brown colored rectangles are the background proteins and green colored rectangles are the chemical ligands of the predicted targets. Purple colored rectangles are the biological processes where the predicted targets are involved. These three colored objects facilitate explaining the molecular function of the predicted targets (orange and yellow). Yellow colored rectangles are the predicted targets that are unexpected to occur in the assay project.

Another target, ND4, a subunit of complex I of the mitochondrial respiratory chain, regulates the expression of PGC-1 α via AMP-activated protein kinase (AMPK) [180]. If complex I is inhibited, ATP will not be produced by the respiratory chain and thus AMP levels will stay high. High AMP levels induce AMPK [181] which up-regulates the expression of PGC-1 α . PGC-1 α activates the broad program of mitochondrial biogenesis, which equips the cell to meet the energy demands (ATP) of the cell [182].

PLA2G4A and ATP1A1 are targets that affect the measured phenotype by increasing the expression of cAMP responsive element binding protein 1 (CREB1), a PGC-1 α transcription factor [183]. PLA2G4A is an enzyme that catalyzes the hydrolysis of cellular phospholipids to liberate arachidonic acid [184]. ATP1A1 is a subunit of the Na⁺/K⁺ ATPase that pumps sodium (Na⁺) out of the cell and potassium (K⁺) into the cell. PLA2G4A activates the cAMP responsive element binding protein 1 (CREB1) [185] and inhibition of ATP1A1 by ouabain leads to low intracellular K⁺ levels [186] which further induce the expression of CREB1 [186].

Modulators of Wnt signaling

Next project “Modulators of Wnt signaling” screened the chemicals modulating the Wnt pathway. Here I predicted 10 significantly over-represented targets that specifically participate on the modulation of Wnt pathway, namely the APP, the Bcr-Abl (ABL1), the mammalian target of rapamycin (MTOR), the dehydrogenase 2 (ALDH2), the melanin-concentrating hormone receptor 1 (MCHR1), the monoamine oxidase B (MAOB), the cytochrome P450, family 1, subfamily A and B, polypeptide 1 (CYP1A1 and CYP1B1), the acetaldehyde 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGCR) and the cytochrome P450, family 19, subfamily A, polypeptide 1 (CYP19A1) (Fig. 6.1.5).

All of these predicted targets are known to modulate the Wnt signaling pathway with β -catenin being a key modulator [187]. For instance, APP is reported to physically interact with presenilin 1 (PSEN1) [188] which negatively regulates β -catenin (also known as CTNNB1) [189]. ABL1 physically interacts with β -catenin and triggers its tyrosine-phosphorylation [190]. It was also shown that the inhibition of MTOR rapidly activates Wnt pathway [191]. The other three targets, ALDH2 [192], MAOB [192] and MCHR1 [193] are reported to regulate serotonin, which is required for Wnt signaling in the early embryo [194]. Inhibition of

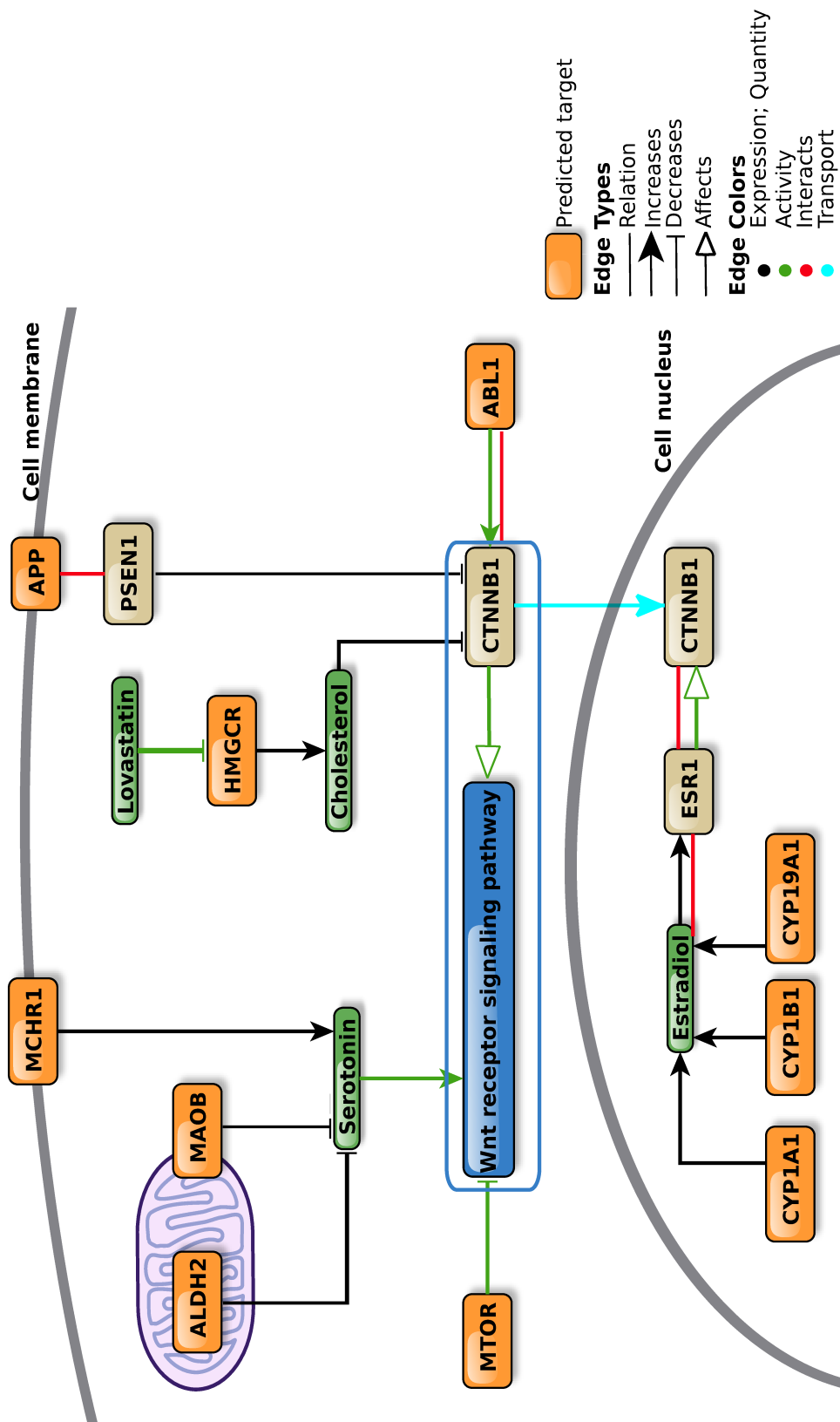


Figure 6.1.5: Assay project of “Modulators of Wnt signaling”. The blue colored rectangle is the measured phenotype. In this graph, the phenotype is the Wnt receptor signaling pathway. Brown colored rectangles are the background proteins and green colored rectangles are the chemical ligands of the predicted targets. These two colored objects facilitate explaining the molecular function of the predicted targets (orange).

HMGCR by the specific hit lovastatin, is known to modulate β -catenin via cholesterol [195,196]. The remaining cytochrome P450 family members, CYP19A1 [197], CYP1A1 [198] and CYP1B1 [199] have been shown to regulate the quantity of estradiol. Binding of estradiol to its receptor (ESR1) leads to the interaction with β -catenin [200] and further modulates the activity of β -catenin [201].

In summary, previous research literature reports confirm the relationship of 23 (88%) out of the 26 significantly over-represented targets in three chemical screening assays, with the biological activity tested in the assay, thereby illustrating the effectiveness of my approach to predict the targets of the biological pathways measured in bioassays.

6.1.3 Validation of the approach based on known activity of hits

As an additional validation, I determined whether compounds with known activity on those predicted targets were part of the specific hits of the assays. For 13 out of 26 targets, I observed that at least one of the compounds with known activity on the targets showed specificity on the assays (Table 6.1.1). Furthermore, for 54% of those targets, the relationship of the compounds (the drugs that are labeled in bold in Table 6.1.1) with the phenotype has been previously reported in the literature, supporting the validity of the approach to capture molecular targets related to phenotypes.

Table 6.1.1: Predicted targets, specific hits with known activity and false negatives of each assay project

Projects	Predicted targets	Among the specific hits, drugs that are known to interact with the predicted target	False negatives
Modulators of lipid transfer	APP	—	—
	NCOA2	—	—
	AGPAT2	—	—
	RXRA	—	—
	FLT3	—	—

Modulators of PGC-1 α expression	ADRB2	salbutamol (PMID: 16239931), noradrenaline (PMID: 21151149), bisoprolol (PMID: 18400557), alprenolol, terbutaline, fenoterol, orciprenaline, tulobuterol, pindolol, isoprenaline, procaterol, dobutamine.	propranolol (PMID: 17446185), clenbuterol (PMID: 22071161)
	MT- ND4	rotenone (PMID: 23930106)	
	ADRB3	noradrenaline (see above), alprenolol, terbutaline, fenoterol, tulobuterol, pindolol, isoprenaline, procaterol	propranolol (see above), clenbuterol (see above)
	ADRB1	salbutamol (see above), noradrenaline (see above), alprenolol, terbutaline, fenoterol, tulobuterol, pindolol, isoprenaline, procaterol, dobutamine, bisoprolol (see above)	propranolol (see above), clenbuterol (see above)
	NR3C1	—	progesterone (PMID: 20133449), spironolactone (PMID: 20211973), dexamethasone (PMID: 17335662), triamcinolone (PMID: 7929119)
	NQO2*	melatonin (PMID: 20557470), primaquine	
	CYP1A1*	albendazole, dicycloverine, primaquine, lansoprazole, pentamidine	fluvastatin (PMID: 19150877), chrysin (PMID: 11343698)
	PLA2G4A	mepacrine	—
	ATP1A1	—	—

	SERPINA6	—	hydrocortisone (PMID: 7929119)
	CYP1B1*	apigenin, primaquine	chrysin (PMID: 11343698), quercetin (PMID: 19211721), luteolin (PMID: 19914244), kaempferol (PMID: 21728151)
Modulators of Wnt signaling	APP	4-(1,3-benzothiazol-2- yl)aniline	—
	CYP19A1	clotrimazole, biochanin A, 7-hydroxyflavone	flavone (PMID: 21652696)
	CYP1B1	acacetin, luteolin (PMID: 20013030)	quercetin (PMID: 19440933), galangin (PMID: 21406604)
	HMGCR	lovastatin (PMID: 17234346)	—
	MCHR1	—	—
	MAOB	—	fluoxetine (PMID: 20979321)
	ALDH2	—	—
	ABL1	—	—
	MTOR	—	—
	CYP1A1	acacetin, tiabendazole, lansoprazole	riluzole (PMID: 21095567), chloroquine (PMID: 23122960), galangin (PMID: 21406604)

Note: Asterisks denote unexpectedly predicted targets involved in the biological process of the assay. “—” denotes no specific hit or false negative result for the target, respectively. The drugs that are labeled in bold are known to influence the phenotype and the according PubMed IDs are given in brackets.

6.2 Discussion

The pure computational methods for target identification have been exploited to predict previously unknown targets for drugs [73, 96, 98]. Here, I propose a new application of these methods in combination with a statistical approach to predict

proteins modulated by organic molecules influencing phenotypic readouts. This is a simple and inexpensive strategy to establish new connections between proteins and phenotypes such as diseases as well as propose novel mechanism of action of hits.

For example, I have found the intriguing connection between FLT3, APP and “Modulators of lipid transfer” process. FLT3 is a gene frequently mutated in acute myeloid leukemia [202] and it is important for lymphocyte (B cell and T cell) development. It has been recently found that ligands of PPAR γ , such as rosiglitazone, are required for regulation of FLT3 that increases the proliferation of hematopoietic stem cells [203] where low levels of cell membrane cholesterol were observed [204]. In addition, it has been shown that the mutation of FLT3 exhibited a 1.5-fold decrease in their plasma HDL-cholesterol levels [205]. In my approach, I have found that ligands of FLT3 modulate lipid transfer mediated by the SCARB1. All this evidence suggests the connection between FLT3 and lipid metabolism. As for APP, recently strong evidence has shown that APP decreases the quantity of cholesterol [206], which then increases the expression of SCARB1 in brain [207], highlighting the role of APP in the lipid metabolism.

In the “Modulators of PGC-1 α expression” project, although the induction of PGC-1 α by NQO2, CYP1A1 and CYP1B1 has not been reported in the literature, indirect evidence shows the functional relationship of these proteins with PGC-1 α . For instance, the ligand of NQO2, melatonin, regulates the activity of CREB1 [208], the PGC-1 α transcription factor. Furthermore, it has been shown that melatonin plays a crucial role in the regulation of rhythmic clock gene expression [209] and PGC-1 α integrates the mammalian clock and the energy metabolism [210], which indicates a possible connection between the target of melatonin and PGC-1 α . The remaining two targets, CYP1A1 and CYP1B1, are both involved in the metabolism of arachidonic acid [211] that is released by PLA2G4A. Besides both CYP1B1 [212] and PLA2G4A [213] affect the development of glaucoma that is caused by mitochondrial dysfunction [212, 214]. Therefore, the regulation of PGC-1 α expression by NQO2, CYP1A1 and CYP1B1 definitely deserves more investigation.

There are two known common issues of chemical screening hits that my approach is limited: (i) The appearance of false negatives is a general problem in the screening that is commonly explained by the degradation of compounds on screening plates, limited compound purity or concentrations [54]. Inglese et

al. [215] quantitatively enumerated the frequency of false negatives from a traditional single-concentration screen and observed that 40% of the actives were scored as false negatives when the 11 μM screening concentration and a three standard deviation were used. In this regard, I noticed that some compounds part of the “Inactive compounds” set have activity on the significant targets and are reported to have activity on the biological process as well. For example, riluzole [216], chloroquine [217], galangin [218] are all the regulators of Wnt signaling (False negatives of the predicted targets see Table 6.1.1) which failed to be captured by my hit identification method. (ii) The second common problem of chemical screening assays is the presence of nonspecific or promiscuous compounds. The nonspecific or promiscuous compounds are the major source of false positives that act non-competitively on the targets [148], leading to that some of the identified targets might have pleotropic activities leading to adverse reactions and not be optimal to be considered as drug targets, or some compounds might have additional drug targets causing toxic effects. In order to explain the molecular activity of specific hits as much as possible, I did not apply any filter to remove the promiscuous compounds. Thus, the biological experiments are still required to validate that the gene product identified actually binds to the small molecule and is associated with the biological process.

In summary, with my approach I was able to confidently explain the molecular activity responsible for the phenotypic effects of around 18% of the hits for which I can derive molecular information. This is a high number considering the scale of current drug target identification methods that are limited to predict targets with already known ligand information. With the rapid increase of drug-target interaction information in public databases, I envisaged a higher coverage of the molecular space related to phenotypes in the near future. Last but not least, the application of this approach to phenotypic assays promise to reveal unexpected connections between drug targets and disease phenotype such as the targets that are found to be related to the modulation of PGC-1 α expression.

6.3 Conclusions

After applying a computational method followed by a statistical approach to three public screens, namely, modulation of lipid transfer, PGC-1 α expression and Wnt signaling, I predicted the association of 26 targets with these phenotypes. I

have validated 23 target-phenotype predicted associations by two different methods. The first one uses previously reported associations, while the second one explores the known activity of specific chemical hits in the screens. Both methods clearly demonstrate the validity of such approach to detect drug targets related to phenotypes. This computational protocol allows me to obtain an overview of druggable molecular repertoire behind the phenotype and to propose novel associations between targets and biological activities.

Chapter 7

Summary and Outlook

Phenotypic chemical genetic screens use small molecules as tools to perturb biological systems with the aim to investigate cellular pathways and identify key protein targets underlying cellular processes [22]. Chemical genetic approaches have been applied to discover novel human therapeutics in many disease areas, such as cancer research [22], stem cell biology [219] and cell death [220]. However, the analysis of these assays is currently challenging due to the limitations of available techniques. In this present thesis I have developed tools to facilitate the analysis of these screens that overcome the current technical and conceptual challenges of chemical biology approaches as well as applied these tools to existing chemical screens to extract novel biological information. In the following sections I summarize the main scientific contributions of this present thesis and discuss the possible extensions and future directions.

7.1 Scientific achievements

Throughout the work of this thesis, the following novel scientific contributions and insights were achieved:

- In order to develop new fast and efficient methods to identify chemical hits of chemical genetic screens, I have compared eight different chemical hit identification approaches to determine the method best suitable to retrieve chemical hits and tested in the ChemBank dataset repository. The best performing method, the modification of B-Score_A, showed 79.6% sensitivity and 97.4% specificity when applied to the positive and negative controls of

all ChemBank assays.

- Another critical issue regarding chemical screening analysis that is addressed here is the insufficient number of easy-to-use online tools for drug target predictive methods. Although both chemical similarity and ligand-based predictive modeling are well-established approaches to classify the compounds and predict the protein targets of small molecules in computational chemistry, only few of them are implemented as easy-to-use online tools. In the present thesis, two of these two approaches are combined into the ligand-based target identification method – HitPick, which was applied in this thesis to provide insights into the mechanisms of action of small molecules. Relying on the ligand-protein interactions from the STITCH 3 database [99], HitPick target prediction first searches the most similar compound to a query compound by 1-nearest-neighbor (1NN) similarity searching [97], and then predicts the targets based on the Laplacian-modified naïve Bayesian target models [98]. On cross-validation, HitPick target prediction performs better than 1NN similarity searching and Bayesian target models methods separately, achieving 60.94% sensitivity, 99.99% specificity and 92.11% precision. To facilitate the analysis of chemical genetic screens the well-known B-Score [61] chemical hit identification method is also implemented in HitPick along with this newly developed approach to predict targets of small molecules. In summary, HitPick can be used to identify hits from chemical screens and predict new ligand-target interactions. HitPick web server can be accessed through <http://mips.helmholtz-muenchen.de/proj/hitpick>.
- The next crucial issue I tackled is the determination of the specificity of the activity of hits of chemical genetic screens. It has been shown that a great number of compounds tend to be promiscuous, that is, the small molecules appear frequently as hits of many assays via interfering with the assay signals or chemically binding to the tested targets of the assay, etc. The promiscuous activity and hence, low selective activity of compounds could translate into toxic effects when applied to complex living systems such as mouse models or humans. Therefore, to select the right level of selectivity to detect specific signals of chemical hits is of the highest relevance. In the present thesis, a computational promiscuous filtering was proposed to detect the selective chemical hits.

- The integration and analysis of chemical screening assays stored in public repositories have proven to be useful so far to determine chemical properties of promiscuous compounds and to predict adverse drug reactions, demonstrating the potential of computational analysis of high-throughput chemical screening to extract novel chemical and biological information. Here, I tested the hypothesis of whether the bioactivities measured in two assays sharing selective chemical hits are related and used the publicly available chemical screening repository ChemBank to validate this hypothesis. Besides, the biological associations between pairs of phenotypic assays are reinforced by the analysis of the biological roles of the predicted molecular targets of shared hits. Furthermore, the analysis of these targets help to better understand the molecular basis of assay relationships. I show that the systematic comparison of the selective hits of chemical screening assays is a promising approach to uncover relationships between biological activities.
- As cell- or organism-based screens preserve native cellular environment of protein function, these phenotypic assays are increasingly used in applications aiming at discovery of new therapeutic targets and new disease biology [57,68]. However, the cost paid for such benefit is that protein targets and mechanism of action responsible for the observed phenotype needed to be determined. In this thesis, I developed a method to detect targets reliably associated to the phenotypic assays based on HitPick target prediction followed by a statistical approach. Strikingly, it was found that 88% of the predicted drug targets are reported to be associated with the measured phenotype in three different phenotypic assays, namely, modulation of lipid transfer, PGC-1 α expression and Wnt signaling, demonstrating that this computational method allows to confidently relate the protein targets to the observed phenotype, and to discover novel molecular mechanisms of action of the chemical hits.

In summary, I have created a variety of powerful tools for tracking and analyzing chemical screening data. These tools are particularly well suited to chemical genetic screens because they offer new insights in identifying chemical hits, linking different assays and associating protein targets with different phenotypes. Especially, the last two above findings of this work demonstrate how powerful the

ability to systematically and effectively integrate chemical genetic screens becomes in understanding the mechanisms of action of small molecules in biological systems. When done in a disciplined and thoughtful manner, the integration of HTS data represents a modern and inexpensive approach to provide insights and clues to novel targets and molecular mechanisms of small chemicals.

7.2 Final conclusions

In this thesis, several approaches to facilitate the analysis of chemical genetic assays have been proposed. An optimal method to retrieve chemical hits from assays was firstly introduced, later a target prediction method was developed to predict the drug targets for the hits. Furthermore, an efficient filtering protocol was proposed to remove the promiscuous compounds. Lastly, the integration of the assays was presented as a valuable tool to find relationships between biological activities, to understand better the molecular mechanisms of the chemical hits, and also to relate drug targets to phenotypes.

7.3 Extensions and future directions

In this thesis, I mainly worked on the analysis of assays stored in ChemBank. PubChem BioAssay [136] is another well-known publically available assay repository that, compared to ChemBank, contains much more data. However, several peculiarities of this database such as the data structure (see Table 7.3.1) make the analysis and integration of the information of this repository fairly complicated.

First of all, it is hard to compare the activity results between different assays due to the inaccessibility of the raw activity values of assays of PubChem and the fact that each assay depositor may use different method to identify the chemical hits. Besides, a preliminary analysis of the assays reveals a high heterogeneous structure of PubChem BioAssay, where projects often comprised assays developed by different research groups that impede the identification of experiment and control assays, complicating the definition of the specific and unspecific assay connections.

To cope with the above-mentioned difficulties, it would be desired to set up collaboration with some of the experimental depositors to get access to the original

Table 7.3.1: Differences between ChemBank and PubChem BioAssay repositories

Method	ChemBank	PubChem BioAssay
Screening data	It stores raw screening data (in total 1,640 primary assays are included).	The screening outcome, such as the bioactivity summary, structure clustering, etc., is available. However, the raw values of HTS experiments are inaccessible. The primary, secondary/confirmatory/counter assays are included (in total 3,295 assays).
Assay description	The assay description is plain and brief, but rigorous.	The lengthy assay description needs to be extracted by users from the protocols.
Projects	The assays of the same experiment are hierarchically organized into screening projects.	According to the goal of the assay, depositors classify the assays into screening projects.

bioassay data and obtain a better understanding of the chemical screens. In the framework of this collaboration, the hit identification method that I developed could be applied to their raw assay data, allowing the comparison of the retrieved hits from different groups and those from ChemBank assays. Furthermore, the knowledge of the experimentalists about the assay protocols will be helpful for the correct classification of the assays into experiment and control assays. Then, the methods that I developed in this thesis could be applied to integrate and analyze these assays and gain further insights into the relationships of their assays to other assays, and also into the molecular activity of specific hits. However, based on my experience and contact with depositors, few depositors would be willing to disclose the assay data not only due to data sensitivity issues but also to the high workload needed to prepare the raw assay data files. For these reasons, the collaboration with the assay depositors is challenging.

Another interesting extension of this thesis will be the possibility to compare the chemical and biological properties of compounds such as the promiscuity and biological activity of hits obtained from the two different repositories. This will reinforce the validity of the methods that rely on the integration and analysis of chemical screening repositories to extract novel chemical and biological informa-

7. SUMMARY AND OUTLOOK

tion of small molecules.

Bibliography

- [1] D. R. Spring. Chemical genetics to chemical genomics: small molecules offer big insights. *Chemical society reviews* **2005**, *34*, 472–482.
- [2] P. Iengar. An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic acids research* **2012**, *40*, 6401–6413.
- [3] F.-T. Kao, T. T. Puck. Genetics of somatic mammalian cells. IX. Quantitation of mutagenesis by physical and chemical agents. *Journal of cellular physiology* **1969**, *74*, 245–257.
- [4] L. Hartwell, L. Hood, M. L. Goldberg, A. E. Reynolds, L. M. Silver, R. C. Veres, *Genetics: from genes to genomes*, McGraw-Hill Higher Education Boston, **2004**.
- [5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, et al., *Studying gene expression and function*, Garland Science, **2002**.
- [6] P. Vallance, T. G. Smart. The future of pharmacology. *British journal of pharmacology* **2006**, *147*, S304–S307.
- [7] M. Kawasumi, P. Nghiem. Chemical genetics: elucidating biological systems with small-molecule compounds. *Journal of investigative dermatology* **2007**, *127*, 1577–1584.
- [8] A. Fleming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae*. *British journal of experimental pathology* **1929**, *10*, 226.
- [9] J. Lederberg. Mechanism of action of penicillin. *Journal of bacteriology* **1957**, *73*, 144.

- [10] Z. A. Knight, K. M. Shokat. Chemical genetics: where genetics and pharmacology meet. *Cell* **2007**, *128*, 425–430.
- [11] J. M. Lehmann, L. B. Moore, T. A. Smith-Oliver, W. O. Wilkison, T. M. Willson, S. A. Kliewer. An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ (PPAR γ). *Journal of biological chemistry* **1995**, *270*, 12953–12956.
- [12] K. Tureyen, R. Kapadia, K. K. Bowen, I. Satriotomo, J. Liang, D. L. Feinstein, R. Vemuganti. Peroxisome proliferator-activated receptor- γ agonists induce neuroprotection following transient focal ischemia in normotensive, normoglycemic as well as hypertensive and type-2 diabetic rodents. *Journal of neurochemistry* **2007**, *101*, 41–56.
- [13] N. Kubota, Y. Terauchi, H. Miki, H. Tamemoto, T. Yamauchi, K. Komeda, S. Satoh, R. Nakano, C. Ishii, T. Sugiyama, et al.. PPAR γ mediates high-fat diet-induced adipocyte hypertrophy and insulin resistance. *Molecular cell* **1999**, *4*, 597–609.
- [14] T. Yamauchi, H. Waki, J. Kamon, K. Murakami, K. Motojima, K. Komeda, H. Miki, N. Kubota, Y. Terauchi, A. Tsuchida, et al.. Inhibition of RXR and PPAR γ ameliorates diet-induced obesity and type 2 diabetes. *Journal of clinical investigation* **2001**, *108*, 1001–1013.
- [15] X. S. Zheng, T.-F. Chan. Chemical genomics: a systematic approach in biological research and drug discovery. *Current issues in molecular biology* **2002**, *4*, 33–44.
- [16] B. R. Stockwell. Frontiers in chemical genetics. *Trends in biotechnology* **2000**, *18*, 449–455.
- [17] B. R. Stockwell. Chemical genetics: ligand-based discovery of gene function. *Nature reviews genetics* **2000**, *1*, 116–125.
- [18] J. Lehár, B. R. Stockwell, G. Giaever, C. Nislow. Combination chemical genetics. *Nature chemical biology* **2008**, *4*, 674–681.
- [19] A. J. Vegas, J. H. Fuller, A. N. Koehler. Small-molecule microarrays as tools in ligand discovery. *Chemical society reviews* **2008**, *37*, 1385–1394.

-
- [20] T. U. Mayer. Chemical genetics: tailoring tools for cell biology. *Trends in cell biology* **2003**, *13*, 270–277.
- [21] R. D. Klausner, J. G. Donaldson, J. Lippincott-Schwartz. Brefeldin A: insights into the control of membrane traffic and organelle structure. *The Journal of cell biology* **1992**, *116*, 1071–1080.
- [22] N. Tolliday, P. A. Clemons, P. Ferraiolo, A. N. Koehler, T. A. Lewis, X. Li, S. L. Schreiber, D. S. Gerhard, S. Eliasof. Small molecules, big players: the National Cancer Institute’s initiative for chemical genetics. *Cancer research* **2006**, *66*, 8935–8942.
- [23] S. J. Haggarty, K. M. Koeller, J. C. Wong, C. M. Grozinger, S. L. Schreiber. Domain-selective small-molecule inhibitor of histone deacetylase 6 (HDAC6)-mediated tubulin deacetylation. *Proceedings of the national academy of sciences* **2003**, *100*, 4389–4394.
- [24] J. R. Sharom, D. S. Bellows, M. Tyers. From large networks to small molecules. *Current opinion in chemical biology* **2004**, *8*, 81–90.
- [25] J. Lehár, G. R. Zimmermann, A. S. Krueger, R. A. Molnar, J. T. Ledell, A. M. Heilbut, G. F. Short, L. C. Giusti, G. P. Nolan, O. A. Magid, et al.. Chemical combination effects predict connectivity in biological systems. *Molecular systems biology* **2007**, *3*.
- [26] J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li, Y. Z. Chen. Mechanisms of drug combinations: interaction and network perspectives. *Nature reviews drug discovery* **2009**, *8*, 111–128.
- [27] I. Smukste, B. R. Stockwell. Advances in chemical genetics. *Annual review of genomics and human genetics* **2005**, *6*, 261–286.
- [28] C. R. Geyer, A. Colman-Lerner, R. Brent. ”Mutagenesis” by peptide aptamers identifies genetic network members and pathway connections. *Proceedings of the national academy of sciences* **1999**, *96*, 8567–8572.
- [29] P. Colas. Combinatorial protein reagents to manipulate protein function. *Current opinion in chemical biology* **2000**, *4*, 54–59.

- [30] Y.-Z. Shu. Recent natural products based drug development: a pharmaceutical industry perspective. *Journal of natural products* **1998**, *61*, 1053–1071.
- [31] L. Di, E. H. Kerns. Profiling drug-like properties in discovery research. *Current opinion in chemical biology* **2003**, *7*, 402–408.
- [32] S. L. Schreiber. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287*, 1964–1969.
- [33] S. Young, N. Ge. Design of diversity and focused combinatorial libraries in drug discovery. *Current opinion in drug discovery & development* **2004**, *7*, 318–324.
- [34] B. R. Stockwell. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846–854.
- [35] R. C. Reid, L. K. Pattenden, J. D. Tyndall, J. L. Martin, T. Walsh, D. P. Fairlie. Countering cooperative effects in protease inhibitors using constrained β -strand-mimicking templates in focused combinatorial libraries. *Journal of medicinal chemistry* **2004**, *47*, 1641–1651.
- [36] F. L. Stahura, L. Xue, J. W. Godden, J. Bajorath. Molecular scaffold-based design and comparison of combinatorial libraries focused on the ATP-binding site of protein kinases. *Journal of molecular graphics and modelling* **1999**, *17*, 1–52.
- [37] M. Sodeoka, R. Sampe, S. Kojima, Y. Baba, T. Usui, K. Ueda, H. Osada. Synthesis of a tetronic acid library focused on inhibitors of tyrosine and dual-specificity protein phosphatases and its evaluation regarding VHR and cdc25B inhibition. *Journal of medicinal chemistry* **2001**, *44*, 3216–3222.
- [38] P. Jimonet, R. Jäger. Strategies for designing GPCR-focused libraries and screening sets. *Current opinion in drug discovery & development* **2004**, *7*, 325–333.
- [39] C. John Harris, R. D Hill, D. W Sheppard, M. J Slater, P. FW Stouten. The design and application of target-focused compound libraries. *Combinatorial chemistry & high throughput screening* **2011**, *14*, 521–531.
- [40] Screening we can believe in. *Nature chemical biology* **2009**, *5*, 127.

-
- [41] E. J. Gordon. Small-molecule screening: it takes a village... *ACS chemical biology* **2007**, *2*, 9.
- [42] J. Inglese, R. L. Johnson, A. Simeonov, M. Xia, W. Zheng, C. P. Austin, D. S. Auld. High-throughput screening assays for the identification of chemical probes. *Nature chemical biology* **2007**, *3*, 466–479.
- [43] R. P. Hertzberg, A. J. Pope. High-throughput screening: new technology for the 21st century. *Current opinion in chemical biology* **2000**, *4*, 445–451.
- [44] J. J. Burbaum. The evolution of miniaturized well plates. *Journal of biomolecular screening* **2000**, *5*, 5–8.
- [45] L. Mere, T. Bennett, P. Coassin, P. England, B. Hamman, T. Rink, S. Zimmerman, P. Negulescu. Miniaturized FRET assays and microfluidics: key components for ultra-high-throughput screening. *Drug discovery today* **1999**, *4*, 363–369.
- [46] B. R. Stockwell, S. J. Haggarty, S. L. Schreiber. High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. *Chemistry & biology* **1999**, *6*, 71–83.
- [47] P. Gut, B. Baeza-Raja, O. Andersson, L. Hasenkamp, J. Hsiao, D. Hesselson, K. Akassoglou, E. Verdin, M. D. Hirschey, D. Y. Stainier. Whole-organism screening for gluconeogenesis identifies activators of fasting metabolism. *Nature chemical biology* **2013**, *9*, 97–104.
- [48] M. V. Boland, M. K. Markey, R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* **1998**, *33*, 366–375.
- [49] R. T. Peterson, B. A. Link, J. E. Dowling, S. L. Schreiber. Small molecule developmental screens reveal the logic and timing of vertebrate development. *Proceedings of the national academy of sciences* **2000**, *97*, 12965–12969.
- [50] G. R. Rosania, Y.-T. Chang, O. Perez, D. Sutherlin, H. Dong, D. J. Lockhart, P. G. Schultz. Myoseverin, a microtubule-binding molecule with novel cellular effects. *Nature biotechnology* **2000**, *18*, 304–308.

- [51] H. E. Pelish, N. J. Westwood, Y. Feng, T. Kirchhausen, M. D. Shair. Use of biomimetic diversity-oriented synthesis to discover galanthamine-like molecules with biological properties beyond those of the natural product. *Journal of the American chemical society* **2001**, *123*, 6740–6741.
- [52] D. Yu, T. Jing, B. Liu, J. Yao, M. Tan, T. J. McDonnell, M.-C. Hung. Overexpression of ErbB2 blocks Taxol-induced apoptosis by upregulation of p21^{Cip1}, which inhibits p34^{Cdc2} kinase. *Molecular cell* **1998**, *2*, 581–591.
- [53] P. Gribbon, A. Sewing. High-throughput drug discovery: what can we expect from HTS? *Drug discovery today* **2005**, *10*, 17–22.
- [54] J. Bajorath. Integration of virtual and high-throughput screening. *Nature reviews drug discovery* **2002**, *1*, 882–894.
- [55] U. S. Eggert. The why and how of phenotypic small-molecule screens. *Nature chemical biology* **2013**, *9*, 206–209.
- [56] P. Dorr, M. Westby, S. Dobbs, P. Griffin, B. Irvine, M. Macartney, J. Mori, G. Rickett, C. Smith-Burchnell, C. Napier, et al.. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrobial agents and chemotherapy* **2005**, *49*, 4721–4732.
- [57] W. Zheng, N. Thorne, J. C. McKew. Phenotypic screens as a renewed approach for drug discovery. *Drug discovery today* **2013**, *18*, 1067–1073.
- [58] J. Kotz. Phenotypic screening, take two. *SciBX: Science-Business eXchange* **2012**, *5*.
- [59] D. C. Swinney, J. Anthony. How were new medicines discovered? *Nature reviews drug discovery* **2011**, *10*, 507–519.
- [60] K. J. Duffy, M. G. Darcy, E. Delorme, S. B. Dillon, D. F. Eppley, C. Erickson-Miller, L. Giampa, C. B. Hopson, Y. Huang, R. M. Keenan, et al.. Hydrazinonaphthalene and azonaphthalene thrombopoietin mimics are nonpeptidyl promoters of megakaryocytopoiesis. *Journal of medicinal chemistry* **2001**, *44*, 3730–3745.

-
- [61] N. Malo, J. A. Hanley, S. Cerquozzi, J. Pelletier, R. Nadon. Statistical practice in high-throughput screening data analysis. *Nature biotechnology* **2006**, *24*, 167–175.
- [62] G. M. Keserű, G. M. Makara. Hit discovery and hit-to-lead approaches. *Drug discovery today* **2006**, *11*, 741–748.
- [63] R. W. Spencer. High-throughput screening of historic collections: Observations on file size, biological targets, and file diversity. *Biotechnology and bioengineering* **1998**, *61*, 61–67.
- [64] H. Strobelt, E. Bertini, J. Braun, O. Deussen, U. Groth, T. U. Mayer, D. Merhof. HiTSEE KNIME: a visualization tool for hit selection and analysis in high-throughput screening experiments for the KNIME platform. *BMC bioinformatics* **2012**, *13*, S4.
- [65] K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, et al.. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic acids research* **2008**, *36*, D351–D359.
- [66] V. Makarenkov, P. Zentilli, D. Kevorkov, A. Gagarin, N. Malo, R. Nadon. An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* **2007**, *23*, 1648–1657.
- [67] B. M. Silber. Driving drug discovery: the fundamental role of academic labs. *Science translational medicine* **2010**, *2*, 30cm16–30cm16.
- [68] M. Schenone, V. Dančik, B. K. Wagner, P. A. Clemons. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology* **2013**, *9*, 232–240.
- [69] L. Burdine, T. Kodadek. Target identification in chemical genetics. *Chemistry & biology* **2004**, *11*, 593–597.
- [70] D. S. Bellows, M. Tyers. Chemical genetics hits "reality". *Science* **2004**, *306*, 67–68.

- [71] U. H. Weidle, D. Maisel, D. Eick. Synthetic lethality-based targets for discovery of new cancer therapeutics. *Cancer genomics-proteomics* **2011**, *8*, 159–171.
- [72] F. L. Muller, S. Colla, E. Aquilanti, V. E. Manzo, G. Genovese, J. Lee, D. Eisensohn, R. Narurkar, P. Deng, L. Nezi, et al.. Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* **2012**, *488*, 337–342.
- [73] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature biotechnology* **2007**, *25*, 197–206.
- [74] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijter, R. C. Matos, T. B. Tran, et al.. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.
- [75] E. Gregori-Puigjané, V. Setola, J. Hert, B. A. Crews, J. J. Irwin, E. Lounkine, L. Marnett, B. L. Roth, B. K. Shoichet. Identifying mechanism-of-action targets for drugs and probes. *Proceedings of the national academy of sciences* **2012**, *109*, 11178–11183.
- [76] C. Laggner, D. Kokel, V. Setola, A. Tolia, H. Lin, J. J. Irwin, M. J. Keiser, C. Y. J. Cheung, D. L. Minor Jr, B. L. Roth, et al.. Chemical informatics and target identification in a zebrafish phenotypic screen. *Nature chemical biology* **2012**, *8*, 144–146.
- [77] W. P. Walters, A. A. Murcko, M. A. Murcko. Recognizing molecules with drug-like properties. *Current opinion in chemical biology* **1999**, *3*, 384–387.
- [78] J. R. Huth, C. Sun, D. R. Sauer, P. J. Hajduk. Utilization of NMR-derived fragment leads in drug design. *Methods in enzymology* **2005**, *394*, 549–571.
- [79] G. M. Rishton. Reactive compounds and in vitro false positives in HTS. *Drug discovery today* **1997**, *2*, 382–384.
- [80] S. L. McGovern, E. Caselli, N. Grigorieff, B. K. Shoichet. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *Journal of medicinal chemistry* **2002**, *45*, 1712–1722.

-
- [81] J. B. Baell, G. A. Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry* **2010**, *53*, 2719–2740.
- [82] F.-J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J.-L. Lavandera, D. E. Vanderwall, D. V. Green, V. Kumar, S. Hasan, et al.. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465*, 305–310.
- [83] K. H. Bleicher, H.-J. Böhm, K. Müller, A. I. Alanine. Hit and lead generation: beyond high-throughput screening. *Nature reviews drug discovery* **2003**, *2*, 369–378.
- [84] H. Kitano. Computational systems biology. *Nature* **2002**, *420*, 206–210.
- [85] E. C. Butcher, E. L. Berg, E. J. Kunkel. Systems biology in drug discovery. *Nature biotechnology* **2004**, *22*, 1253–1259.
- [86] J. K. Nicholson, I. D. Wilson. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nature reviews drug discovery* **2003**, *2*, 668–676.
- [87] H. Kitano. Systems biology: a brief overview. *Science* **2002**, *295*, 1662–1664.
- [88] S. Dibyajyoti, E. T. Bin, P. Swati. Bioinformatics: The effects on the cost of drug discovery. *Galle medical journal* **2013**, *18*, 44–50.
- [89] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, et al.. Functional discovery via a compendium of expression profiles. *Cell* **2000**, *102*, 109–126.
- [90] W. Chavuenet, *A manual of spherical and practical astronomy*, **1871**.
- [91] J. W. Tukey. Exploratory data analysis. **1977**.
- [92] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters* **2006**, *27*, 861–874.
- [93] E. K. Wagner, J. Ramirez, S. Stingley, S. Aguilar, L. Buehler, G. Devi-Rao, P. Ghazal. Practical approaches to long oligonucleotide-based DNA

- microarray: lessons from herpesviruses. *Progress in nucleic acid research and molecular biology* **2002**, *71*, 445.
- [94] T. Forster, D. Roy, P. Ghazal. Experiments using microarray technology: limitations and standard operating procedures. *Journal of endocrinology* **2003**, *178*, 195–204.
- [95] A. Birmingham, L. M. Selfors, T. Forster, D. Wrobel, C. J. Kennedy, E. Shanks, J. Santoyo-Lopez, D. J. Dunican, A. Long, D. Kelleher, et al.. Statistical methods for analysis of high-throughput RNA interference screens. *Nature methods* **2009**, *6*, 569–575.
- [96] X. Liu, I. Vogt, T. Haque, M. Campillos. HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics* **2013**, *29*, 1910–1912.
- [97] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby. Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of chemical information and computer sciences* **2003**, *43*, 391–405.
- [98] Nidhi, M. Glick, J. W. Davies, J. L. Jenkins. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *Journal of chemical information and modeling* **2006**, *46*, 1124–1133.
- [99] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, P. Bork. STITCH 3: zooming in on protein–chemical interactions. *Nucleic acids research* **2012**, *40*, D876–D880.
- [100] P. Willett. Chemoinformatics–similarity and diversity in chemical libraries. *Current opinion in biotechnology* **2000**, *11*, 85–88.
- [101] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *Journal of chemical information and computer sciences* **2003**, *43*, 1218–1225.
- [102] R. D. Brown. Descriptors for diversity analysis. *Perspectives in drug discovery and design* **1997**, *7*, 1–31.

-
- [103] G. M. Downs, J. M. Barnard. Techniques for generating descriptive fingerprints in combinatorial libraries §. *Journal of chemical information and modeling* **1997**, *37*, 59–61.
- [104] H. Morgan. The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. *Journal of chemical documentation* **1965**, *5*, 107–113.
- [105] W. P. Walters, M. T. Stahl, M. A. Murcko. Virtual screening—an overview. *Drug discovery today* **1998**, *3*, 160–178.
- [106] B. Waszkowycz, T. D. J. Perkins, R. A. Sykes, J. Li. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM systems journal* **2001**, *40*, 360–376.
- [107] M. A. Miller. Chemical database techniques in drug discovery. *Nature reviews drug discovery* **2002**, *1*, 220–227.
- [108] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, L. E. Weinberger. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of medicinal chemistry* **1996**, *39*, 3049–3059.
- [109] R. D. Brown, Y. C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of chemical information and computer sciences* **1996**, *36*, 572–584.
- [110] R. D. Brown, Y. C. Martin. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *Journal of chemical information and computer Sciences* **1997**, *37*, 1–9.
- [111] D. K. Agrafiotis. Diversity of chemical libraries. *Encyclopedia of computational chemistry* **1998**.
- [112] P. Willett, J. M. Barnard, G. M. Downs. Chemical similarity searching. *Journal of chemical information and computer sciences* **1998**, *38*, 983–996.
- [113] R. P. Sheridan, S. K. Kearsley. Why do we need so many chemical similarity search methods? *Drug discovery today* **2002**, *7*, 903–911.

- [114] N. Nikolova, J. Jaworska. Approaches to measure chemical similarity—a review. *QSAR & combinatorial science* **2003**, *22*, 1006–1026.
- [115] A. Bender, R. C. Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry* **2004**, *2*, 3204–3218.
- [116] X. Xia, E. G. Maliski, P. Gallant, D. Rogers. Classification of kinase inhibitors using a Bayesian model. *Journal of medicinal chemistry* **2004**, *47*, 4463–4470.
- [117] H. Y. Mussa, J. B. Mitchell, R. C. Glen. Full "Laplacianised" posterior naive Bayesian algorithm. *Journal of cheminformatics* **2013**, *5*, 37.
- [118] J. A. Townsend, R. C. Glen, H. Y. Mussa. Note on naive Bayes based on binary descriptors in Cheminformatics. *Journal of chemical information and modeling* **2012**, *52*, 2494–2500.
- [119] F. Nigsch, A. Bender, J. L. Jenkins, J. B. Mitchell. Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *Journal of chemical information and modeling* **2008**, *48*, 2313–2325.
- [120] F. Martínez-Jiménez, G. Papadatos, L. Yang, I. M. Wallace, V. Kumar, U. Pieper, A. Sali, J. R. Brown, J. P. Overington, M. A. Marti-Renom. Target prediction for an open access set of compounds active against *Mycobacterium tuberculosis*. *PLoS computational biology* **2013**, *9*, e1003253.
- [121] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *Journal of chemical information and modeling* **2006**, *46*, 462–470.
- [122] M. Ashton, J. Barnard, F. Casset, M. Charlton, G. Downs, D. Gorse, J. Holliday, R. Lahana, P. Willett. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative structure-activity relationships* **2002**, *21*, 598–604.
- [123] H. W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K. F. Mayer, V. Stümpflen, et al.. MIPS: curated

- databases and comprehensive secondary data resources in 2010. *Nucleic acids research* **2011**, *39*, D220–D224.
- [124] R. D. C. Team. R: A language and environment for statistical computing. *R foundation for statistical computing* **2005**.
- [125] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, I. Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **2001**, *125*, 279–284.
- [126] P. Dragiev, R. Nadon, V. Makarenkov. Systematic error detection in experimental high-throughput screening. *BMC bioinformatics* **2011**, *12*, 25.
- [127] D. Kevorkov, V. Makarenkov. Statistical analysis of systematic errors in high-throughput screening. *Journal of biomolecular screening* **2005**, *10*, 557–567.
- [128] C. Brideau, B. Gunter, B. Pikounis, A. Liaw. Improved statistical methods for hit selection in high-throughput screening. *Journal of biomolecular screening* **2003**, *8*, 634–647.
- [129] C. Ahlberg. Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug discovery today* **1999**, *4*, 370–376.
- [130] M. F. Engels, L. Wouters, R. Verbeeck, G. Vanhoof. Outlier mining in high throughput screening experiments. *Journal of biomolecular screening* **2002**, *7*, 341–351.
- [131] B. Gunter, C. Brideau, B. Pikounis, A. Liaw. Statistical and graphical methods for quality control determination of high-throughput screening data. *Journal of biomolecular screening* **2003**, *8*, 624–633.
- [132] B. P. Kelley, M. R. Lunn, D. E. Root, S. P. Flaherty, A. M. Martino, B. R. Stockwell. A flexible data analysis tool for chemical genetic screens. *Chemistry & biology* **2004**, *11*, 1495–1503.
- [133] D. E. Root, B. P. Kelley, B. R. Stockwell. Detecting spatial patterns in biological array experiments. *Journal of biomolecular screening* **2003**, *8*, 393–398.

- [134] S. Amberkar, N. A. Kiani, R. Bartenschlager, G. Alvisi, L. Kaderali. High-throughput RNA interference screens integrative analysis: Towards a comprehensive understanding of the virus-host interplay. *World journal of virology* **2013**, *2*, 18.
- [135] V. Makarenkov, D. Kevorkov, P. Zentilli, A. Gagarin, N. Malo, R. Nadon. HTS-Corrector: software for the statistical analysis and correction of experimental high-throughput screening data. *Bioinformatics* **2006**, *22*, 1408–1409.
- [136] Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B. A. Shoemaker, T. O. Suzek, J. Wang, J. Xiao, J. Zhang, S. H. Bryant. An overview of the PubChem BioAssay resource. *Nucleic acids research* **2010**, *38*, D255–D266.
- [137] D. W. Young, A. Bender, J. Hoyt, E. McWhinnie, G.-W. Chirn, C. Y. Tao, J. A. Tallarico, M. Labow, J. L. Jenkins, T. J. Mitchison, et al.. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature chemical biology* **2008**, *4*, 59–68.
- [138] M. Kuhn, M. Campillos, P. González, L. J. Jensen, P. Bork. Large-scale prediction of drug–target relationships. *FEBS letters* **2008**, *582*, 1283–1290.
- [139] J.-C. Wang, P.-Y. Chu, C.-M. Chen, J.-H. Lin. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic acids research* **2012**, *40*, W393–W399.
- [140] R. H. Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews cancer* **2006**, *6*, 813–823.
- [141] C. P. Austin, L. S. Brady, T. R. Insel, F. S. Collins. NIH molecular libraries initiative. *Science* **2004**, *306*, 1138–9.
- [142] S. A. Canny, Y. Cruz, M. R. Southern, P. R. Griffin. PubChem promiscuity: a web resource for gathering compound promiscuity data from PubChem. *Bioinformatics* **2012**, *28*, 140–141.
- [143] B. Chen, D. Wild, R. Guha. PubChem as a source of polypharmacology. *Journal of chemical information and modeling* **2009**, *49*, 2044–2055.

-
- [144] S. C. Schürer, U. Vempati, R. Smith, M. Southern, V. Lemmon. BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *Journal of biomolecular screening* **2011**, *16*, 415–426.
- [145] Y. Pouliot, A. P. Chiang, A. J. Butte. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clinical pharmacology & therapeutics* **2011**, *90*, 90–99.
- [146] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al.. Gene Ontology: tool for the unification of biology. *Nature genetics* **2000**, *25*, 25–29.
- [147] D. Lin. An information-theoretic definition of similarity. *Proceedings of the 15th international conference on machine learning* **1998**, *98*, 296–304.
- [148] B. Y. Feng, A. Shelat, T. N. Doman, R. K. Guy, B. K. Shoichet. High-throughput assays for promiscuous inhibitors. *Nature chemical biology* **2005**, *1*, 146–148.
- [149] P. J. Barnes. Anti-inflammatory actions of glucocorticoids: molecular mechanisms. *Clinical science* **1998**, *94*, 557–572.
- [150] P. W. Cook, K. T. Swanson, C. P. Edwards, G. L. Firestone. Glucocorticoid receptor-dependent inhibition of cellular proliferation in dexamethasone-resistant and hypersensitive rat hepatoma cell variants. *Molecular and cellular biology* **1988**, *8*, 1449–1459.
- [151] X. Xu, J. Hoebeke, P. Björntorp. Progesterin binds to the glucocorticoid receptor and mediates antiglucocorticoid effect in rat adipose precursor cells. *Journal of steroid biochemistry* **1990**, *36*, 465–471.
- [152] C. Hetz. The unfolded protein response: controlling cell fate decisions under ER stress and beyond. *Nature reviews molecular cell biology* **2012**, *13*, 89–102.
- [153] R. A. Newman, P. Yang, A. D. Pawlus, K. I. Block. Cardiac glycosides as novel cancer therapeutic agents. *Molecular interventions* **2008**, *8*, 36.
- [154] K. Gajjar, P. L. Martin-Hirsch, F. L. Martin. CYP1B1 and hormone-induced cancer. *Cancer letters* **2012**, *324*, 13–30.

- [155] D.-F. Ma, T. Kondo, T. Nakazawa, D.-F. Niu, K. Mochizuki, T. Kawasaki, T. Yamane, R. Katoh. Hypoxia-inducible adenosine A2B receptor modulates proliferation of colon carcinoma cells. *Human pathology* **2010**, *41*, 1550–1557.
- [156] J. W. Brewer, J. A. Diehl. PERK mediates cell-cycle exit during the mammalian unfolded protein response. *Proceedings of the national academy of sciences* **2000**, *97*, 12625–12630.
- [157] T. Ozdemir, R. Nar, V. Kilinc, H. Alacam, O. Salis, A. Duzgun, S. Gulden, A. Bedir. Ouabain targets the unfolded protein response for selective killing of HepG2 cells during glucose deprivation. *Cancer biotherapy and radiopharmaceuticals* **2012**, *27*, 457–463.
- [158] J. M. Wagner, B. Hackanson, M. Lübbert, M. Jung. Histone deacetylase (HDAC) inhibitors in recent clinical trials for cancer therapy. *Clinical epigenetics* **2010**, *1*, 117–136.
- [159] E. K. Rowinsky, J. J. Windle, D. D. Von Hoff. Ras protein farnesyltransferase: a strategic target for anticancer therapeutic development. *Journal of clinical oncology* **1999**, *17*, 3631–3652.
- [160] S. Shangary, S. Wang. Targeting the MDM2-p53 interaction for cancer therapy. *Clinical cancer research* **2008**, *14*, 5318–5324.
- [161] L. Madi, A. Ochaion, L. Rath-Wolfson, S. Bar-Yehuda, A. Erlanger, G. Ohana, A. Harish, O. Merimski, F. Barer, P. Fishman. The A3 adenosine receptor is highly expressed in tumor versus normal cells potential target for tumor growth inhibition. *Clinical cancer research* **2004**, *10*, 4472–4479.
- [162] F. Ye, Y. Chen, T. Hoang, R. L. Montgomery, X.-h. Zhao, H. Bu, T. Hu, M. M. Taketo, J. H. van Es, H. Clevers, et al.. HDAC1 and HDAC2 regulate oligodendrocyte differentiation by disrupting the β -catenin–TCF interaction. *Nature neuroscience* **2009**, *12*, 829–838.
- [163] X.-H. Fan, Y.-Y. Cheng, Z.-L. Ye, R.-C. Lin, Z.-Z. Qian. Multiple chromatographic fingerprinting and its application to the quality control of herbal medicines. *Analytica chimica acta* **2006**, *555*, 217–224.

- [164] P. M. Petrone, B. Simms, F. Nigsch, E. Lounkine, P. Kutchukian, A. Cornett, Z. Deng, J. W. Davies, J. L. Jenkins, M. Glick. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS chemical biology* **2012**, *7*, 1399–1409.
- [165] V. Dančák, H. Carrel, N. E. Bodycombe, K. P. Seiler, D. Fomina-Yadlin, S. T. Kubicek, K. Hartwell, A. F. Shamji, B. K. Wagner, P. A. Clemons. Connecting small molecules with similar assay performance profiles leads to new biological hypotheses. *Journal of biomolecular screening* **2014**, 1087057113520226.
- [166] R. T. Jacob, M. J. Larsen, S. D. Larsen, P. D. Kirchhoff, D. H. Sherman, R. R. Neubig. MScreen: an integrated compound management and high-throughput screening data storage and analysis system. *Journal of biomolecular screening* **2012**, *17*, 1080–1087.
- [167] E. Aydar, P. Onganer, R. Perrett, M. B. Djamgoz, C. P. Palmer. The expression and functional characterization of sigma (σ) 1 receptors in breast cancer cell lines. *Cancer letters* **2006**, *242*, 245–257.
- [168] J. S. Cisar, B. F. Cravatt. Fully functionalized small-molecule probes for integrated phenotypic screening and target identification. *Journal of the American chemical society* **2012**, *134*, 10385–10388.
- [169] C. E. Lipscomb. Medical subject headings (MeSH). *Bulletin of the medical library association* **2000**, *88*, 265.
- [170] T. J. Nieland, M. Penman, L. Dori, M. Krieger, T. Kirchhausen. Discovery of chemical inhibitors of the selective transfer of lipids mediated by the HDL receptor SR-BI. *Proceedings of the national academy of sciences* **2002**, *99*, 15422–15427.
- [171] A. R. Subauste, A. K. Das, X. Li, B. Elliot, C. Evans, M. El Azzouny, M. Treutelaar, E. Oral, T. Leff, C. F. Burant. Alterations in lipid signaling underlie lipodystrophy secondary to AGPAT2 mutations. *Diabetes* **2012**, *61*, 2922–2931.

- [172] V. Chandra, P. Huang, Y. Hamuro, S. Raghuram, Y. Wang, T. P. Burris, F. Rastinejad. Structure of the intact PPAR- γ -RXR- α -nuclear receptor complex on DNA. *Nature* **2008**, *456*, 350–356.
- [173] J. Osz, M. V. Pethoukhov, S. Sirigu, D. I. Svergun, D. Moras, N. Rochel. Solution structures of PPAR γ 2/RXR α complexes. *PPAR research* **2012**, *2012*.
- [174] J.-W. Jeong, I. Kwak, K. Y. Lee, L. D. White, X.-P. Wang, F. Brunicaudi, B. W. O'Malley, F. J. DeMayo. The genomic analysis of the impact of steroid receptor coactivators ablation on hepatic metabolism. *Molecular endocrinology* **2006**, *20*, 1138–1152.
- [175] S. Soyala, F. Krempler, H. Oberkofler, W. Patsch. PGC-1 α : a potent transcriptional cofactor involved in the pathogenesis of type 2 diabetes. *Diabetologia* **2006**, *49*, 1477–1488.
- [176] A. Zhou, Z. Wei, P. L. Stanley, R. J. Read, P. E. Stein, R. W. Carrell. The S-to-R transition of corticosteroid-binding globulin and the mechanism of hormone release. *Journal of molecular biology* **2008**, *380*, 244–251.
- [177] J. Baxter, G. Rousseau. Glucocorticoid hormone action: an overview. *Monographs on endocrinology* **1978**, *12*, 1–24.
- [178] P. Barnes. Scientific rationale for inhaled combination therapy with long-acting β 2-agonists and corticosteroids. *European respiratory journal* **2002**, *19*, 182–191.
- [179] P. Puigserver, Z. Wu, C. W. Park, R. Graves, M. Wright, B. M. Spiegelman. A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **1998**, *92*, 829–839.
- [180] S. B. Jørgensen, J. F. Wojtaszewski, B. Viollet, F. Andreelli, J. B. Birk, Y. Hellsten, P. Schjerling, S. Vaulont, P. D. Neuffer, E. A. Richter, et al.. Effects of α -AMPK knockout on exercise-induced gene activation in mouse skeletal muscle. *The FASEB journal* **2005**, *19*, 1146–1148.
- [181] M. Sanders, P. Grondin, B. Hegarty, M. Snowden, D. Carling. Investigating the mechanism for AMP activation of the AMP-activated protein kinase cascade. *Biochem. J* **2007**, *403*, 139–148.

-
- [182] B. N. Finck, D. P. Kelly. PGC-1 coactivators: inducible regulators of energy metabolism in health and disease. *Journal of clinical investigation* **2006**, *116*, 615–622.
- [183] A. Karamitri, A. M. Shore, K. Docherty, J. R. Speakman, M. A. Lomax. Combinatorial transcription factor regulation of the cyclic AMP-response element on the Pgc-1 α promoter in white 3T3-L1 and brown HIB-1B preadipocytes. *Journal of biological chemistry* **2009**, *284*, 20738–20752.
- [184] B. Nanda, A. Nataraju, R. Rajesh, K. Rangappa, M. Shekar, B. Vishwanath. PLA2 mediated arachidonate free radicals: PLA2 inhibition and neutralization of free radicals by anti-oxidants-a new role as anti-inflammatory molecule. *Current topics in medicinal chemistry* **2007**, *7*, 765–777.
- [185] Z. Hazan-Eitan, Y. Weinstein, N. Hadad, A. Konforty, R. Levy. Induction of Fc γ RIIA expression in myeloid PLB cells during differentiation depends on cytosolic phospholipase A2 activity and is regulated via activation of CREB by PGE2. *Blood* **2006**, *108*, 1758–1766.
- [186] G. Wang, K. Kawakami, G. Gick. Regulation of Na, K-ATPase α 1 subunit gene transcription in response to low K $^{+}$: Role of CRE/ATF-and GC box-binding proteins. *Journal of cellular physiology* **2007**, *213*, 167–176.
- [187] K. Willert, R. Nusse. β -catenin: a key mediator of Wnt signaling. *Current opinion in genetics & development* **1998**, *8*, 95–102.
- [188] L. Pradier, N. Carpentier, L. Delalonde, N. Clavel, M.-D. Bock, L. Buée, L. Mercken, B. Tocqué, C. Czech. Mapping the APP/presenilin (PS) binding domains: the hydrophilic N-terminus of PS2 is sufficient for interaction with APP and can displace APP/PS1 interaction. *Neurobiology of disease* **1999**, *6*, 43–55.
- [189] S. Soriano, D. E. Kang, M. Fu, R. Pestell, N. Chevallier, H. Zheng, E. H. Koo. Presenilin 1 negatively regulates β -catenin/T cell factor/lymphoid enhancer factor-1 signaling independently of β -amyloid precursor protein and notch processing. *The Journal of cell biology* **2001**, *152*, 785–794.
- [190] A. M. L. Coluccia, A. Vacca, M. Duñach, L. Mologni, S. Redaelli, V. H. Bustos, D. Benati, L. A. Pinna, C. Gambacorti-Passerini. Bcr-Abl stabilizes

- β -catenin in chronic myeloid leukemia through its tyrosine phosphorylation. *The EMBO journal* **2007**, *26*, 1456–1466.
- [191] J. Zhou, P. Su, L. Wang, J. Chen, M. Zimmermann, O. Genbacev, O. Afonja, M. C. Horne, T. Tanaka, E. Duan, et al.. mTOR supports long-term self-renewal and suppresses mesoderm and endoderm activities of human embryonic stem cells. *Proceedings of the national academy of sciences* **2009**, *106*, 7840–7845.
- [192] N. Rooke, D.-J. Li, J. Li, W. M. Keung. The mitochondrial monoamine oxidase-aldehyde dehydrogenase pathway: a potential site of action of daidzin. *Journal of medicinal chemistry* **2000**, *43*, 4169–4179.
- [193] M. Roy, N. K. David, J. V. Danao, H. Baribault, H. Tian, M. Giorgetti. Genetic inactivation of melanin-concentrating hormone receptor subtype 1 (MCHR1) in mice exerts anxiolytic-like behavioral effects. *Neuropsychopharmacology* **2006**, *31*, 112–120.
- [194] T. Beyer, M. Danilchik, T. Thumberger, P. Vick, M. Tisler, I. Schneider, S. Bogusch, P. Andre, B. Ulmer, P. Walentek, et al.. Serotonin signaling is required for Wnt-Dependent GRP specification and leftward flow in *Xenopus*. *Current biology* **2011**, 1–7.
- [195] C. G. Baptiste, M.-C. Battista, A. Trottier, J.-P. Baillargeon. Insulin and hyperandrogenism in women with polycystic ovary syndrome. *The Journal of steroid biochemistry and molecular biology* **2010**, *122*, 42–52.
- [196] C. S. Mermelstein, D. M. Portilho, F. A. Mendes, M. L. Costa, J. G. Abreu. Wnt/ β -catenin pathway activation and myogenic differentiation are induced by cholesterol depletion. *Differentiation* **2007**, *75*, 184–192.
- [197] M. H. Tao, Q. Cai, Z.-F. Zhang, W.-H. Xu, N. Kataoka, W. Wen, Y.-B. Xiang, W. Zheng, X. O. Shu. Polymorphisms in the CYP19A1 (aromatase) gene and endometrial cancer risk in Chinese women. *Cancer epidemiology biomarkers & prevention* **2007**, *16*, 943–949.
- [198] D. C. Spink, H.-P. Eugster, D. W. Lincoln II, J. D. Schuetz, E. G. Schuetz, J. A. Johnson, L. S. Kaminsky, J. F. Gierthy. 17β -Estradiol hydroxylation catalyzed by human cytochrome P450 1A1: A comparison of the activities

- induced by 2, 3, 7, 8-tetrachlorodibenzo- p-dioxin in MCF-7 cells with those from heterologous expression of the cDNA. *Archives of biochemistry and biophysics* **1992**, *293*, 342–348.
- [199] D. N. Li, A. Seidel, M. P. Pritchard, C. R. Wolf, T. Friedberg. Polymorphisms in P450 CYP1B1 affect the conversion of estradiol to the potentially carcinogenic metabolite 4-hydroxyestradiol. *Pharmacogenetics and genomics* **2000**, *10*, 343–353.
- [200] A. P. Kouzmenko, K.-i. Takeyama, S. Ito, T. Furutani, S. Sawatsubashi, A. Maki, E. Suzuki, Y. Kawasaki, T. Akiyama, T. Tabata, et al.. Wnt/ β -catenin and estrogen signaling converge in vivo. *Journal of biological chemistry* **2004**, *279*, 40255–40258.
- [201] S. Mohibi, S. Mirza, H. Band, V. Band, et al.. Mouse models of estrogen receptor-positive breast cancer. *Journal of carcinogenesis* **2011**, *10*, 35.
- [202] H. Dombret. Gene mutation and AML pathogenesis. *Blood* **2011**, *118*, 5366–5367.
- [203] S. Avagyan, F. Aguilo, K. Kamezaki, H.-W. Snoeck. Quantitative trait mapping reveals a regulatory axis involving peroxisome proliferator-activated receptors, PRDM16, transforming growth factor- β 2 and FLT3 in hematopoiesis. *Blood* **2011**, *118*, 6078–6086.
- [204] S. Dessi, B. Batetta, A. Carrucciu, D. Pulisci, S. Laconi, A. Fadda, C. Anchisi, P. Pani. Variations of serum lipoproteins during cell proliferation induced by lead nitrate. *Experimental and molecular pathology* **1989**, *51*, 97–102.
- [205] M. Westerterp, S. Gourion-Arsiquaud, A. J. Murphy, A. Shih, S. Cremers, R. L. Levine, A. R. Tall, L. Yvan-Charvet. Regulation of hematopoietic stem and progenitor cell mobilization by cholesterol efflux pathways. *Cell stem cell* **2012**, *11*, 195–206.
- [206] M. O. Grimm, T. L. Rothhaar, T. Hartmann. The role of APP proteolytic processing in lipid metabolism. *Experimental brain research* **2012**, *217*, 365–375.

- [207] R. A. K. Srivastava. Scavenger receptor class B type I expression in murine brain and regulation by estrogen and dietary cholesterol. *Journal of the neurological sciences* **2003**, *210*, 11–18.
- [208] I. Bazwinsky-Wutschke, S. Wolgast, E. Mühlbauer, E. Albrecht, E. Peschke. Phosphorylation of cyclic AMP-response element-binding protein (CREB) is influenced by melatonin treatment in pancreatic rat insulinoma β -cells (INS-1). *Journal of pineal research* **2012**, *53*, 344–357.
- [209] C. GALL, D. R. Weaver, J. Moek, A. Jilg, J. H. Stehle, H.-W. KORF. Melatonin plays a crucial role in the regulation of rhythmic clock gene expression in the mouse pars tuberalis. *Annals of the New York academy of sciences* **2005**, *1040*, 508–511.
- [210] C. Liu, S. Li, T. Liu, J. Borjigin, J. D. Lin. Transcriptional coactivator PGC-1 α integrates the mammalian clock and energy metabolism. *Nature* **2007**, *447*, 477–481.
- [211] D. Choudhary, I. Jansson, I. Stoilov, M. Sarfarazi, J. B. Schenkman. Metabolism of retinoids and arachidonic acid by human and mouse cytochrome P450 1b1. *Drug metabolism and disposition* **2004**, *32*, 840–847.
- [212] G. Lascaratos, D. F. Garway-Heath, C. E. Willoughby, K.-Y. Chau, A. H. Schapira. Mitochondrial dysfunction in glaucoma: understanding genetic influences. *Mitochondrion* **2012**, *12*, 202–212.
- [213] M. Helin, S. Rönkkö, T. Puustjärvi, M. Teräsvirta, H. Uusitalo. Phospholipases A2 in normal human conjunctiva and from patients with primary open-angle glaucoma and exfoliation glaucoma. *Graefe's archive for clinical and experimental ophthalmology* **2008**, *246*, 739–746.
- [214] V. Chrysostomou, F. Rezaia, I. A. Trounce, J. G. Crowston. Oxidative stress and mitochondrial dysfunction in glaucoma. *Current opinion in pharmacology* **2013**, *13*, 12–15.
- [215] J. Inglese, D. S. Auld, A. Jadhav, R. L. Johnson, A. Simeonov, A. Yassar, W. Zheng, C. P. Austin. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in

- large chemical libraries. *Proceedings of the national academy of sciences* **2006**, *103*, 11473–11478.
- [216] T. L. Biechele, N. D. Camp, D. M. Fass, R. M. Kulikauskas, N. C. Robin, B. D. White, C. M. Taraska, E. C. Moore, J. Muster, R. Karmacharya, et al.. Chemical-genetic screen identifies riluzole as an enhancer of Wnt/ β -catenin signaling in melanoma. *Chemistry & biology* **2010**, *17*, 1177–1182.
- [217] R. Dobrowolski, P. Vick, D. Ploper, I. Gumper, H. Snitkin, D. D. Sabatini, E. M. De Robertis. Presenilin deficiency or lysosomal inhibition enhances Wnt signaling through relocalization of GSK3 to the late-endosomal compartment. *Cell reports* **2012**, *2*, 1316–1328.
- [218] J. Gwak, J. Oh, M. Cho, S. K. Bae, I.-S. Song, K.-H. Liu, Y. Jeong, D.-E. Kim, Y.-H. Chung, S. Oh. Galangin suppresses the proliferation of β -catenin response transcription-positive cancer cells by promoting adenomatous polyposis coli/Axin/glycogen synthase kinase-3 β -independent β -catenin degradation. *Molecular pharmacology* **2011**, *79*, 1014–1022.
- [219] S. Ding, P. G. Schultz. A role for chemistry in stem cell biology. *Nature biotechnology* **2004**, *22*, 833–840.
- [220] N. M. Gangadhar, B. R. Stockwell. Chemical genetic approaches to probing cell death. *Current opinion in chemical biology* **2007**, *11*, 83–87.

Appendix A

Additional data

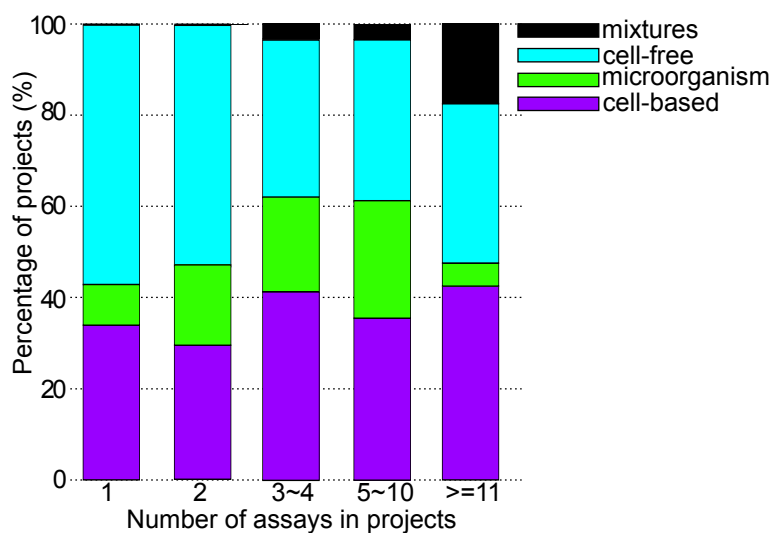


Figure A.0.1: Distribution of cell-free, cell-based and microorganism assays for projects.

A. ADDITIONAL DATA

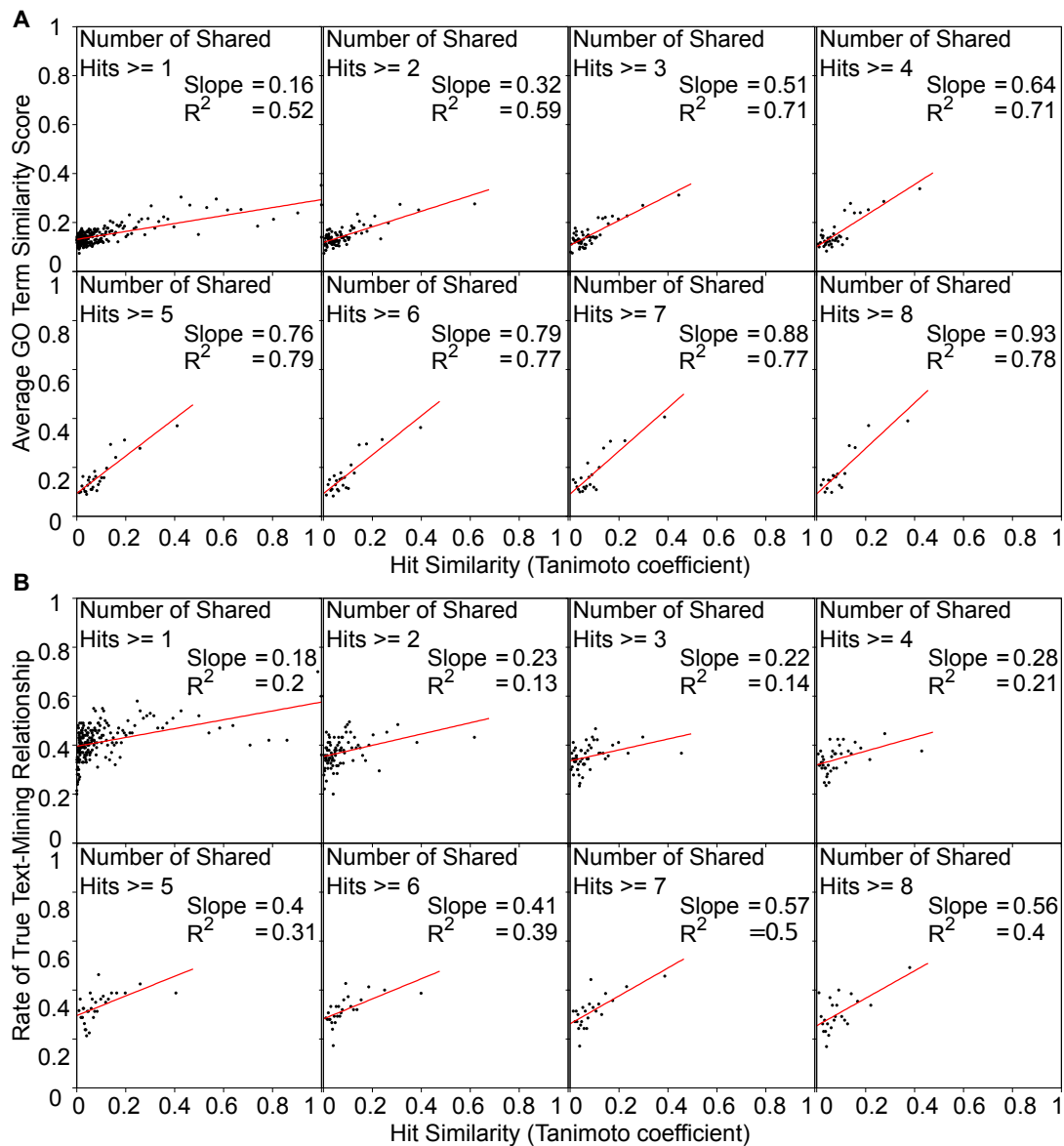


Figure A.0.2: Different cut-offs on the number of shared hits in F3 of Chem-Bank assay pairs. (A) Relationships indicated by GO terms and (B) relationships indicated by text-mining.

Table A.0.1: Shared hits with predicted targets between two assays.

The shared ChemBank compound IDs along with HitPick predicted targets (precision >50%) are shown below each assay combination. The other hits without information are not shown.		
M.Tuberculosis Macrophage	GSI Synthetic Lethal	Share 7 chemical hits
1000270	NR3C1;CYP1A1;CYP1A2;	
1001967	NR3C1;CYP3A4;ABCB1;	
3043	NR3C1;	
3187752	SERPINA6;NR3C1;	
3189180	SERPINA6;NR3C1;SLCO1A2;NR3C2;	
2080102	NR3C1;SERPINA6;SLCO1A2;SHBG;	
3616626	NR3C1;CYP3A4;	
GSI Synthetic Lethal	Adipocyte Differentiation	Share 8 chemical hits
1045	SERPINA6;NR3C1;	
1101	NR3C1;CYP3A4;	
1136	NR3C1;CYP3A4;	
1234	NR3C1;CYP3A4;ABCB1;	
1455	NR3C1;SERPINA6;SLCO1A2;CYP3A4;	
1457	NR3C1;ANXA1;	
694	SERPINA6;NR3C1;CYP3A4;	
3046112	NR3C1;SHBG;	
Unfolded Protein Response	Wnt Inhibitors	Share 14 chemical hits
3555011	ATP1A1;	
3558737	ADORA2B;ADORA2A;	
3559706	CYP1B1;	
3616386	ATP1A1;	
3616405	ATP1A1;	
Wnt Inhibitors	Histone Modification	Share 15 chemical hits
3213088	HDAC1;	
3214216	HDAC1;	
3214224	HDAC3;	
3214240	HDAC1;	
3214248	HDAC1;	
3214418	HTR6;	

A. ADDITIONAL DATA

3214452	HDAC1;	
3214496	HDAC1;	
Histone Modification	Stem Cell Differentiation	Share 15 chemical hits
3214216	HDAC1;	
3214224	HDAC3;	
3214248	HDAC1;	
3214452	HDAC1;	
3214478	HDAC1;	
3214496	HDAC1;	
Histone Modification	Genotype Specific Inhibitors NSCLC	Share 10 chemical hits
3214216	HDAC1;	
3214224	HDAC3;	
3214240	HDAC1;	
3214248	HDAC1;	
3214452	HDAC1;	
Genotype Specific Inhibitors NSCLC	Stem Cell Differentiation	Share 9 chemical hits
3214216	HDAC1;	
3214224	HDAC3;	
3214248	HDAC1;	
3214370	HDAC3;	
3214452	HDAC1;	
Genotype Specific Inhibitors NSCLC	Glioblastoma Modulators	Share 195 chemical hits
1102980	TOP2A;TUBA4A;	
1134906	ADORA3;	
1247906	HTR6;	
1393431	CNR2;	
1502789	TRPV1;	
1511593	CA2;	
1513405	CETP;	
1524322	FLT3;	
1612458	NPY5R;	
1643016	MDM2;	
1687915	ALOX5;	
1862850	CETP;	
1862851	CETP;	

1882538	MDM2;
1917199	CNR2;
1921581	OXTR;
1921765	MMP8;
3020708	HTR1A;
3026939	PTGS2;
3069277	MMP1;
3070282	MMP1;
3178168	ALOX5;
3179496	ADORA3;
3179707	DRD4;
3179710	GCGR;
3185117	ATP1A1;ATP1B1;
3214216	HDAC1;
3214224	HDAC3;
3214248	HDAC1;
3552203	PTGS1;
3554106	JUN;
3554291	ATP2A1;
3554523	FNTA;
3554525	FNTA;
3554998	ATP1A1;
3555011	ATP1A1;
3557708	ATP1A1;SLCO1A2;SLCO4C1;CYP11A1;
3557710	ATP1A1;
3557735	ATP1A1;
3558167	ATP2A1;
3558437	ATP1A1;
3558638	NR1H4;
3558693	FNTA;
3558697	FNTA;
3558877	FNTA;
3559060	MT-ND4;
3559282	SLC5A2;
3559706	CYP1B1;
3559721	FNTA;
3559755	UGT2B7;
3559863	SLC18A2;
3614482	DRD2;DRD4;
3615241	MBL2;
3616405	ATP1A1;
3622930	HDAC1;
3624592	CNR2;

A. ADDITIONAL DATA

3625232	CETP;
3625266	MDM2;
3625359	FLT3;
3625448	CNR2;
3625502	HTR6;
3625991	ADORA2A;
3635064	HDAC6;HDAC1;
3635080	HDAC3;HDAC2;HDAC6;HDAC1;
3635093	HDAC6;HDAC1;
3635098	HDAC6;HDAC1;
3644448	DRD4;
3652467	CCR4;
3652509	HDAC6;
3652551	HDAC6;HDAC1;
3652574	HTR2C;
3544	DRD4;HTR7;HTR2A;SIGMAR1;
2082296	CYP2D6;ORM1;KCNH2;CHRM2;

Glioblastoma Modulators	Wnt Inhibitors	Share 114 chemical hits
1054556	CA1;CA2;	
1134906	ADORA3;	
1241281	ALDOA;	
1247906	HTR6;	
1393431	CNR2;	
1464378	SIGMAR1;	
1511593	CA2;	
1612458	NPY5R;	
1687915	ALOX5;	
3020702	DAPK3;	
3020708	HTR1A;	
3021158	IMPDH2;	
3178168	ALOX5;	
3185117	ATP1A1;ATP1B1;	
3214216	HDAC1;	
3214224	HDAC3;	
3214248	HDAC1;	
3554106	JUN;	
3554523	FNTA;	
3554525	FNTA;	
3554998	ATP1A1;	
3555011	ATP1A1;	
3557708	ATP1A1;SLCO1A2;SLCO4C1;CYP11A1;	

3557710	ATP1A1;
3557735	ATP1A1;
3558167	ATP2A1;
3558437	ATP1A1;
3558638	NR1H4;
3558693	FNTA;
3558697	FNTA;
3558877	FNTA;
3559282	SLC5A2;
3559706	CYP1B1;
3615060	TUBB1;TUBA4A;
3615241	MBL2;
3616341	KCNA3;
3616405	ATP1A1;

Beta-Catenin	Histone Modification	Share 8 chemical hits
3214216	HDAC1;	
3214224	HDAC3;	
3214248	HDAC1;	
3214452	HDAC1;	
Wnt And Lithium Modulators	E-Cadherin Synthetic Lethal	Share 40 chemical hits
1021994	ESRRG;	
1111942	SIGMAR1;	
1112646	GRM5;	
1118692	AVPR2;	
1227426	PREP;	
1285366	SIGMAR1;	
1311944	HRH3;	
1464378	SIGMAR1;	
1570064	F2;	
1733044	SCD;	
3065778	HSD11B1;	
3178349	SCN10A;	
3554525	FNTA;	
3558623	NR1H4;	
3558693	FNTA;	
3558697	FNTA;	
3634663	HDAC1;	
3634669	HDAC1;	

A. ADDITIONAL DATA

3634737

HDAC1;
