TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Proteomik und Bioanalytik

Computational Proteomics

Harald Marx

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.

Vorsitzende: Univ.-Prof. Dr. I. Antes

Prüfer der Dissertation: 1. Univ.-Prof. Dr. D. Frischmann
2. Univ.-Prof. Dr. B. Küster

Die Dissertation wurde am 17.04.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 02.06.2014 angenommen.

Cogito ergo sum
René Descartes, Discours De la Méthode, 1637

Table of contents

Abstract

The key high-throughput technology to interrogate the proteome on a large scale is Mass spectrometry (MS) based proteomics. To interpret the resulting experimental data, computational proteomics plays a critical role. A major challenge is the reliable identification of protein sequence, function and post translational modifications. To this end, a popular approach is database searching, strongly relying on protein sequence databases to reflect the actual biological protein space and subsequent statistical result validation.

However, current protein sequence databases are not complete and build with sequence clustering, not illustrating the peptide centric and inference-prone nature of proteomics data. To reduce indistinguishable proteins by MS-based proteomics in protein sequence databases, a peptide centric clustering algorithm was developed. The algorithm is implemented as a module in a pipeline, termed mass spectrometric centric database (MScDB), accepting various source protein sequence databases as input and generates a single consensus database as result. MScDB increases the peptide to protein ratio in databases in comparison to sequence clustering and also enables the identification of peptides and putative single amino acid polymorphisms not present in UniProtKB.

The merit of including multiple databases to increase the theoretical search space, is also extendable to nucleotide databases. With the advent of next-generation sequencing more genome and transcriptome data is readily available and subject of proteogenomics. To alleviate an issue of proteogenomics, to derive a valid set of peptide and protein identifications from multiple database searches, a tailored strategy was conceived, including peptide spectrum match (PSM) grouping, an objective PSM quality criteria and the notion of genome inference. In applying the strategy on a porcine biological sample comprising nine juvenile organs and six embryonic stages, enabled refinement of known and identification of novel gene models. Additional is the unprecedented protein evidence useful to supplement the ongoing functional and structural genome annotation process.

To validate results of database searching in MS-based proteomics, a common statistical measure is the false discovery rate (FDR). A large (> 200,000) peptide and phosphopeptide reference library was synthesized to enable the objective assessment of the FDR and other analytical parameters. The synthetic peptides were fragmented with higher-energy collisional dissociation (HCD) and electron-transfer dissociation (ETD), searched with Mascot and Andromeda to derive local and global FDR models as a function of the search engine score. Furthermore is the library a valuable resource to benchmark phosphorylation site localization tools (MD-Score, PTM-Score, PhosphoRS) and derive false localization rate models. The design of the library, also gives the means to compare the retention time behaviour of modified and unmodified peptides in a reverse phase liquid chromatography system.

Zusammenfassung

Die zentrale Hochdurchsatztechnologie, um das Proteom im grossen Stil zu erforschen, ist Massenspektrometrie (MS) basierte Proteomik. Angesichts der Interpretation der daraus resultierenden experimentellen Daten, spielt Computational Proteomics eine entscheidende Rolle. Eine große Herausforderung ist dabei die zuverlässige Identifikation von Proteinsequenz, -funktion und posttranslationalen Modifikationen. Zu diesem Zweck ist ein beliebter Ansatz, die Datenbanksuche, welche sich stark auf Proteinsequenzdatenbanken zum Widerspiegeln des tatsächlichen Proteinraums und nachfolgender statistischer Validierung verlässt.

Allerdings sind aktuelle Proteinsequenzdatenbanken nicht vollständig und mit Hilfe von Sequenzgruppierung konstruiert, die nicht den Peptid-zentrischen und Inferenz-geneigten Charakter von Proteomikdaten veranschaulichen. Ein Peptid-zentrischer Gruppierungsalgorithmus wurde entwickelt, um ununterscheidbare Proteine für MS-basierte Proteomik zu reduzieren. Der Algorithmus wurde als Modul einer Pipeline implementiert, welche als Massenspektrometrie zentrische Datenbank (MScDB) bezeichnet wird, und verschiedene Quellproteinsequenzdatenbanken als Eingabe akzeptiert und als Ergebnis eine einzelne Konsensusdatenbank generiert. MScDB erhoeht das Verhältnis von Peptid zu Protein in Datenbanken im Vergleich zur Sequenzgruppierung und ermöglicht zusätzlich die Identifizierung von Peptiden und vermeintlichen Einzelaminosäure-Polymorphismen die nicht in UniProtKB vorhanden sind.

Der Vorzug mehrere Datenbanken einzubinden, um den theoretischen Suchraum zu vergrössern, kann auch auf Nukleotiddatenbanken erweitert werden. Mit dem Aufkommen von Sequenzierung der naechsten Generation sind mehr Genom- und Transkriptomdaten ohne Weiteres verfügbar und somit Thema der Proteogenomik. Zur Erleichterung des Proteogenomikaspekts, eine valide Menge an Peptid- und Proteinidentifikationen von multiplen Datenbanksuchen abzuleiten, wurde eine maßgeschneiderte Strategie konzipiert, einschließlich Peptid Spektrum Match (PSM) Gruppierung, objektives PSM Qualitätskriterium und der Begriff der Genominferenz. Unter Verwendung der Strategie an einer biologischen Schweineprobe, die neun jugendliche Organe und sechs Embryonenstadien umfasst, ermöglichte diese die Verfeinerung bekannter und die Identifikation neuer Genmodelle. Zusätzlich ist der neuartige Nachweis von Proteinen nützlich, um den laufenden funktionellen und strukturellen Genomannotierungsprozess zu ergänzen.

Zur Validierung der Datenbanksuchergebnisse in der MS-basierten Proteomik, ist die False Discovery Rate (FDR) ein wichtiges statistisches Mass. Eine große (> 200.000) Peptid- und Phosphopeptidbibliothek wurde synthetisiert, um die objektive Beurteilung der FDR und anderer analytische Parameter zu ermöglichen. Die synthetischen Peptide wurden mittels stossinduzierter Dissoziation mit hoher Energie (HCD) und Elektronentransferdissoziation (ETD) fragmentiert, und mit Mascot und Andromeda gesucht, um lokale und globale FDR Modelle als Funktion des Suchmaschinenscores abzuleiten. Darüber hinaus ist die Bibliothek eine wertvolle Ressource für das Benchmarking von Tools zur Lokalisierung der Phosphorylierungsstelle (MD-Score, PTM-Score, PhsophoRS) und um False Localization Rate Modelle abzuleiten. Das Design der Bibliothek, ermöglicht auch das Retentionszeitverhalten von modifizierten und unmodifizierten Peptiden in einer Umkehrphasen-Flüssigkeitschromatographie zu vergleichen.

Chapter 1

General introduction

Genomics, transcriptomics and proteomics

The ultimate goal of biological sciences is a holistic view on the building blocks of life, cells, in particular their components and interplay on a molecular level. The key component is the deoxyribonucleic acid (DNA), a molecule that encodes the instructions to sustain the cellular functions. The instructions, also referred to as genes, are fundamental to express other vital components, including ribonucleic acid (RNA) and protein molecules.
To systematically interrogate structure and function of the myriads of components in cell types, body fluids, tissues and organs, necessitates large scale and high throughput technologies. Hence, the last decade brought forth major technological advances leading to the omics era with key disciplines such as genomics, transcriptomics and proteomics, that facilitate routine DNA- and RNA-sequencing (RNA-Seq) [1] of more than 100 million base pairs (bp) and profiling of thousands of proteins per day [2, 3].

In 2001, the Human Genome Project was the result of a 10 year concerted effort to sequence a single 3.3 billion bp genome [4]. With the advent of next generation sequencing (NGS) a few years later, even personal genome sequencing was feasible, as in the 1000 Genomes Project, resulting in a genome per day [5]. In addition has NGS also revolutionized gene expression analysis, giving the means to characterize and quantify RNA in the dynamic range of six orders of magnitude [6].
In comparison [7], is the challenge for complex proteomes even more difficult, exceeding seven in cells to ten orders of magnitude in body fluids [8], requiring highly sensitive methods. The emerging technology over the last years for in-depth proteomics profiling, was mass spectrometry (MS) based proteomics [9].

The amount and complexity of data generated by these high throughput technologies poses a challenge to manual curation and requires automatic mechanisms. Therefore the field of bioinformatics [10] plays a critical role, covering major research areas such as data integration, sequence analysis (structure), annotation and expression analysis.
In particular, MS-based proteomics requires novel, sophisticated bioinformatics algorithms and tools to tackle the associated computational challenges. This new area of bioinformatics is referred to as computational proteomics [11 - 14].

Mass spectrometry based proteomics

MS-based proteomics covers protein sequencing, quantification, post translational modifications (PTM), protein-protein interactions (PPI), localization and structure [15]. In contrast to the DNA- and RNA- NGS methods featuring single nucleotide resolution and upcoming direct readout, is MS-based proteomics a multi-step process to characterize protein sequences and modifications.
MS-based proteomics (Fig. 1) is divided in two prevalent paradigms, namely “top-down” and “bottom-up”. In “top-down” are the intact proteins subject of the analysis, whereas the more common “bottom-up” proteomics approach requires proteolytic digestion of the proteins into peptides prior to the analysis (shotgun proteomics) [16].
In general a mass spectrometric measurement is, the analysis of an ionized analyte (e.g. peptide,

protein) in gas phase. To this end, the analyte is ionized in an ion source, measured by a mass analyzer to determine the mass-to-charge ratio (m/z) and a detector to register the number of ions at each m/z value (intensity). To depict the result, the intensity values are plotted as a function of the m/z values, and is referred to as mass spectrum. To derive in MS-based proteomics the amino acid sequence the mass spectrum is processed with various computational strategies.
The next paragraphs outline briefly the major experimental and computational steps of the "bottom-up" approach.

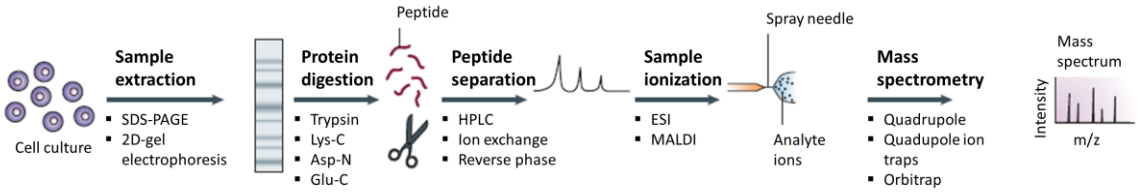

Figure 1. The default MS-based proteomics workflow (adapted from 16). The workflow comprises, sample preparation, protein digestion, peptide separation, sample ionization and the mass spectrometric analysis.

Sample extraction

The biological samples analyzed in a MS-based proteomics experiment consist of various sources such as cell culture systems, tissues, body fluids or organs. Each sample type may be subject of methods to sort, dissect or isolate a subpopulation of the originating source, depending on the sample homogeneity [17, 18]. The subsequent sample preparation to extract the proteome, comprises diverse strategies to access insoluble proteins and proteins localized in organelles or membranes [19].

Protein separation

The limited analytical capacity of the mass spectrometer requires pre-fractionation of the extracted proteins to scale down the complexity of the proteome. The most common techniques are one- or two-dimensional gel electrophoresis or affinity chromatography [20]. Gel electrophoresis is a generic technique to separate the complete proteome. In contrast, affinity purification reduces the complexity of the sample by capturing specific proteins, complexes or protein classes [21].
In gel electrophoresis are proteins separated in native or denatured conformation based on their size and charge. In one-dimensional gel electrophoresis, polyacrylamide gel electrophoresis (PAGE) is the method of choice in conjunction with sodium dodecyl sulfate (SDS). SDS is a strong detergent (denaturation of proteins) that wraps around the polypeptide chain to provide a uniform charge to mass ratio. The coated proteins are separated by size in the acrylamide gel applying an electric field to enable the migration of the proteins. Subsequent staining with coomassie brilliant blue or silver visualizes the separated proteins as bands. A merit of gel electrophoresis is the removal of unwanted byproducts of the biochemical protein (sample) preparation such as salts and detergents.

Protein digestion

The gel bands are excised and usually digested by proteases, i.e. enzymatical cleaving of the protein

sequence at specific sites into peptides. The most common protease is trypsin, cleaving the carboxyl-terminal side of arginine and lysine residues, generating suitable peptides for the mass spectrometric measurement. Notable features of trypsin are the high specificity (few miss cleavages), and the length of peptides generated, due to the frequency of arginine and lysine in protein sequences. Alternative proteases with consistent and predictable cleavage patterns are among others Lys-C, Asp-N and Glu-C.

Even though prior protein separation by gel electrophoresis is beneficial, alternative methods have been developed, which employ a digestion of proteins in solution or on filter devices. However, these approaches require additional dimensions of peptide separation to cope with the increase of sample complexity.

Peptide separation

The complexity of the resulting peptide mixture is unmanageable for an in depth mass spectrometric measurement and therefore subject to an additional separation step, namely one- and two-dimensional (2D) chromatographic fractionation. The chromatography is column based and makes use of the physicochemical properties inherent to each peptide, such as electrical charge (ion exchange) and hydrophobicity (ion-pairing reverse phase). PTMs require additional methods to separate and enrich peptides, due to the low stoichiometry and / or low abundance [22 - 26].

The general concept of chromatography comprises a stationary phase and a mobile phase (solvent and analyte). The physicochemical interaction of the analyte and the stationary phase influences the elution order and the retention time.

In ion exchange chromatography, in particular strong anion exchange (SAX) chromatography, contains the stationary phase cationic groups which interact with anionic analytes. SAX is frequently used in combination with ion-pairing reverse phase chromatography to achieve an orthogonal separation (2D). Ion-pairing reversed columns separate analytes due to the non-polar stationary phase and an aqueous (polar) mobile phase (gradient) with an ion pairing reagent such as formic acid modulating the retention time of ionic analytes.

Ionization

In general, the analytes have to be ionized and transferred into the gas phase, to be amenable for the mass spectrometric analysis. To ionize the analyte, so called soft ionization methods, which do not fragment the analyte upon ionization, such as matrix assisted laser desorption / ionization (MALDI) [27] and electrospray ionization (ESI) [28] are common in MS-based proteomics (Fig. 2).

In ESI the eluting peptide fractions pass through a capillary with an electrostatic potential generating charged droplets. In the transition the droplet solvent evaporates and contains mostly peptides with two or more protons ($[M+nH]^{n+}$). Ionization occurs following two major theories, the Ion Evaporation Model (IEM) [29] suggests that ionized peptides at the surface of the droplet are extracted and ionized by field desorption or as in the Charge Residue Model (CRM) [30] is the solvent evaporating almost completely, leaving the charges of the droplet on the peptide.

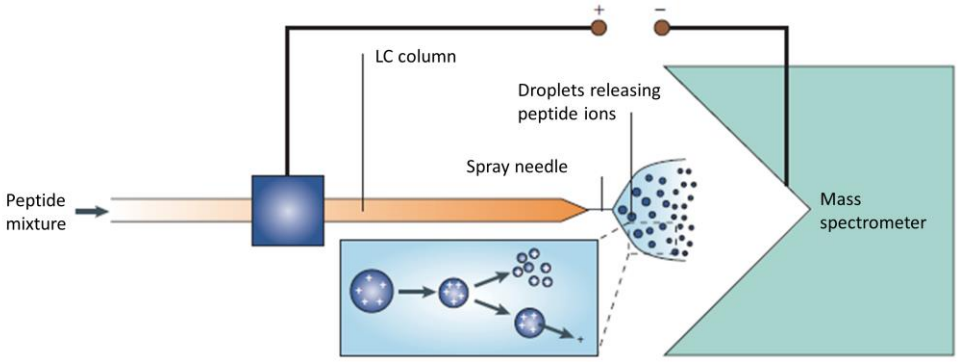

Figure 2. Electrospray ionization (adapted from 16). A peptide mixture is separated with liquid chromatography and subjected to the ion source resulting in ionized peptides based on the IEM and CRM model.

Mass analyzer

To direct the ionized analytes through the mass spectrometer electrostatic and or magnetic fields are applied by mass analyzers. In general the mass analyzer determines the mass to charge ratio (m/z), but also enables the ion transfer, selection and fragmentation. Mass analyzers with different principles of functions and features are commonly employed in MS-based proteomics, such as time-of-flight (TOF) [31], quadrupole, ion traps, orbitrap and fourier transform ion cyclotron resonance (FT-ICR) mass analyzer.

Quadrupole mass analyzers comprise two pairs of parallel rods with an oscillating electrostatic field (quadrupolar field), resulting in a spiral trajectory of the analyte. To select a specific m/z, the field frequency and amplitude are modulated leading to stable trajectories for the selected m/z value, while undesired ions are ejected or collide with the rods [32]. Quite similar is the working principle of ion traps, where ions are kept in a (quadrupolar) field, the modulation of the electrostatic field parameters ejects ions of specific m/z value that can be detected. A particular advantage is the accumulation of ions, while ion traps and quadrupole mass analyzers suffer from low resolution [33-35].

Orbitraps are based on the principle of the analyte oscillating around a central spindle, separating ions with different m/z values due to the oscillation frequency. The frequency is inversely proportional to the analyte m/z value and applying Fourier Transformation will result in the actual m/z value [36, 37].

Tandem mass spectrometry

To identify peptides different paradigms are used, namely peptide mass fingerprinting (PMF) [38-40], accurate mass and time tag (AMT) and tandem mass spectra (MS/MS). In contrast to PMF is MS/MS not solely relying on the mass of the peptide but also acquiring sequence level information to derive a valid peptide sequence.

In MS/MS, ions of a particular m/z are selected in the first mass analyzer, subsequently fragmented and resulting fragment ions are then measured by a second mass analyzer. The use of two separated mass analyzers is referred to as ‘tandem in space’, while sequential use of the same analyzer can be viewed as ‘tandem in time’.

A commonly employed ‘tandem in space’ setup is the combination of an ion trap and a FT-ICR or

Orbitrap mass analyzer, featuring high accuracy and resolution of the Orbitrap readout and, fast scanning time for the readout of fragment spectra in the ion trap. Other tandem configurations are the triple quadrupole (QQQ) comprising a Q_1 mass analyzer as m/z selection, Q_2 as fragmentation and Q_3 as filter or scan of fragments as well as, quadrupole and TOF (Q-TOF) and quadrupole and ion trap (Q-TRAP).

Fragmentation

To derive the peptide sequence information, the peptide backbone has to be fragmented, common techniques are resonance-type or beam-type collision induced dissociation (CID), electron transfer dissociation (ETD), electron capture dissociation (ECD) and post source decay (PSD). The peptide sequence is fragmented in CID based on the collision with inert gas molecules (He, N_2, Ar), the resulting kinetic energy is converted to internal energy ultimately breaking the weakest peptide bonds. The resulting fragments are referred to as b- and y- ions depending on whether they contain the N- or C- terminus of the peptide. In ETD an electron donor reacts with the peptide ions and transfers an electron. This leads to a radical anion and the unpaired electron configuration is so unstable that peptide bonds are rapidly cleaved. The resulting ions are c- and z- fragment ions. The process can diminish the sensitivity because of charge reduction [41 - 43]. In general is CID more advantageous for peptide sequence determination and ETD for modified or large peptides due to marginal cleavage bias.

Liquid Chromatography tandem mass spectrometry (LC-MS/MS)

The coupling of liquid chromatography and tandem mass spectrometry enables the separation and mass spectrometric analysis of highly complex peptide mixtures. Most commonly, LC and MS are coupled via an ESI interface (LC-ESI-MS/MS). Miniaturization of the ESI interface and the LC flow rates lead to the development of nanoLC and nanoESI systems which feature a dramatically increased sensitivity compared to standard ESI approaches [43, 44]. NanoLC is advantageous for low sample quantities and nanoESI features a 100% ionization rate with no ion suppression, i.e. reduction of ionization efficiency of the analyte of interest due to competing molecules.

The prevalent chromatography in LC-MS/MS is ion-pair reverse phase. Advantage is the volatile solvents used in reverse phase and the direct coupling to the ESI source (on-line configuration).

Peptide and protein quantification

MS-based proteomics is also used for peptide and protein quantification. To quantify, the main strategies are the relative and absolute quantification [45] of the analyte. In recent years absolute quantification gains popularity with strategies such as selected reaction monitoring (SRM) and multiple reaction monitoring (MRM) [46]. A comprehensive and critical review is available by Bantscheff et al [47].

Computational proteomics

MS-based proteomics relies on computational proteomics to process the raw data into interpretable biological information, comprising spectra pre-processing, identification, quantification and statistical validation [48, 49].

Spectra preprocessing

The experimental mass spectra comprise noise as well as isotopic and charge state variants of the analyte, requiring pre-processing to simplify the spectrum for further analysis [50].

First, the spectrum is subject to noise reduction, discriminating electronic noise, chemical noise (e.g. sample handling) and the signal of interest. The signal of interest should ideally have the highest intensity. In MS-based proteomics various approaches are used to remove noise and / or select a peak, such as local maximum detection to wavelet analysis [51, 52] In addition the signal intensity in the mass spectrum has to be normalized to distinguish noise from real signals (signal to noise ratio) [53, 54].

Second, the peptide of interest may occur at multiple m/z values due to the natural isotope distribution (isotope envelope) of its atomic constituents, requiring the reduction to one monoisotopic peak (deisotoping) [55, 56]. In tandem mass spectra an additional step is to collate fragments in different charge states to one state, referred to as charge state deconvolution [57, 58].

The MS preprocessing is error-prone and in cases of false monoisotopic peaks or undeterminable charge states, precursor mass errors can be corrected based on the MS/MS spectrum.

Peptide sequence assignment to MS/MS spectra

The processed MS/MS spectra (peak lists) are subject to various peptide identification strategies. To assign a peptide sequence to a spectrum, the experimental MS/MS spectra are correlated against theoretical spectra from a protein sequence database (database search approach) or against spectra from a set of interpreted spectra (spectral library searching) [59-62]. Other strategies, directly infer the peptide sequence from the spectrum (*de novo* approach) [63, 64] or combine strategies (hybrid approaches). In hybrid approaches, a short segment of the spectra (highest intensity) is interpreted with *de novo* sequencing (sequence-tag) reducing the number of theoretical candidate peptides from a sequence database to search against [65, 66].

Sequence database searching

The most common peptide identification approach is database searching (Fig. 3). To this end, the search engine compares the peak lists against theoretical peptide fragment mass lists generated with a proteolytic *in silico* digest of a (protein) sequence database and several search criteria. The most important criteria comprise, enzyme digestion, post-translational or chemical modifications, parent ion mass tolerance, type of fragment ions and the fragment ion tolerance. As a result, the search engine reports a set of peptide spectrum matches (PSM) ranked by a score [67].

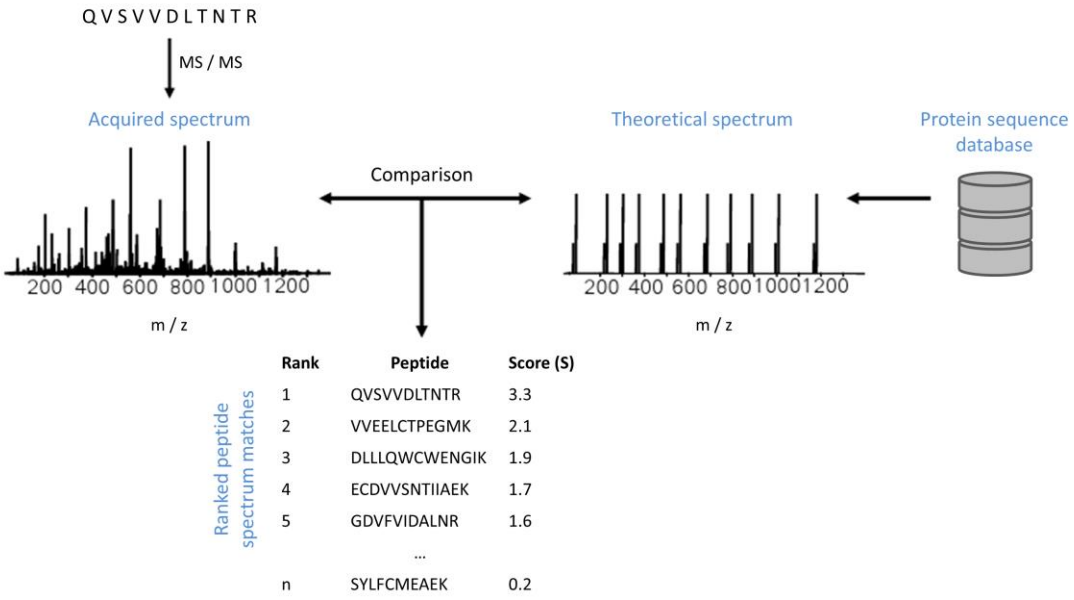

Figure 3. Sequence database searching (adapted from 49). Correlation of experimental (acquired) and theoretical spectrum constructed from a protein sequence database. The candidate peptides are ranked based on a score, reflecting the quality of the match.

Databases

A sequence database is indispensable to database searching, illustrating the potential identifiable entities of a single or multiple organisms. Additional is the content of databases varying, due to manual or automatic curation, mostly reflecting the quality of the database [68].

In MS-based proteomics the default database type, are protein sequence databases, provided by consortia such as UniProt, European Bioinformatics Institute (EBI), Wellcome Trust Sanger Institute or National Center for Biotechnology (NCBI). In general, are protein sequence databases, the result of integrating genome and transcriptome information, namely expressed sequence tags (EST), complementary DNA (cDNA) or messenger RNA (mRNA), and at times supplemented with protein level evidence.

With the advent of next generation sequencing, arises a plethora of genomes and transcripts even down to a cell specific level. To make use of this genomic information in proteomics, the field of proteogenomics is providing the means to combine sequence data from various sources and make it attainable for database searching [69 - 71].

In database searching, sequence databases constitute the theoretical search space, and therefore restrict the number of peptides assigned to a spectrum, also referred to as candidate peptides.

In silico digest

To generate the candidate peptide list and respective theoretical fragment spectra to compare the experimental against, the amino acid sequence database is subject to *in silico* digestion, considering the protease cleavage pattern, missed cleavages as well as fixed and variable modifications.

The *in silico* digest, simulates a biological protease, including incomplete digestion caused by unread cleavage sites, also referred to as missed cleavages. Another parameter to consider are modifications, occurring on all acceptor amino acids (fixed) or on some (variable). The parameters

influence the combinatorial peptide space and therefore the list of candidate peptides for each spectrum.
To filter the candidate peptide masses, the database searching parameter ion tolerance is considering the accuracy of the mass spectrometer to detect a parent (precursor).

Construction theoretical spectrum
To construct a spectrum from the candidate peptide list, the fragmentation type is a vital parameter. The theoretical fragmentation, results in a complete spectrum that would consist of all possible peaks or in a sparse spectrum under consideration of fragment type probabilities. To derive the peak intensities, three types are used, namely the uniform theoretical spectra (UT spectra), fragment theoretical spectra (FT spectra) and residue theoretical spectra (RT spectra).
After filtering the peptide fragment spectra based on the fragment ion tolerance, the resulting candidate list is ranked based on a score.

Scoring
The score (S) is a measure for the matching similarity or quality of the experimental and theoretical MS/MS spectrum. The spectral comparison algorithms, include simple dot product, cross-correlation, empirical rules and statistical fragmentation frequencies. The cross-correlation function is the most popular, implemented in search engines, such as Mascot [72], Andromeda [73], Sequest [74] and OMSSA [75].
A naive approach is the dot product or inner product of all matching peak intensities of the theoretical (T) and experimental spectra (R), following

$$S = \sum_{j=1}^{n} I_j^R I_j^T$$

where n is the number of matching peaks. The cross correlation function, introduces a relative displacement factor to measure similarity more exact between spectra ($\tau \neq 0$).

$$S = \sum_{j=1}^{n} I_j^R I_{j+\tau}^T$$

Each candidate MS/MS spectrum is scored, resulting in multiple score-ranked peptide spectrum matches (PSM) to each experimental MS/MS spectrum.

Statistical confidence scores and error rates for peptide to spectrum matches

A major issue to assign correct peptide sequences to a spectrum is among others data generated by low mass accuracy mass spectrometers. Multiple approaches are in use to distinguish true and false identifications. To assess the confidence, a single PSM or the whole dataset can be considered.

Single spectrum confidence scores
The best hit of a PSM, can be converted into a p-value or E-value [76], in looking at the score distribution of all candidate peptides to the spectrum (null distribution). The significance can be

derived in comparing the best match against the null distribution [77 - 80]. The E-value takes into account the scores of all expected peptides with equal or higher score than the observed, assuming peptides are matching the MS/MS by chance.

Posterior probabilities and false discovery rates (FDR)

In the presence of many MS/MS spectra is the p-value not discriminative alone and requires multiple testing correction. A common approach in MS-based proteomics is the false discovery rate (FDR) [81, 82] as correction. Two concepts are prevalent, the global FDR on the entire PSM collection and the local FDR (posterior "error" probability) for individual PSMs.

Target-decoy strategy for FDR assessment

To assess the (global) FDR a common approach is the target-decoy strategy (TDA) [83]. Therefore the MS data has to be searched against a target and decoy database (Fig. 4). A basic assumption for the approach is the decoy peptide sequences and false matches in the target database follow the same distribution [84 - 86]. After filtering the PSMs (e.g. ion score cutoff Mascot), the target and decoy PSMs are used to calculate the FDR:

$FDR = N_d / N_t$, where d (decoy) and t (target)

Decoy hits correspond to false positive and target hits are true positive identifications. The construction of the decoy database (e.g. random or reverse) can vary but in general does not influence the outcome [87]. The target and decoy can be searched separately or concatenated.

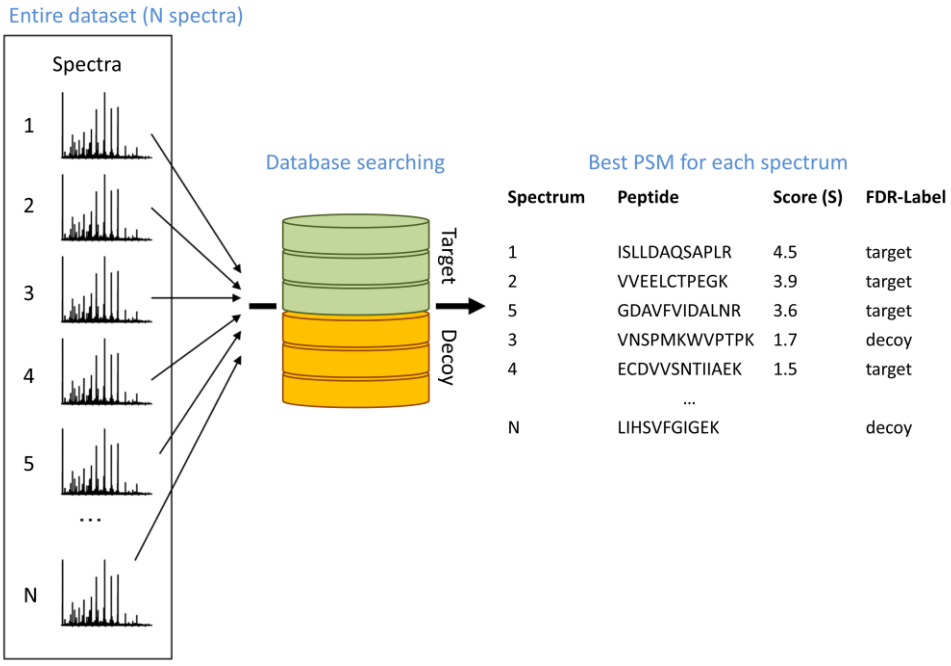

Figure 4. Target-decoy approach (adapted from 49). Comparison of experimental spectra against theoretical spectra from a target and decoy database to estimate the false discovery rate.

Mixture model methods for computing posterior probabilities and FDR

PeptideProphet [88] uses the EM-algorithm to fit a bimodal distribution (unsupervised) into the distribution of PSMs to distinguish true and false PSMs (mixture model), referred to as empirical Bayes approach [89]. The correct identifications are a gaussian distribution and the incorrect a gamma one. PeptideProphet may be also supplemented by decoy hits to make the estimation more robust (semi-supervised) [90].

Post translational discovery (PTM) discovery

MS-based proteomics is capable of identifying and localizing thousands of transient and stable PTMs, such as O-GlcNAcylation, methylation, acetylation, ubiquitination and phosphorylation. Modifications play an important role as cellular regulatory mechanism.

The strategies range from the conventional database search with user-defined PTMs to unrestrictive or "blind" searches trying to cover all possible post-translational or chemical modifications [91 - 94].

The most interesting PTM is phosphorylation, involved in many aspects of the cell regulatory response [95 - 97].

In contrast to default database searching, are multi stage search strategies, first assessing high quality spectra and sequence, second identify PTM and position. This post-processing strategy is common to supplement the search engine score with a measure of the reliability of the modification site localization. Two strategies are quite common, probabilistic approaches [98 - 101] (e.g. PTM-Score, A-Score, phosphoRS) to assess the chance of randomness in the site assignment and a delta approach (e.g. Mascot Delta Score), search engine score difference for different localizations [102, 103].

A concept to validate the localization of a PTM is the false localization rate (FLR), similar to the FDR. The concept requires decoy residue site localizations, except in the case of synthetic peptides, where a true FLR can be derived.

Objective and outline of the thesis

Computational proteomics is an upcoming field with a multitude of challenges to solve due to the ever increasing technological advance in mass spectrometry. A central task of computational proteomics is the correct assignment of peptide fragment spectra to a sequence. In the common database searching approach, assignment issues arise from the database size, unexpected peptide modifications, sequence conflicts or variants, number of missed cleavages or complete absence of the sequence in the database [104 - 106]. The objectives of this thesis were to address a few of these issues, namely improve the sensitivity of peptide identification in respect to the database size and content, validation of peptide identification and phosphorylation site localization.

Chapter 2 describes a novel clustering approach to build a protein sequence database representing the peptide centric and inference-prone character of MS-based proteomics data. The peptide centric clustering algorithm is part of a pipeline, referred to as mass spectrometry centric protein sequence database (MScDB). In contrast to common sequence clustering approaches, is MScDB increasing the peptide to protein ratio in a comparable protein sequence space and hence enables the identification of peptides and putative single amino acid polymorphisms not present in UniProtKB.

In chapter 3 a large (> 200,000 peptides) synthetic peptide and phosphopeptide library was generated to derive objective false discovery rate (FDR) and false localization rate (FLR) models. The library is a valuable resource for the evaluation of peptide identification algorithms, such as Mascot and Andromeda. Additional benchmarks of common phosphorylation site localization tools, namely Mascot Delta (MD) Score, PTM-Score and PhosphoRS, were performed. The information about true and false identifications helps to address fundamental issues in MS-based proteomics, such as database search result validation, phosphorylation localization and the chromatographic behavior of modified and unmodified peptides in a reverse phase HPLC system.

Chapter 4 covers the merits of MS-based proteomics for annotation of the recently sequenced porcine genome. The proteogenomic analysis, is based on a tailored strategy to combine search results from multiple database searches, an objective criteria to filter low scoring peptide identifications and the notion of genome inference. The results suggest improvement in existing and identification of novel gene models with unprecedented protein evidence from juvenile organs and embryonic stages.

Abbreviations

AMT	accurate mass and time tag
bp	base pairs
cDNA	complementary DNA
CID	beam-type collision induced dissociation
CRM	charge residue model
DNA	deoxyribonucleic acid
EBI	European bioinformatics institute
ECD	electron capture dissociation
ESI	electrospray ionization
EST	expressed sequence tags
ETD	electron transfer dissociation
FDR	false discovery rate
FLR	false localization rate
FT spectra	fragment theoretical spectra
FT-ICR	Fourier transform ion cyclotron resonance
IEM	ion evaporation model
LC-MS/MS	liquid Chromatography tandem mass spectrometry
m/z	mass-to-charge ratio
MALDI	matrix assisted laser desorption / ionization
MD Score	Mascot delta score
MRM	multiple reaction monitoring
mRNA	messenger RNA
MS	mass spectrometry
MS/MS	tandem mass spectra
MScDB	mass spectrometry centric protein sequence database
NCBI	national center for biotechnology
NGS	next generation sequencing
PAGE	polyacrylamide gel electrophoresis
PMF	peptide mass fingerprinting
PPI	protein protein interaction
PSD	post source decay
PSM	peptide spectrum match
PTM	post translational modification
QQQ	triple quadrupole
Q-TOF	quadrupole and TOF
Q-TRAP	quadrupole and ion trap
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RT spectra	residue theoretical spectra
SAX	strong anion exchange
SDS	sodium dodecyl sulfate
SRM	selected reaction monitoring
TDA	target-decoy strategy
TOF	time-of-flight
UT spectra	uniform theoretical spectra

References

1. Wang, Z., Gerstein, M., & Snyder, M. (2009, Jan). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10(1), 57-63.
2. Mallick, P., & Kuster, B. (2010, Jul). Proteomics: a pragmatic perspective. Nat Biotechnol, 28(7), 695-709.
3. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., et al. (2011). The quantitative proteome of a human cell line. Mol Syst Biol, 7, 549.
4. Consortium, I. H. (2004, Oct). Finishing the euchromatic sequence of the human genome. Nature, 431(7011), 931-945.
5. Consortium, 1. G., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012, Nov). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65.
6. Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., et al. (2012, Nov). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res, 40(20), 10084-10097.
7. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., et al. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol, 7, 548.
8. Mitchell, P. (2010, Jul). Proteomics retrenches. Nat Biotechnol, 28(7), 665-670.
9. Aebersold, R., & Mann, M. (2003, Mar). Mass spectrometry-based proteomics. Nature, 422(6928), 198-207.
10. Hogeweg, P. (2011, Mar). The roots of bioinformatics in theoretical biology. PLoS Comput Biol, 7(3), e1002021.
11. Colinge, J., & Bennett, K. L. (2007, Jul). Introduction to computational proteomics. PLoS Comput Biol, 3(7), e114.
12. Matthiesen, R. (2007, Aug). Methods, algorithms and tools in computational proteomics: a practical point of view. Proteomics, 7(16), 2815-2832.
13. Vitek, O. (2009, May). Getting started in computational mass spectrometry-based proteomics. PLoS Comput Biol, 5(5), e1000366.
14. Käll, L., & Vitek, O. (2011, Dec). Computational mass spectrometry-based proteomics. PLoS Comput Biol, 7(12), e1002277.
15. Patterson, S. D., & Aebersold, R. H. (2003, Mar). Proteomics: the first decade and beyond. Nat Genet, 33 Suppl, 311-323.
16. Steen, H., & Mann, M. (2004, Sep). The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol, 5(9), 699-711.
17. Svensson, M., Boren, M., Sköld, K., Fälth, M., Sjögren, B., Andersson, M., et al. (2009, Feb). Heat stabilization of the tissue proteome: a new technology for improved proteomics. J Proteome Res, 8(2), 974-981.
18. Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R., et al. (1996, Nov). Laser capture microdissection. Science, 274(5289), 998-1001.
19. Weber, G., & Wildgruber, R. (2008). Free-flow electrophoresis system for proteomics applications. Methods Mol Biol, 384, 703-716.

20. Anderson, N. L., & Anderson, N. G. (2002, Nov). The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics, 1(11), 845-867.
21. Thulasiraman, V., Lin, S., Gheorghiu, L., Lathrop, J., Lomas, L., Hammond, D., et al. (2005, Sep). Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. Electrophoresis, 26(18), 3561-3571.
22. Porath, J. (1992, Aug). Immobilized metal ion affinity chromatography. Protein Expr Purif, 3(4), 263-281.
23. Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., et al. (2002, Mar). Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol, 20(3), 301-305.
24. McNulty, D. E., & Annan, R. S. (2008, May). Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. Mol Cell Proteomics, 7(5), 971-980.
25. Pinkse, M. W., Uitto, P. M., Hilhorst, M. J., Ooms, B., & Heck, A. J. (2004, Jul). Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. Anal Chem, 76(14), 3935-3943.
26. Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., et al. (2005, Jan). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat Biotechnol, 23(1), 94-101.
27. Karas, M., & Hillenkamp, F. (1988, Oct). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem, 60(20), 2299-2301.
28. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989, Oct). Electrospray ionization for mass spectrometry of large biomolecules. Science, 246(4926), 64-71.
29. Wilm, M., and Mann, M. (1994) Electrospray and taylor-cone theory, Dole's beam of macromolecules at last? Int J Mass Spectrom Ion Process [epub ahead of print, 167-180.
30. Iribarne, J. V., and Thompson, B. A. (1976) On the evaporation of small ions from charged droplets. J Chem Phys 64, 2287-2294.
31. Whittal, R. M., & Li, L. (1995, Jul). High-resolution matrix-assisted laser desorption/ionization in a linear time-of-flight mass spectrometer. Anal Chem, 67(13), 1950-1954.
32. Jonscher, K. R., & Yates, 3. J. (1997, Jan). The quadrupole ion trap mass spectrometer--a small solution to a big challenge. Anal Biochem, 244(1), 1-15.
33. March, R. E. (2009). Quadrupole ion traps. Mass Spectrom Rev, 28(6), 961-989.
34. Schwartz, J. C., Senko, M. W., & Syka, J. E. (2002, Jun). A two-dimensional quadrupole ion trap mass spectrometer. J Am Soc Mass Spectrom, 13(6), 659-669.
35. Douglas, D. J., Frank, A. J., & Mao, D. (2005). Linear ion traps in mass spectrometry. Mass Spectrom Rev, 24(1), 1-29.
36. Scigelova, M., & Makarov, A. (2006, Sep). Orbitrap mass analyzer--overview and applications in proteomics. Proteomics, 6 Suppl 2, 16-21.
37. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Müller, M., et al. (2012, Mar). Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol Cell Proteomics, 11(3), O111.013698.
38. James, P., Quadroni, M., Carafoli, E., & Gonnet, G. (1993, Aug). Protein identification by mass profile fingerprinting. Biochem Biophys Res Commun, 195(1), 58-64.

39. Mann, M., Højrup, P., & Roepstorff, P. (1993, Jun). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom, 22(6), 338-345.
40. Pappin, D. J., Hojrup, P., & Bleasby, A. J. (1993, Jun). Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol, 3(6), 327-332.
41. McAlister, G. C., Phanstiel, D., Good, D. M., Berggren, W. T., & Coon, J. J. (2007, May). Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. Anal Chem, 79(10), 3525-3534.
42. Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E., & Coon, J. J. (2007, Jan). Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. Anal Chem, 79(2), 477-485.
43. Wilm, M., and Mann, M. (1996) Analytical properties of the nanoelectrospray ion source. Anal Chem 68, 1-8.
44. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature 379, 466-469.
45. Kirkpatrick, D. S., Gerber, S. A., & Gygi, S. P. (2005, Mar). The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. Methods, 35(3), 265-273.
46. Lange, V., Picotti, P., Domon, B., & Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol, 4, 222.
47. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007, Oct). Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem, 389(4), 1017-1031.
48. Nesvizhskii, A. I., Vitek, O., & Aebersold, R. (2007, Oct). Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods, 4(10), 787-797.
49. Nesvizhskii, A. I. (2010, Oct). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics, 73(11), 2092-2123.
50. Matthiesen, R. (2007). Extracting monoisotopic single-charge peaks from liquid chromatography-electrospray ionization-mass spectrometry. Methods Mol Biol, 367, 37-48.
51. Lange, E., Gröpl, C., Reinert, K., Kohlbacher, O., & Hildebrandt, A. (2006). High-accuracy peak picking of proteomics data using wavelet techniques. Pac Symp Biocomput, 243-254.
52. Du, P., Kibbe, W. A., & Lin, S. M. (2006, Sep). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics, 22(17), 2059-2065.
53. Kapp, E. A., Schütz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O'Hair, R. A., et al. (2003, Nov). Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Anal Chem, 75(22), 6251-6264.
54. Na, S., & Paek, E. (2006, Dec). Quality assessment of tandem mass spectra based on cumulative intensity normalization. J Proteome Res, 5(12), 3241-3248.
55. Wolters, D. A., Washburn, M. P., & Yates, 3. J. (2001, Dec). An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem, 73(23), 5683-5690.
56. Senko, M. W., Beu, S. C., & McLaffertycor, F. W. (1995, Apr). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J Am Soc Mass Spectrom, 6(4), 229-233.

57. Zhang, Z., & Marshall, A. G. (1998, Mar). A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. J Am Soc Mass Spectrom, 9(3), 225-233.
58. Wehofsky, M., & Hoffmann, R. (2002, Feb). Automated deconvolution and deisotoping of electrospray mass spectra. J Mass Spectrom, 37(2), 223-229.
59. Yates, 3. J., Morgan, S. F., Gatlin, C. L., Griffin, P. R., & Eng, J. K. (1998, Sep). Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem, 70(17), 3557-3565.
60. Craig, R., Cortens, J. C., Fenyo, D., & Beavis, R. C. (2006, Aug). Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res, 5(8), 1843-1849.
61. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., & MacCoss, M. J. (2006, Aug). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem, 78(16), 5678-5684.
62. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., et al. (2007, Mar). Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics, 7(5), 655-667.
63. Seidler, J., Zinn, N., Boehm, M. E., & Lehmann, W. D. (2010, Feb). De novo sequencing of peptides by MS/MS. Proteomics, 10(4), 634-649.
64. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., et al. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 17(20), 2337-2342.
65. Mann, M., & Wilm, M. (1994, Dec). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem, 66(24), 4390-4399.
66. Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., et al. (2005, Jul). InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem, 77(14), 4626-4639.
67. Nesvizhskii, A. I. (2007). Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol, 367, 87-119.
68. Nesvizhskii, A. I., & Aebersold, R. (2005, Oct). Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics, 4(10), 1419-1440.
69. Choudhary, J. S., Blackstock, W. P., Creasy, D. M., & Cottrell, J. S. (2001, May). Interrogating the human genome using uninterpreted mass spectrometry data. Proteomics, 1(5), 651-667.
70. Colinge, J., Cusin, I., Reffas, S., Mahé, E., Niknejad, A., Rey, P.-A., et al. (2005). Experiments in searching small proteins in unannotated large eukaryotic genomes. J Proteome Res, 4(1), 167-174.
71. Edwards, N. J. (2007). Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol, 3, 102.
72. Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999, Dec). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18), 3551-3567.
73. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res., 10(4), 1794-805.

74. Eng, J. K., McCormack, A. L., & Yates, J. R. (1994, Nov). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom, 5(11), 976-989.
75. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., et al. (2004). Open mass spectrometry search algorithm. J Proteome Res, 3(5), 958-964.
76. Fenyö, D., & Beavis, R. C. (2003, Feb). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem, 75(4), 768-774.
77. Sadygov, R. G., & Yates, 3. J. (2003, Aug). A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem, 75(15), 3792-3798.
78. Alves, G., Ogurtsov, A. Y., Wu, W. W., Wang, G., Shen, R.-F., & Yu, Y.-K. (2007). Calibrating E-values for MS2 database search methods. Biol Direct, 2, 26.
79. Alves, G., Ogurtsov, A. Y., & Yu, Y.-K. (2007). RAId_DbS: peptide identification using database searches with realistic statistics. Biol Direct, 2, 25.
80. Klammer, A. A., Park, C. Y., & Noble, W. S. (2009, Apr). Statistical calibration of the SEQUEST XCorr function. J Proteome Res, 8(4), 2106-2113.
81. Benjamini Y, Hochberg Y. (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological, 57, 289–300.
82. Efron B, Tibshirani R, Storey JD, Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association, 96, 1151–60.
83. Elias, J. E., & Gygi, S. P. (2007, Mar). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 4(3), 207-214.
84. Choi, H., & Nesvizhskii, A. I. (2008, Jan). False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res, 7(1), 47-50.
85. Käll, L., Storey, J. D., MacCoss, M. J., & Noble, W. S. (2008, Jan). Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res, 7(1), 40-44.
86. Käll, L., Storey, J. D., MacCoss, M. J., & Noble, W. S. (2008, Jan). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res, 7(1), 29-34.
87. Blanco, L., Mead, J. A., & Bessant, C. (2009, Apr). Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. J Proteome Res, 8(4), 1782-1791.
88. Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002, Oct). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 74(20), 5383-5392.
89. Choi, H., & Nesvizhskii, A. I. (2008, Jan). Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res, 7(1), 254-265.
90. Choi, H., Ghosh, D., & Nesvizhskii, A. I. (2008, Jan). Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res, 7(1), 286-292.
91. Dasari, S., Chambers, M. C., Slebos, R. J., Zimmerman, L. J., Ham, A.-J. L., & Tabb, D. L. (2010, Apr). TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res, 9(4), 1716-1726.

92. Nielsen, M. L., Savitski, M. M., & Zubarev, R. A. (2006, Dec). Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. Mol Cell Proteomics, 5(12), 2384-2391.
93. Na, S., Jeong, J., Park, H., Lee, K.-J., & Paek, E. (2008, Dec). Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. Mol Cell Proteomics, 7(12), 2452-2463.
94. Liu, C., Yan, B., Song, Y., Xu, Y., & Cai, L. (2006, Jul). Peptide sequence tag-based blind identification of post-translational modifications with point process model. Bioinformatics, 22(14), e307--e313.
95. Chalkley, R. J., & Clauser, K. R. (2012, May). Modification site localization scoring: strategies and performance. Mol Cell Proteomics, 11(5), 3-14.
96. Boersema, P. J., Mohammed, S., & Heck, A. J. (2009, Jun). Phosphopeptide fragmentation and analysis by mass spectrometry. J Mass Spectrom, 44(6), 861-878.
97. Grimsrud, P. A., Swaney, D. L., Wenger, C. D., Beauchene, N. A., & Coon, J. J. (2010, Jan). Phosphoproteomics for the masses. ACS Chem Biol, 5(1), 105-119.
98. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., et al. (2006, Nov). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell, 127(3), 635-648.
99. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., & Gygi, S. P. (2006, Oct). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol, 24(10), 1285-1292.
100. Bailey, C. M., Sweet, S. M., Cunningham, D. L., Zeller, M., Heath, J. K., & Cooper, H. J. (2009, Apr). SLoMo: automated site localization of modifications from ETD/ECD mass spectra. J Proteome Res, 8(4), 1965-1971.
101. Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., et al. (2011, Dec). Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res, 10(12), 5354-5362.
102. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., et al. (2011, Feb). Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics, 10(2), M110.003830.
103. Baker, P. R., Trinidad, J. C., & Chalkley, R. J. (2011, Jul). Modification site localization scoring integrated into a search engine. Mol Cell Proteomics, 10(7), M111.008078.
104. Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., et al. (2006, Apr). Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics, 5(4), 652-670.
105. Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzihradszky, K. F., Allen, N. P., Rexach, M., et al. (2005, Aug). Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: I. How much of the data is theoretically interpretable by search engines? Mol Cell Proteomics, 4(8), 1189-1193.
106. Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., et al. (2005, Aug). Comprehensive analysis of a multidimensional liquid chromatography mass

spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol Cell Proteomics, 4(8), 1194-1204.

Chapter 2

MScDB: A mass spectrometry-centric protein sequence database for proteomics

Abstract

Protein sequence databases are indispensable tools for life science research including mass spectrometry (MS)-based proteomics. In current database construction processes, sequence similarity clustering is used to reduce redundancies in the source data. Albeit powerful, it ignores the peptide centric nature of proteomic data and the fact that MS is able to distinguish similar sequences. Therefore, we introduce an approach that structures the protein sequence space at the peptide level using theoretical and empirical information from large-scale proteomic data to generate a mass spectrometry centric protein sequence database (MScDB). The core modules of MScDB are an in-silico proteolytic digest and a peptide centric clustering algorithm that groups protein sequences that are indistinguishable by mass spectrometry.
Analysis of various MScDB use cases against five complex human proteomes, results in 69 peptide identifications not present in UniProtKB as well as 79 putative single amino acid polymorphisms. MScDB retains ~99% of the identifications in comparison to common databases despite a 3 - 48% increase in the theoretical peptide search space (but comparable protein sequence space). In addition MScDB enables cross-species applications such as human/mouse graft models and our results suggest that the uncertainty in protein assignments to one species can be smaller than 20%.

Introduction

The most common approach in mass spectrometry (MS)-based bottom-up proteomics for the large-scale identification of proteins is the matching of peptide tandem mass spectra (MS/MS) to theoretical spectra constructed from an in-silico digested protein sequence database [1]. In this process the protein sequence database constitutes the available sequence space which should ideally be both complete and correct. The most widely used and well annotated databases in the proteomics field are the International Protein Index (IPI) [2], the UniProt Knowledgebase (UniProtKB) [3], the NCBI Protein sequence database [4] and the NCBI Reference Sequence collection (RefSeq) [5]. RefSeq and UniProtKB derive protein sequences from the International Nucleotide Sequence Collaboration (INSDC) [6], a repository comprising the European Nucleotide Archive [7], the DNA Data Bank of Japan [8] and GenBank [9] whereas IPI is a consensus of different source databases, including UniProtKB, Ensembl [10], RefSeq, H-InvDB [11], Vega [12] and TAIR [13]. A more recent consensus approach is the consensus coding sequence (CCDS) project to track protein sequences in the context of the genome sequence. The CCDS collaboration consists of the Ensembl Genome Browser, NCBI Map Viewer, the University of California Santa Cruz (UCSC) Genome Browser and the Wellcome Trust Sanger Institute (WTSI) Vertebrate Genome Annotation (Vega) Genome Browser [14].
In general, the source database content reflects alternate views and/or versions of a protein sequence resulting in often significant redundancies (mainly homologous sequences) which, in turn, can complicate protein identification in proteomics. To provide a less redundant sequence space, simple rules like one record per gene in one species (UniProtKB/SwissProt) or one record for 100% identical full-length sequences in one species (UniProtKB/TrEMBL) can be employed. To group closely homologous sequences, IPI and UniProt reference clusters (UniRef) [15] use sequence similarity

clustering algorithms, such as CD-HIT [16]. In principle, current sequence clustering algorithms [17, 18, 19, 20, 21, 22, 23] perform an all-against-all comparison of all source sequences, calculating pair-wise alignments scores using the degree of sequence similarity [24] as a distance measure. The composition of a sequence cluster is therefore affected by the similarity threshold used and directly influenced by the relatedness of sequences and/or members of a protein family. Therefore, the UniRef provides releases with different levels of granularity on the clustered sequences. In particular, sequence similarity clustering is useful for generating databases of higher eukaryotes due to the high abundance of similar sequences from e.g. alternative splicing, speciation, duplication or other transcriptional and evolutionary events.

In shotgun proteomics, proteins are digested into peptides which are then analyzed by mass spectrometry generating thousands of tandem mass spectra. In the process of protein identification, these spectra are then matched to peptide sequences in the search database and proteins are assembled from the list of identified peptides. The latter is a significant challenge because a single peptide may point to more than one protein. The loss of information between peptides and proteins arising from the proteolytic digest leads to ambiguities in the identification and quantification of proteins due to the set of shared (degenerate) peptides. This so-called protein inference problem [25] is obviously more pronounced the more redundant the information in the underlying protein sequence database is. Therefore, Nesvizhskii and Aebersold proposed to report a minimal list of proteins, consisting of groups of indistinguishable proteins explaining all of the experimentally observed peptides [25]. This may be achieved in a number of ways and several mostly a posteriori approaches have been published using techniques such as expectation-maximization [26,27] bayesian- [28,29] non-probabilistic- [30] or deterministic methods [31,32] as well as graph theory [33,34] and a heuristic approach utilizing empirical information on peptide detectability by mass spectrometry [35, 36, 37].In a recent publication, the tool IsoformResolver [38] combines a priori and posteriori information by matching experimentally observed peptides to connected components (i.e. proteins sharing one peptide) derived from a peptide centric database (mapping peptide to protein). Other peptide centric databases use the raw genome information to identify novel protein isoforms or single nucleotide polymorphism (SNP) variations [39]. In addition, some specialized databases extend the sequence space by modifying existing sequences in the source sequence databases [40] or construct a database with sequence clustering from genome sequences [41] to cover isoforms, N-terminal peptides and single amino acid polymorphisms (SAPs) not present in underlying protein databases used.

All of the above approaches have in common that they employ protein sequence similarity in some shape or form to construct the database used for searching proteomic data. Albeit powerful and widely used, this neither reflects the peptide centric nature of MS-based bottom-up proteomics nor does it acknowledge the general ability of mass spectrometry to distinguish similar sequences by virtue of the masses of the constituent amino acids (with the exception of the isomers L/I). In keeping with the above, we here propose to restructure the protein sequence space in a peptide centric fashion to create a mass spectrometry centric sequence database (MScDB) for proteomics. In contrast to current protein sequence databases, MScDB features the differences of proteins on the level of peptide sequence instead of the protein sequence.

In doing so, MScDB reduces the protein sequence redundancy by 17% in comparison to IPI although MScDB increases the theoretical search space by 35,975 in-silico peptides. It also provides an efficient way to resolve sequences over time, a situation created by the fact that the content of the major source sequence databases are in constant flux. Our experimental data using MScDB show

that hundreds of true peptide identifications missed by the most current UniProtKB complete proteome can be recovered in this way. The peptide centric core of MScDB also enables cross-species applications such as human/mouse graft models and our results suggest that the uncertainty in species assignment in such studies can be smaller than 20%.
We believe that MScDB has the potential to complement or replace the discontinued IPI [42] as the standard protein database in proteomics and we provide ready to use MS searchable FASTA files and the corresponding cluster XML files for a number of use cases. We encourage the community to participate in the further development of MScDB by making the source code available and providing pre-built MScDB versions to download (https://sourceforge.net/projects/mscdb/ and https://www.wzw.tum.de/proteomics/).

Material and methods

Datasets

Dataset I : Kinobead pull-downs of HeLa and K562 cells

To identify and explore the peptide parameters relevant for the construction process of MScDB, we used data from published kinobead pull-down experiments comprising over 500,000 identified tryptic peptides from HeLa and K562 cells [43].

Dataset II: Cancer cell line proteomes

To obtain complex proteome datasets for the experimental validation of the MScDB pipeline, shotgun proteomics experiments were performed on four human cancer cell lines (SKNBE2, OVCAR8, Colo205 and K562) and a post-delivery human placenta (Supplemental Experimental Procedure). Cell lines were lysed using Tris-HCl buffer containing 4 % SDS. The lysate was ultracentrifuged for 1h at 20 °C and 52000x g. Samples were reduced and alkylated by 10 mM DTT and 55 mM iodoacetamide. The protein extract was digested using Filter Aided Sample Preparation (FASP) as described previously [44]. 200 µg protein extract was digested using Trypsin (Promega) at final ratio of 1:100 (here 2 µg). To reduce sample complexity peptides were separated into six fractions using Strong Anion Exchange Chromatography (SAX) in tip columns. Peptides were fractionated at pH 11, pH 8, pH 6, pH 5, pH 4 and pH 3. Samples were desalted using C18 stage tips [45].
Nanoflow LC-MS/MS was performed by coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent, Dublin, CA) to an Oribtrap Velos (Thermo Scientific, Bremen, Germany). Peptides were delivered to a trap column (100 µm i.d. × 2 cm, packed with 5 µm C18 resin, Reprosil PUR AQ, Dr. Maisch, Ammerbuch, Germany) at a flow rate of 5 µL/minute in 100% buffer A (0.1% FA in HPLC grade water). After 10 minutes of loading and washing, peptides were transferred to an analytical column (75 µm x 40 cm C18 column Reprosil PUR AQ, 3µm, Dr. Maisch, Ammerbuch, Germany) and separated using a 220 minute gradient from 7% to 35% of buffer B (0.1% FA in acetonitrile) at 300 nL/minute flow rate. The Orbitrap Velos was operated in data dependent mode, automatically switching between MS and MS2. Full scan MS spectra were acquired in the Orbitrap at 30,000 resolution. Internal calibration was performed using the ion signal (Si (CH3) 2O) 6 H + at m/z 445.120025 present in ambient laboratory air. The 10 most intense precursors were selected for Higher energy Collisional

Dissociation (HCD) fragmentation with normalized collision energy of 30% at an AGC target setting of 40,000. HCD spectra were acquired in the Orbitrap at 7,500 resolution.
The raw files were processed into mascot generic format (mgf) files using Mascot Distiller 2.4.2.0 and searched against Mascot 2.3.0 (MatrixScience, London, UK) using the following search parameters: peptide charge 2+ and 3+, maximum missed cleavage 2, variable modifications – oxidation of methionine, fixed modifications - carbamidomethyl cysteine, peptide tolerance 5 ppm, MS/MS tolerance 0.02 Da, instrument ESI-Trap (HCD), decoy search enabled, searched against various databases.

Dataset III: Human Placenta Proteome
To obtain a complex proteome dataset for the experimental validation of the MScDB pipeline, a shotgun proteomics experiment was performed on a post-delivery human placenta from a healthy female (provided by Freising hospital based on informed consent).
The tissue was homogenized and lysed using 8M urea, 100mM TEAB, 1% Triton X-100, 10mM NaF and 20 mM nitrophenylphosphate. The protein extract was reduced with 10 mM dithiothreitol and alkylated with 55 mM iodoacetamide. Proteins were separated by 1D SDS gel electrophoresis and each lane was cut into 16 regions. The regions were digested with trypsin according to the procedure described in Shevchenko et al. Peptides were separated by a 120' gradient on a nanoscale reversed phase liquid chromatography system (Agilent nanoLC G2226 with G1376 loading pump) coupled online to an amaZon ion trap mass spectrometer (BrukerDaltonik, Bremen, Germany). Each protein digest was measured eight times (one full mass range and three gas phase fractions; m/z 350-580, m/z 575-800, m/z 795-1300; each by CID and ETD). For gas phase fractionation measurements, peptides already identified from the full scan range measurements were excluded from fragmentation based on m/z and retention time. The raw files were processed into mascot generic format (mgf) files using Data Analysis 4.0 (BrukerDaltonik, Bremen, Germany) and searched against Mascot 2.3.0 (MatrixScience, London, UK) using the following search parameters: peptide charge 2+ and 3+, maximum missed cleavage 3, variable modification of carbamidomethylation of cysteine residues, monoisotopic peptide mass (considering up to two 13C isotopes) tolerance 0.3Da, MS/MS tolerance 0.5Da, instrument ESI-Trap (CID) or ETD-Trap (ETD), decoy search enabled, searched against various versions of the IPI and MScDB databases. The resulting Mascot results files were exported to XML with default options but including same and sub set identifications (to investigate/demonstrate the extent of the protein inference problem). These data are available from the Tranche data repository using the following hash key: VnFzDb1JmH6ghvMP0YMcQbggv8nsuMirg9eRFEGjSCRWkm+aHqumS7Aj4D8ybtVIh5WhFszAgrqmcXLLWDjlhWQ/nQ4AAAAAAABbRg==.

Dataset IV: Xenograft tumors
For the illustration of cross-species applications of MScDB we used a published dataset of primary tumor xenografts of human lung adeno and squamous cell carcinoma (in a mouse background) [45]. 112 of the 122 Orbitrap MS raw files (in mzXML format) were downloaded from the Tranche data repository (hash key: bMNHuK72kXYUnt0X1a/+xRu1ZCjvq/hWhgmVQgvRE1a/X3ocf2/bBa0jPmyD4g0V/+Dv/QAAu206bNIi74/7AVDaYMEAAAAAAAA49Q==) (the remaining 10 files were corrupt). The raw files were processed into mascot generic format (mgf) files using Mascot Distiller 2.4.2.0and searched against Mascot 2.3.0 (MatrixScience, London, UK) using the following search parameters: peptide charge 2+ and 3+, maximum missed cleavage 3, variable modifications –

oxidation of methionine, fixed modifications - carbamidomethyl cysteine, peptide tolerance 10ppm, MS/MS tolerance 0.4Da, instrument ESI-Trap (CID), decoy search enabled, searched against MScDB version of IPI Human v3.72 and IPI Mouse v3.72.

Peptide and Protein identification acceptance criteria

The Mascot result files were processed with the software Scaffold (Proteome Software, Portland, US) using default options. Scaffold reports for each peptide and protein identification a peptide [26] and protein prophet probability [27]. We exported the results as Peptide Reports filtering with 75% peptide and 50% protein probability to correct in a post-processing step for false positive exclusive identifications to a database. These occur due to the influence of the database size on the Mascot identity score ($p = 0.05$) respective on the Peptide Prophet discriminant score (Mascot Ion Score – Identity score). All peptide identifications were ≥ 94.5% peptide probability and exclusive identifications had to be unique to a Scaffold protein group (i.e. the peptide occurring only in the group). In contrast we counted in-silico peptides as unique if they occur once in the underlying database. The raw and Scaffold files are available to download (https://www.wzw.tum.de/proteomics/) and accessible with the free Scaffold Viewer (Proteome Software, Portland, US).

MScDB Implementation

The MScDB pipeline was written in the programming language Java. The distinct modules of the pipeline, including the import of protein sequence databases, the in-silico digest, the clustering of indistinguishable proteins, and the output of result files are extendable (e.g. other file formats and or clustering algorithms). The implementation of the Data Access Object (DAO) design pattern facilitates a flexible input layer for various file formats such as FASTA and IPI-EMBL. To centralize the configuration, parameters are stored in property files, enabling users to customize the pipeline to individual requirements. The pipeline generates a MS searchable FASTA file suitable for search engines from different vendors and well-formed, valid XML files to affiliate the FASTA entries to the respective protein clusters and optional information, e.g. gene loci and cross references.
The source code of MScDB as well as pre-built FASTA formatted databases can be downloaded from https://sourceforge.net/projects/mscdb/ and https://www.wzw.tum.de/proteomics/.

MScDB Pipeline

The main components of the MScDB pipeline are the in-silico digest and the clustering (Fig. 1). The initial step is the sequential parsing of the source protein sequence database(s) entries.

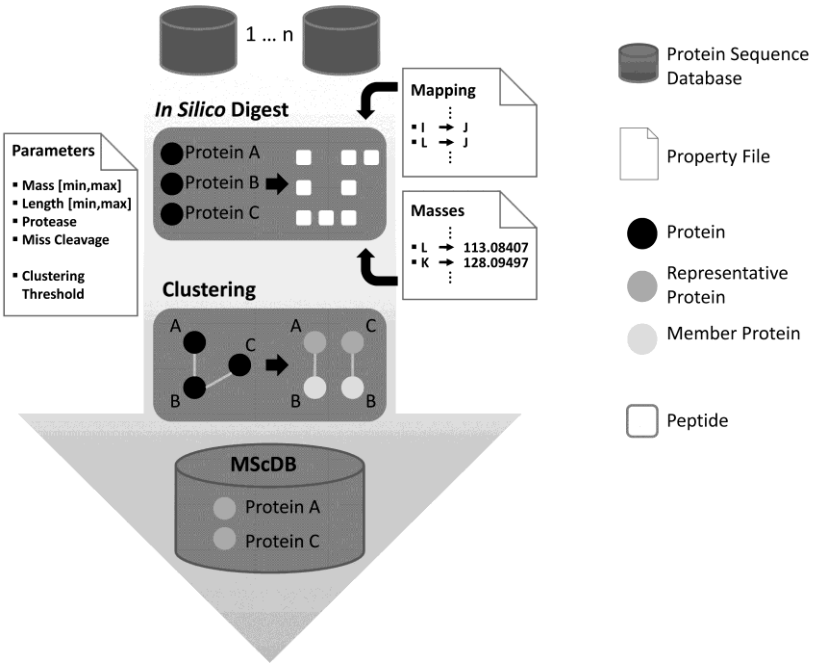

Figure 1. Workflow of the MScDB pipeline. The pipeline accepts protein databases in different file formats as input. Application of configurable parameters for the in-silico digest processes database entries into peptide lists. The peptide lists are clustered in a two-step process based on thresholds allowing their discrimination and, in the last step, the MScDB database itself is generated.

In-silico digest

To transform each entry protein sequence into a mass centric representation, the isomeric amino acids I and L are substituted by J. The sequence is then cut considering the cleavage site pattern of a protease in the form of a regular expression and n missed cleavages. Next, all peptides outside a certain mass (monoisotopic) and/or length interval are excluded (Fig. 1, see default parameters below).The resulting list of peptides is made non-redundant so that the peptides are only used once in the subsequent distance calculations (Fig. 2, and see below). To store and process peptides efficiently, distinct numbers are used to represent each peptide sequence.

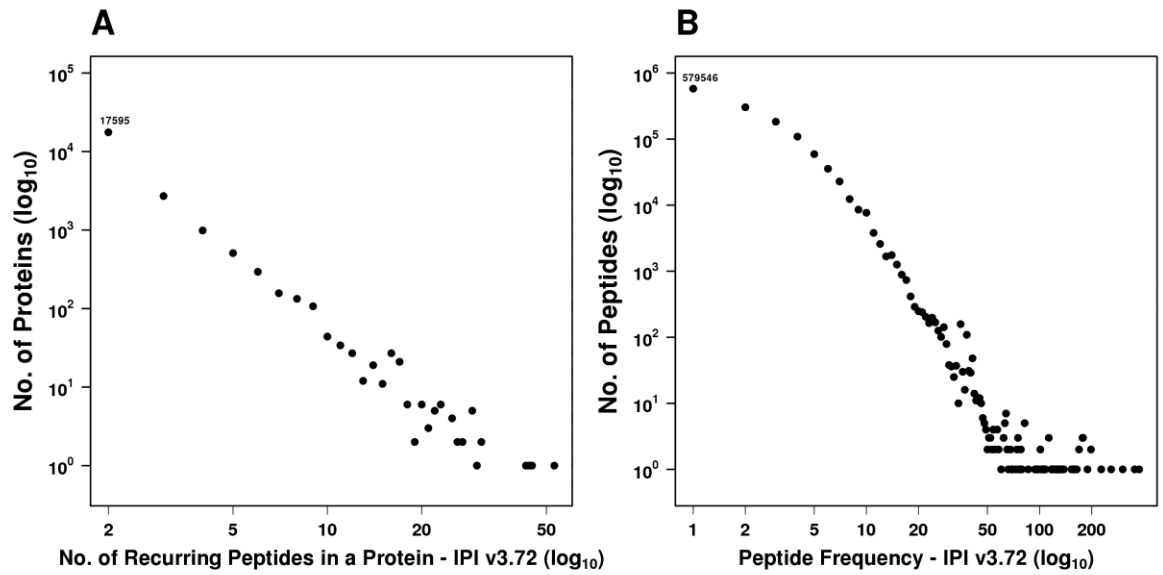

Figure 2. Peptide frequency in and over IPI proteins. (A) Redundancy of peptides in a single protein. 22738 out of 86392 proteins contain multiple peptides of the same sequence (repeat sequences). (B) Frequency of peptides in the database, resulting in 3,345,248 redundant peptides, 1,337,103 non-redundant and 579,546 unique peptides.

Peptide-centric clustering

The storage structure for the peptide lists, each representing a source protein sequence, is a non-redundant list on the protein sequence (100% identical) and peptide level (100% same peptides) associating each peptide list entry with the identifiers of identical protein sequences. Identical peptide lists form trivial initial clusters of absolutely indistinguishable proteins. Empty peptide lists, i.e. proteins with no protease specific cleavage site, are not considered for the peptide centric clustering step.

To address the incomplete sequence coverage in shotgun proteomics, the clustering algorithm calculates the asymmetric distances in an all against all comparison of the list of peptide lists validating each distance against a threshold. To calculate the distance d between peptide list $Y = \{y_1,\ldots,y_n\}$ and $X = \{x_1,\ldots,x_m\}$ each peptide is assigned a probabilistic weighting w. A simple and discrete weighting assigns each peptide an identical probability ($w = 1$) using the term Mismatch (MM). To introduce a continuous weighting ($w = [0, 1]$), we used the Peptide Sieve (PS) score [35] which assesses a proteotypic probability for each peptide, i.e. how likely it will be detected by the mass spectrometer.

$$\boldsymbol{d(Y,X) = \sum_{i=1}^{n} w_{y_i}; y_i \notin X}$$

$$\boldsymbol{d(X,Y) = \sum_{i=1}^{m} w_{x_i}; x_i \notin Y}$$

The asymmetric distance is the sum of peptide weightings for the peptides not matching between the peptide lists, including the condition of at least one matching peptide. A large distance indicates distinguishable proteins. To decide on the ability to discriminate X and Y, the distances are compared against a threshold T (a value associated with the weighting, see default parameters), as follows:

$$\min\big(\boldsymbol{d}(\boldsymbol{Y},\boldsymbol{X}),\boldsymbol{d}(\boldsymbol{X},\boldsymbol{Y})\big) \leq \boldsymbol{T}$$

The resulting intermediate clusters are connected components in which nodes represent the peptide list entries and edges are the connection of the indistinguishable ones (Fig. 1).To derive representative clusters (RC) from the connected component clusters (CCC), the final step of the clustering is the affiliation of the indistinguishable peptide list entry(s) to the respective representative(s). A representative peptide list features the largest distance to each of the nearest neighbors. Therefore the algorithm iterates through the CCC graph calculating the asymmetric distances and storing the representative to member affiliation on condition of a member not being a representative. In case the selected representative is a redundant entry (d(Y,X) = d(X,Y) = 0) on the peptide level in the list of peptide lists (see above), we chose the longest protein sequence. To create the MScDB output files from the RC, a second in-silico digest is performed (to save memory) assigning the source database annotation to the processed peptide lists. The representative protein and the non-proteolytic proteins (i.e. those not producing any peptides within the parameter range upon digestion) are written into a MS searchable FASTA file and the respective cluster information is available in a corresponding XML file. The FASTA file comprises the unmodified protein sequences from the input (source) databases. Peptide sequences are written into a separate XML file.

Biological Homogeneity Index (BHI)

To evaluate the peptide centric clustering algorithm and derive a meaningful default threshold for the distance calculation, we used the BHI [46] which is a measure for the biological homogeneity in clusters as a function of a certain class (annotation). In this work the class is the gene locus. To derive the BHI, we used the highest abundant gene locus of the distribution of gene loci in a cluster, excluding singletons and cluster without gene loci information and divide by the total number of gene loci in the cluster. We include all the gene loci for an entry in the case that multiple ones exist. Instead of averaging, we calculate the median over all clusters to derive the BHI. The range of the BHI is from 0.0 to 1.0, a high value indicating good performance of the clustering.

MScDB Default Parameters

The default parameters for the MScDB pipeline were chosen by extensive analysis of different parameter settings and experimental data. We note that users can customize the default settings to fit specific purposes. To derive the parameters for the in-silico digest, we analyzed the features of the peptides in Dataset I. The analysis comprised the peptide monoisotopic mass distribution (Fig. 3A) and peptide length distribution (Fig. 3B) as well as the occurrence of missed protease cleavages (Fig. 3C). The quantile-quantile plot shown in Supplemental Fig. 3D depicts only marginal differences between mass and length. Therefore, we exclusively use the mass criteria in this work. The peptide mass [883.4605, 2825.4963], length [8, 26] and missed cleavages [0, 1] default parameter of MScDB includes 90% (median and 45% quantiles) of the experimental data which prevents

overrepresentation of peptides that are unlikely to be detected. To deduce the clustering threshold for MM and PS, we used the Biological Homogeneity Index (BHI) [46], resulting in MM = 2 and PS = 1.0. We reasoned that different gene products from the same gene locus as well as proteins arising from paralogous genes would be hard to distinguish by mass spectrometry. We therefore set the cutoff for the biological homogeneity index based on the maximum number of contributing clusters (for each threshold) resulting in 100% (median) for MM and PS (Fig. 4). We note again that the in-silico digest and clustering parameters can be configured by the user and are only used in these respective steps. All plots in this manuscript were generated using R [47] and networks were built using Cytoscape [48].

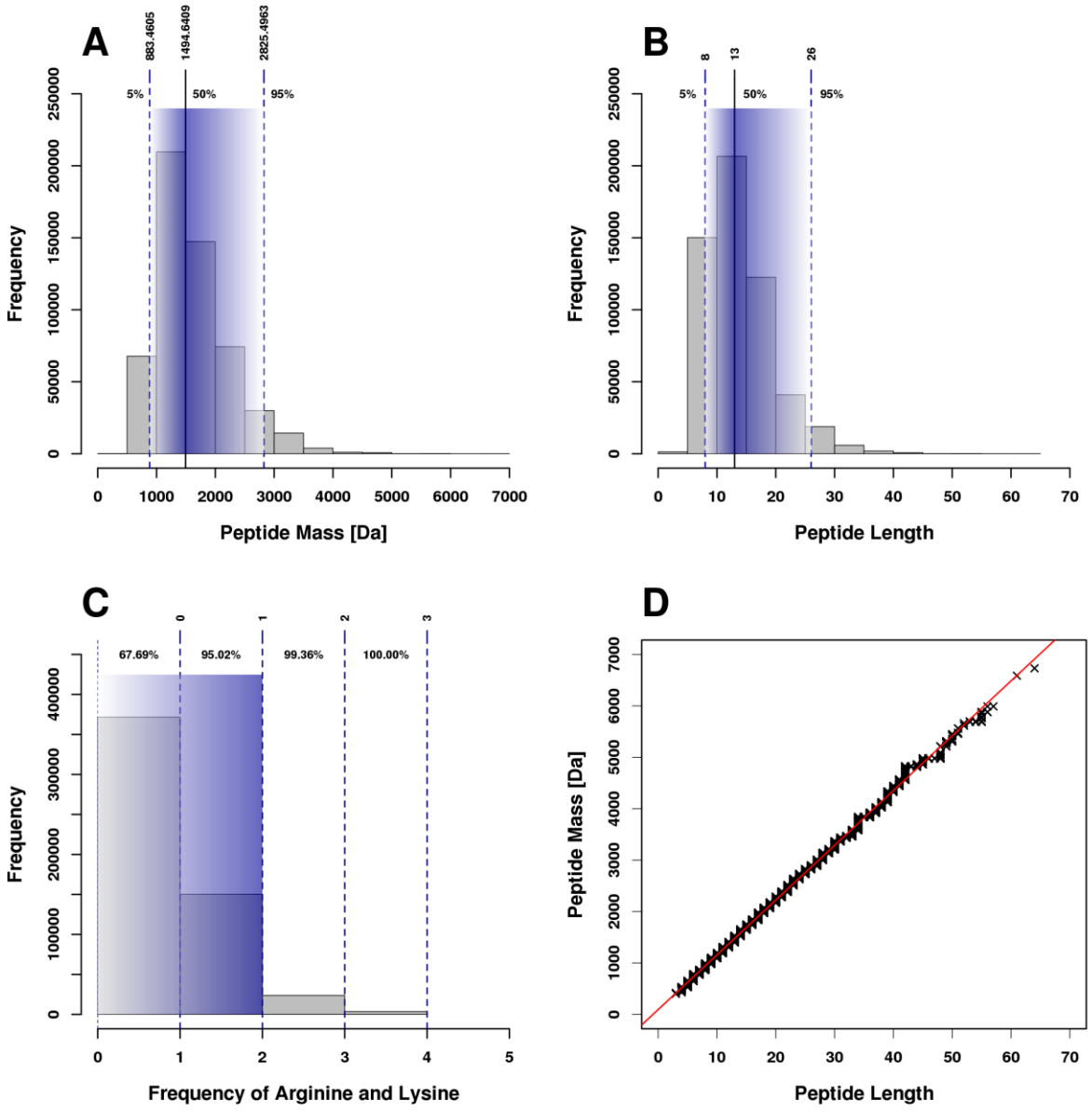

Figure 3. Properties of experimentally verified peptides to deduce default parameters for the in silico digest. (A) Monoisotopic mass distribution with median and 45% quantiles. (B) Length distribution with median and 45% quantiles. (C) Frequency of Arginine and Lysine, corresponding miss cleavages and cumulative percentage of frequency. (D) Quantile-quantile plot comparing the

monoisotopic mass and length distributions.

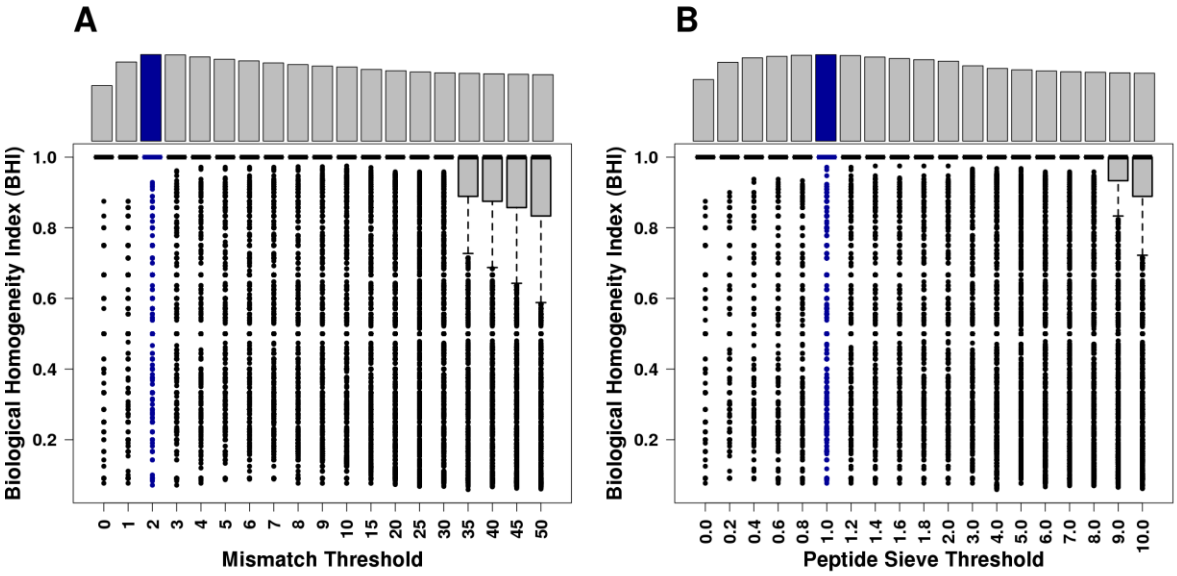

Figure 4. Biological Homogeneity Index (BHI) of gene loci in cluster to derive the default parameter for the clustering. (A) Distribution of the BHI for each mismatch threshold with the number of contributing clusters. (B) Distribution of the BHI for each peptide sieve threshold with the number of contributing clusters. The maximum of contributing clusters (blue; MM - 19727, PS - 19702) indicates the default parameter for the respective threshold.

Database reconstruction

To compare peptide centric and sequence clustering, we used the source databases (RefSeq v40, UniProtKB v2010_05, Ensembl v56 2010, H-InvDB v6.2 and Vega Mar 2009) of the IPI HUMAN version v3.72. We were able to download Ensembl v56 2010 and Vega Mar 2009 for human, as well as the canonical release of UniprotKB v2010_05 (i.e. without isoforms) and extract all human entries for Swiss-Prot and TrEMBL. The European Bioinformatics Institute (EBI) provided us with the original H-InvDB v6.2 and the National Center for Biotechnology Information the RefSeq v40 because these databases were not anymore available online. To compensate for minor discrepancies between the source data and the IPI database, we extracted the master entries of IPI.

Databases

In this study we use various human protein sequence databases to benchmark these against MScDB (Table 1). The databases vary in the underlying source databases and the construction process (sequence and peptide clustering). Our implementation of a basic sequence clustering (SC) algorithm groups proteins with 100 % sequence identity.

Name	Description	Clustering
UniProtKB [cpl]	UniProtKB (11/03/2012) complete including isoform	Sequence

	sequences	
UniProtKB [all]	UniProtKB (11/03/2012) all including isoform sequences	Sequence
IPI [src]	IPI v3.72 source database sequences; UniProtKB (04/20/2010), RefSeq 40, Ensembl 56, H-InvDB 6.2, Vega (Mar 2009)	Sequence
MScDB [src]	UniProtKB (04/20/2010), RefSeq 40, Ensembl 56, H-InvDB 6.2, Vega (Mar 2009)	Peptide
MScDB [ver]	IPI v3.00 – v3.72	Peptide
SC [ver]	IPI v3.00 – v3.72	Sequence
MScDB (T = 1, 2, 3, 4, 5)	UniProtKB (11/03/2012), RefSeq (11/05/2012), Ensembl 69 (all and ab inito), H-InvDB 8.0, Vega 49, CCDS (11/05/2012)	Peptide

Table 1. Human MScDB and common sequence databases. Nomenclature (in this work) for each database with description of sequence source or content and the respective clustering algorithm.

UniProtKB protein evidence and sequence annotation

In an attempt to classify the exclusive peptide identifications not present in IPI, UniProtKB [cpl] and UniProtKB [all], we used the web BLAST with the respective full length protein sequence and default parameters against UniProtKB, considering the highest ranking human hit.

The classification includes the protein evidence categories 'Evidence at protein level', 'Evidence at transcript level', 'Inferred from homology', 'Predicted', 'Uncertain' and an additional category 'Unknown' for BLAST searches with no result. Furthermore the classification includes the sequence annotation categories 'Sequence conflict' and 'Alternative sequence' (http://www.uniprot.org/manual/sequence_annotation).

Results

Building a mass spectrometry centric protein sequence database

MScDB is the result of an attempt to restructure the protein sequence space on the level of peptides to make it conceptually better aligned with the mass spectrometric analysis of peptides in bottom-up proteomics. Figure 1 depicts the process in which MScDB is built. Protein sequences are parsed from protein sequence databases and in-silico digested to generate lists of peptides. These lists are clustered using mass spectrometry centric parameters such as the protease used, the mass or length of the generated peptides as well as the number of allowed missed protease cleavages. We learned the latter parameters from a retrospective analysis of a large-scale proteomic study containing >500,000 high quality peptide identifications and found that 90% of the data are within a monoisotopic mass interval of 883.5 and 2,825.5 Da and that these peptides contained a maximum of one missed cleavage site. Clustering the peptide lists for all sequences in the public databases lead to the identification of representative protein sequences that can be used for protein identification by database searching (see below). This MScDB pipeline runs in quadratic time and using an Intel Core 2 - Quad@2.66 GHz machine with 8 GB RAM, the pipeline performs the in-silico digest in 13 s and the clustering (3.5 x 109 comparisons) in 23 min using default parameters and IPI Human v3.72 (86392 proteins, 82893 peptide lists) as the source database. The performance of the pipeline is

largely governed by the size of the non-redundant list of peptide lists (i.e. the size and number of input sequence databases) but the above figures show that MScDB can be built using ordinary personal computers.

A priori minimal list of protein groups

The peptide centric representation of the protein sequence space in MScDB reduces its size compared to the source database (IPI Human v3.72) because many of the protein sequences are indistinguishable by mass spectrometry and are hence clustered together (Fig. 5 A).The applied clustering parameter thresholds (here mismatch, i.e. a peptide sequence is present in one list but not the other) obviously determines the extent of the reduction. Small values for the mismatch parameter threshold (i.e. T = 0-2 corresponding to > 95 % biological homogeneity index, BHI, see Fig. 4 A) group together peptide lists of very similar proteins. Interestingly, 3499 non-identical protein sequences produce identical peptide lists upon in-silico digestion. Allowing for larger differences between peptide lists generated from proteins in the original sequence collection, the MScDB is further reduced in size (31105 distinguishable proteins at T=50). The observed reduction in database size with increasing clustering thresholds obviously correlates with an increase of protein sequences in a cluster because the process will cluster more and more dissimilar proteins (Fig. 5 B, also indicated by the increasing size of the quartiles). At our chosen default mismatch threshold of T=2, 62 % of the clusters contain a single protein sequence (singletons). This indicates that MScDB strikes a good balance between reduction of redundancy and completeness at the peptide level, resulting in an a priori minimal list of protein groups for database searching.

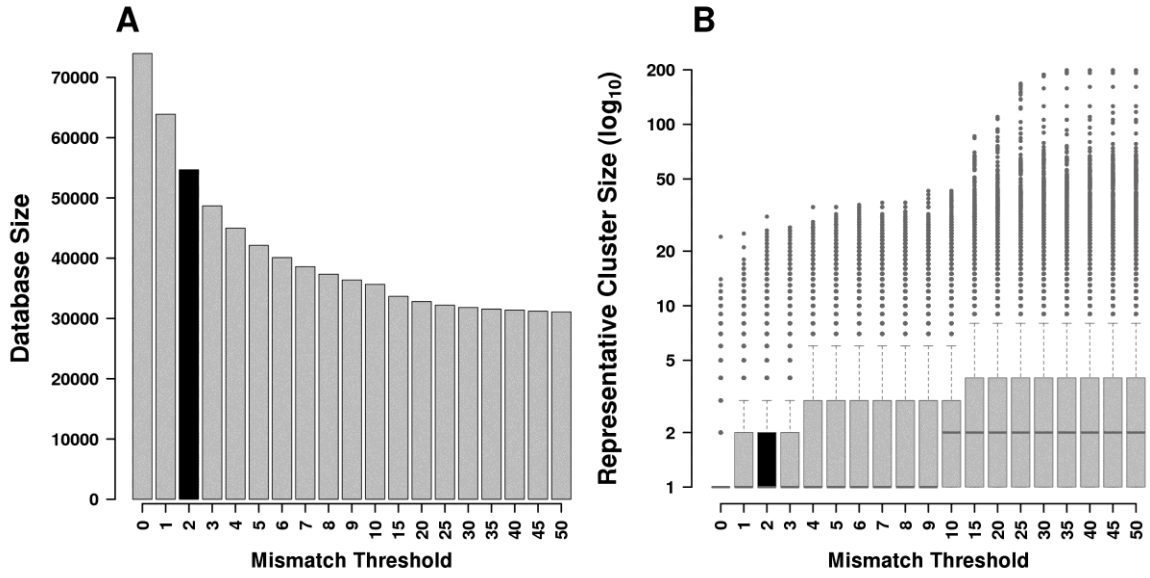

Figure 5. A priori minimal lists of protein groups. (A) Reduction of database size as a function of the mismatch threshold applied during clustering. Allowing for an increasing number of differences between peptide lists generated from proteins clusters more and more dissimilar proteins together. A mismatch threshold of two (black) strikes a good balance between redundancy and completeness of the sequence space. (B) Distribution of representative cluster sizes as a function of the mismatch threshold applied during clustering. The cluster size obviously increases with increasing the mismatch threshold. The median is indicated by a horizontal dark grey line, quartiles by boxes,

interquartile ranges by whiskers and outliers by grey dots.

Cluster characterization

The clustering module of MScDB initially creates connected component clusters (CCC) by comparing all proteins against all others in the source database and by grouping these proteins by shared peptides considering the applied clustering threshold (Fig. 6). Consequently, a CCC does, to some degree, contain distinguishable proteins. To account for this, the next step of the clustering partitions the CCC into representative clusters (RC) where a representative protein (black) illustrates the greatest distance to the nearest neighbors not being representatives themselves (grey; i.e. sharing many of its peptides with its neighbors). As an example, Figure 6A shows the CCC of glycerate kinase family (GLYCTK) sequences in which the size of the cluster nodes corresponds to the degree of connectivity. The CCC contains all known isoforms of the GLYCTK family and also includes the two uncharacterized 17 kDa and 23 kDa proteins associated with the same gene name indicating that the chosen clustering thresholds are meaningful (IPI v3.72, default parameters, T=2 allowing two mismatches). The alignment shown in Fig. 6B illustrates the process of representative selection in more detail: Isoform 2 contains 3 exclusive peptides in comparison to Isoform 7 which only contains 2 exclusive peptides and, therefore, Isoform 2 represents Isoform 7.The selection of a representative sequence increases the number of peptides in the database in comparison to sequence similarity clustering where, in general, the longest sequence is chosen as the representative.

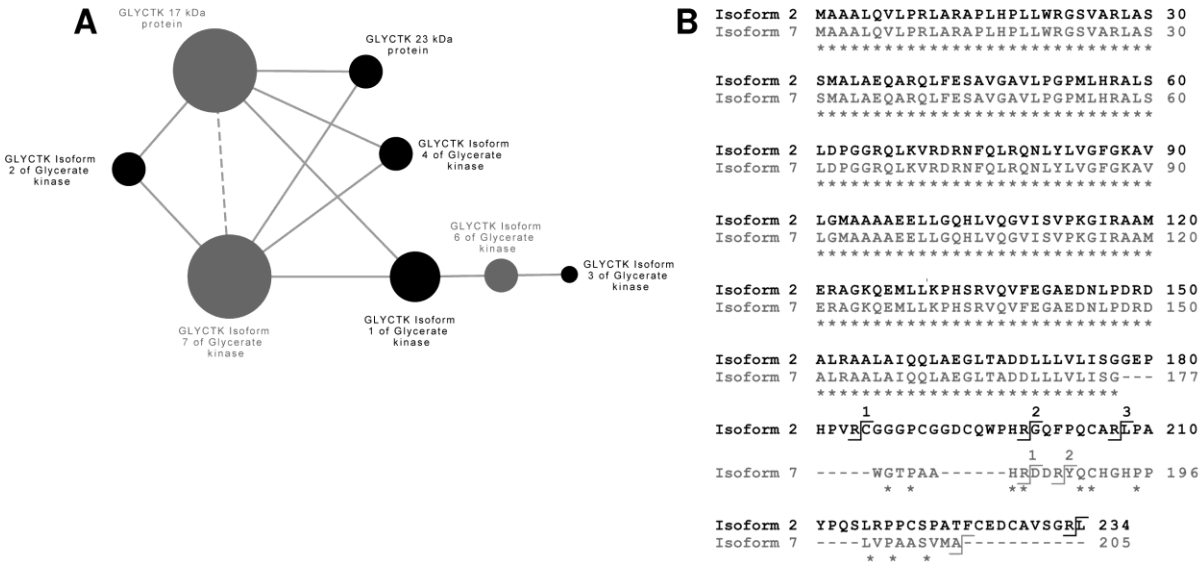

Figure 6. Connected component clusters and representative clusters. (A) Proteins of the glycerate kinase family are shown as an example for a connected component cluster and for the transition to the respective representative clusters (black nodes are representatives, grey nodes are members and the size of a node represents its degree of connectivity). (B) Sequence alignment of the C-terminal regions of Isoform 2 and 7 of glycerate kinase which highlighting the tryptic peptides of the two proteins. Using mismatch parameter of two, Isoform 2 represents Isoform 7 because the former contains more peptide information (three peptides in isoforms 2 but not Isoform 7; two peptides in

Isoform 7 but not Isoform 2).

The representation of gene loci within and across clusters can serve as an indicator for the overall quality of the initial clustering and the transition process from CCCs to RCs. Figure 7A shows that the vast majority of CCCs and RCs point to only a single gene locus. The occurrence of a limited number of clusters with more than one gene locus can be rationalized by the occurrence of MS-indistinguishable proteins (notably paralogs). It is also apparent that the representative selection process further improves the gene locus homogeneity in RCs in comparison to CCCs. In addition, we analysed how often the same gene locus is found in different clusters (Fig. 7B). Again, the majority of gene loci are found in just a single cluster. However, a significant number of gene loci can be found in multiple clusters showing that the proteins arising from these gene loci can often be distinguished by mass spectrometry. Collectively, the data indicate that the clustering method is working well and that the way in which representative proteins are chosen from a cluster (i.e. maximizing MS-compatible peptides) may be advantageous to the more traditional approach in which mostly the longest sequence in a cluster is chosen.

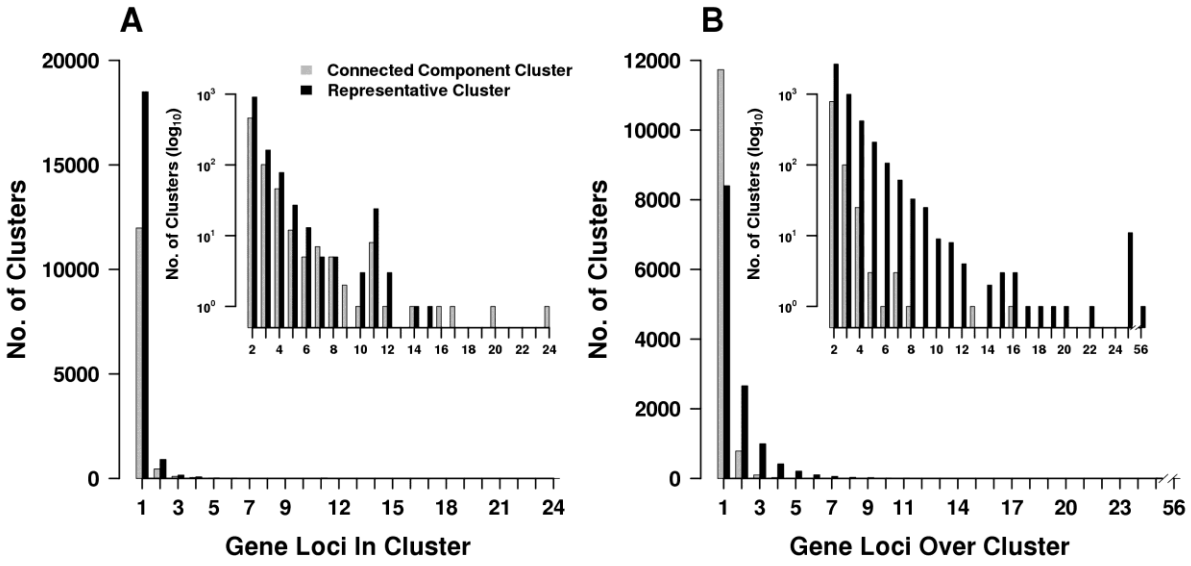

Figure 7. Cluster characterization. (A) Distribution of gene loci within clusters for connected component clusters (grey) and representative clusters (black). Proteins within a cluster tend to originate from an identical gene locus and the representative cluster selection further decreases the diversity of gene loci within a cluster. (B) Similarly, very few gene loci are distributed over several clusters (e.g. because of protein isoforms that can be distinguished by mass spectrometry).

Sequence clustering versus peptide clustering

To compare peptide and sequence clustering, we built an MScDB [src] version with default parameters of the six IPI [src] source databases (Table 1). We first evaluated if the contents of the source databases available for download are consistent with what is described for constructing the IPI [src] database and found a near to 100% consistency (Fig. 8A). The distribution of the MScDB [src]

representatives to the master entries of IPI [src] is also quite similar except for those derived from the UniProtKB/Swiss-Prot database which is default selected by the IPI [src] algorithm while MScDB [src] choses representatives on the bases of peptide sequence information (Fig. 8B).

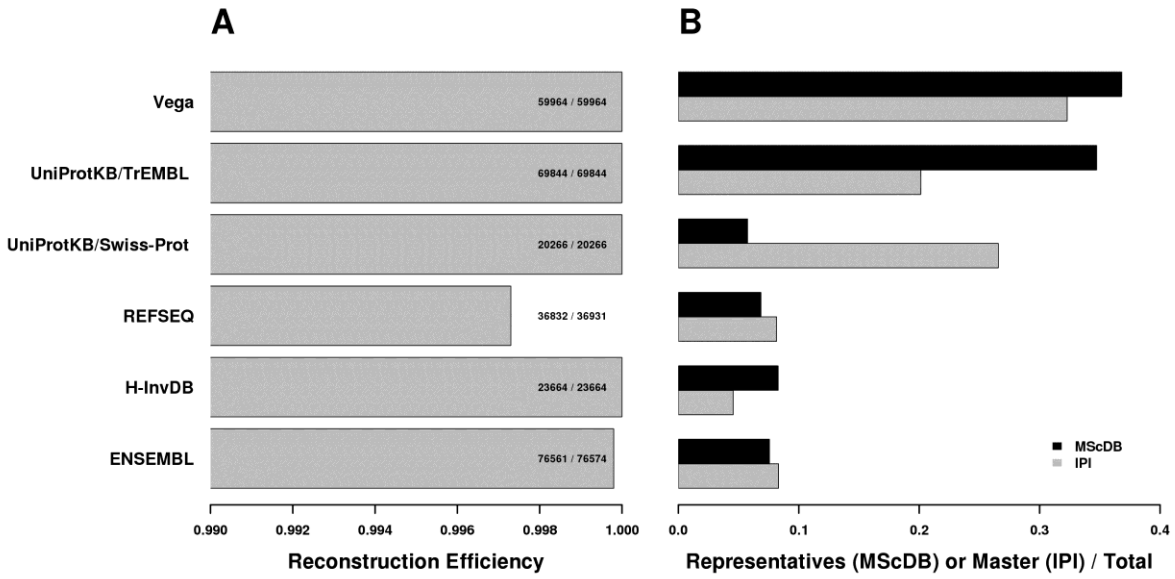

Figure 8. The IPI source databases. (A) Reconstruction efficiency of the source databases of IPI v3.72. In general, the efficiency was higher than 99.97 % (B) The distribution of master entry identifiers and representative identifiers over the underlying source databases.

In comparison the MScDB clustering algorithm reduces sequence redundancy by 17% (15282 proteins; Fig. 9A) and moderately (parameter driven) extends the theoretical search space compared to IPI [src] (35975 more in-silico peptides, 2.6 %, Fig. 9B) resulting in a slightly more comprehensive view of the MS-accessible sequence space. The ratio of theoretical peptides to proteins increases from 15.46 in IPI [src] to 19.31 in MScDB [src] (Fig. 10A). The ratio of unique (once in database) to shared peptides is comparable (Fig. 10B).

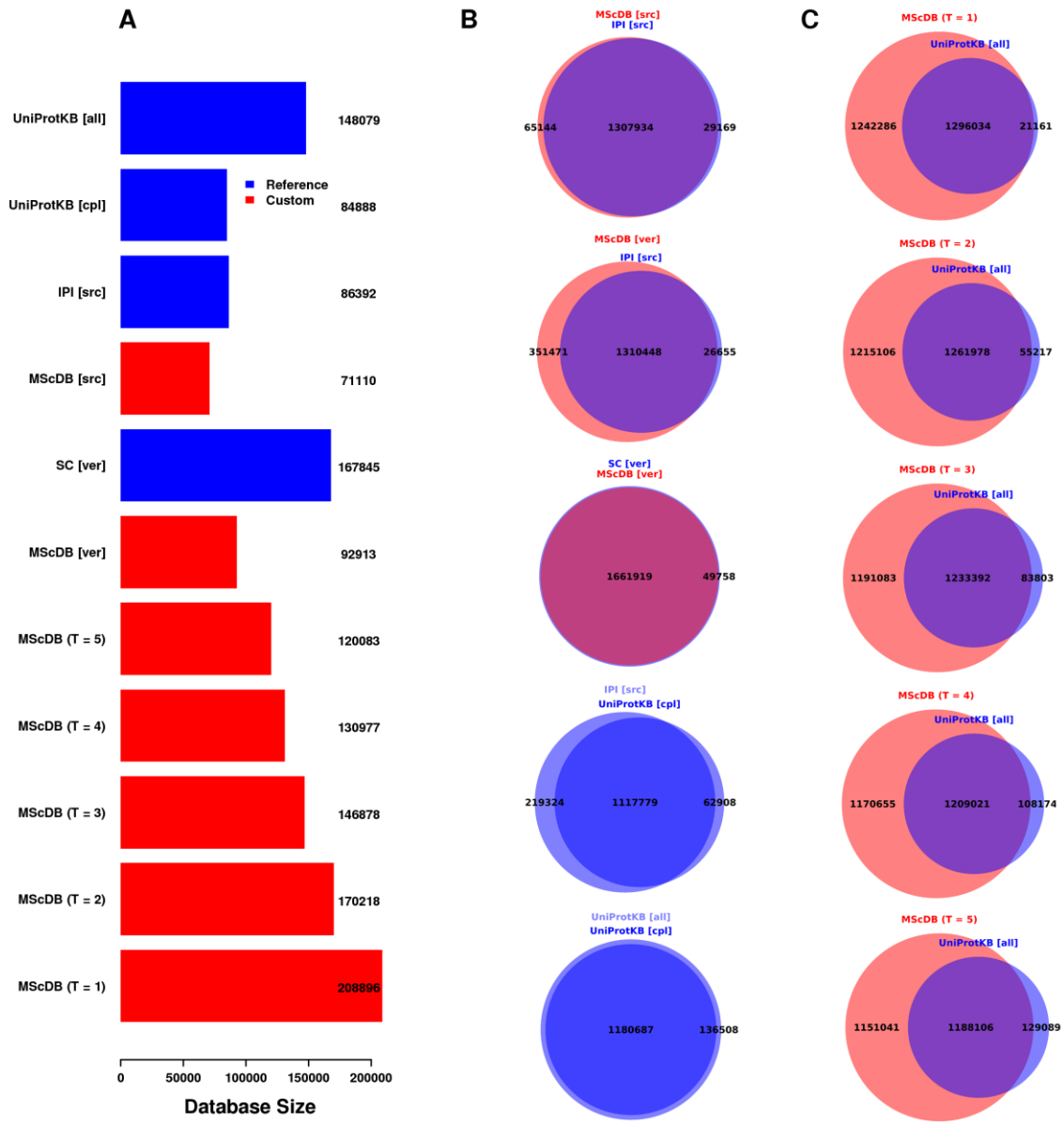

Figure 9. Features of sequence and peptide clustered databases. (A) Database sizes of common protein sequence and custom databases (Table 1). Databases to benchmark MScDB against are referred to as reference databases. (B) Comparison of in-silico peptides of reference and custom databases. (C) Influence of the clustering parameter (T) on the number of in-silico peptides in the consensus MScDB version of the latest proteome databases in comparison to UniProtKB [all].

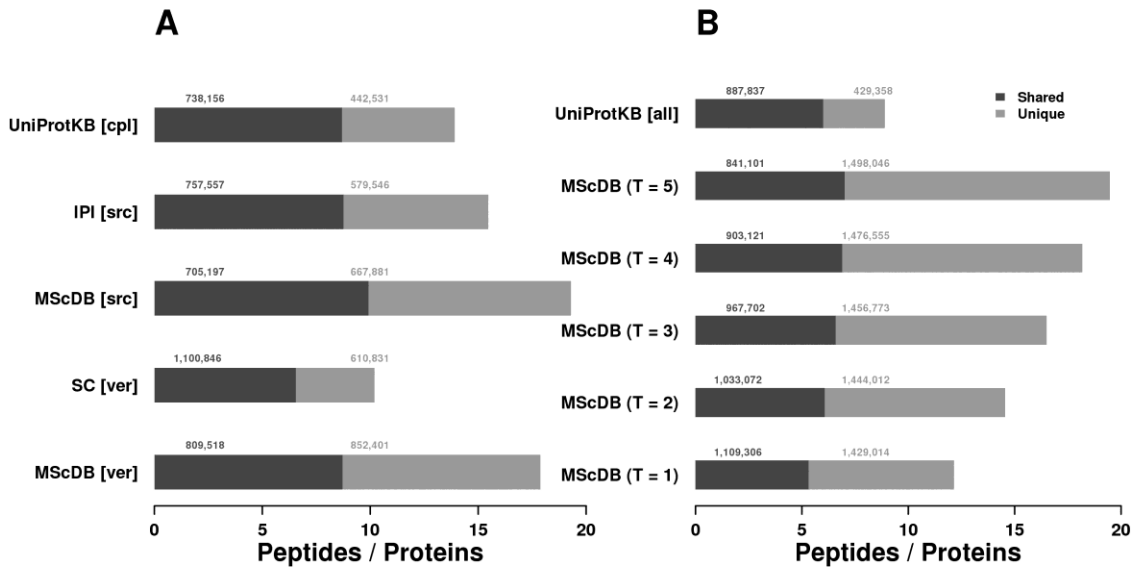

Figure 10. Peptide to protein ratio of common and custom protein sequence databases. Peptide to protein ratio of shared, unique and all peptides. (A) Ratio for common and custom databases. (B) Ratio of MScDB versions with clustering parameter T= 1, 2, 3, 4, 5 as well as UniProtKB [all].

When searching the LC-MS/MS data derived from the human cancer cell lines and the placenta against both databases resulted in > 99% identical peptides (in cell lines 34221 peptides; in placenta 8576 peptides) showing that MScDB [src] is equally useful for protein identification (Table 2). The remaining small number of exclusive identifications (Fig. 11A and Fig. 12A) in either database arises from the alternative perspectives of the clustering algorithms on the sequence space. We note that MScDB [src] identifies, in the cell lines, 106 (in placenta 44) exclusive peptides (corresponding to 91 out of 6607 protein groups; in placenta 40 out of 1547 protein groups) and IPI [src] identifies 86 (in placenta 35) exclusive peptides (corresponding to 82 out of 6573 protein groups; in placenta 31 out of 1578 protein groups).

Sample	Database	Proteins	Protein FDR(< 0.05)	Peptide FDR(< 0.05)
Cell Lines	UniProtKB [cpl]	6506	0.016	0.002
	IPI [src]	6573	0.017	0.001
	MScDB [src]	6607	0.017	0.002
	MScDB [ver]	6663	0.018	0.002
	SC [ver]	6622	0.019	0.002
	UniProtKB [all]	6588	0.015	0.001
	MScDB (T = 1)	6681	0.022	0.01
	MScDB (T = 2)	6657	0.022	0.009
	MScDB (T = 3)	6661	0.024	0.009
	MScDB (T = 4)	6654	0.024	0.009
	MScDB (T = 5)	6710	0.025	0.002
Placenta	IPI [src]	1578	0.044	0.003
	MScDB [src]	1547	0.044	0.003
	MScDB [ver]	1561	0.041	0.002

Table 2. Protein identifications in cell lines and placenta over various databases. Number of proteins over databases at a given Peptide and Protein FDR.

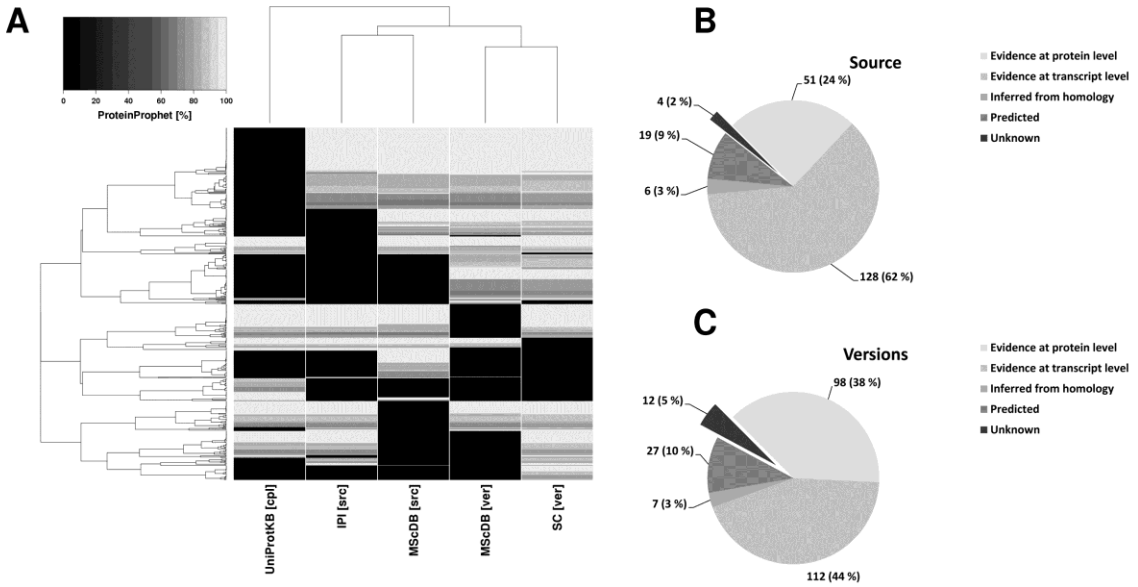

Figure 11. Complement identifications and classifications of the peptide centric MScDB clustering against common sequence clustered databases. (A) The heat map is a representation of the protein probabilities of complement peptides (not present in at least one database) over various protein sequence databases (Table 1) to show the overall confidence in peptide identifications (light grey indicates a high probability and black the absence of the peptide). Peptides clustering together are likely to origin from the same protein (left dendrogram). Groups in the databases point to similar source databases (top dendrogram). (B) Classifications based on the UniprotKB protein evidence scheme for MScDB [src] and IPI [src] identifications not present in UniProtKB [cpl] against UniProtKB [all]. (C) Classifications for MScDB [ver] and SC [ver] identifications not present in UniProtKB [cpl] against UniProtKB [all].

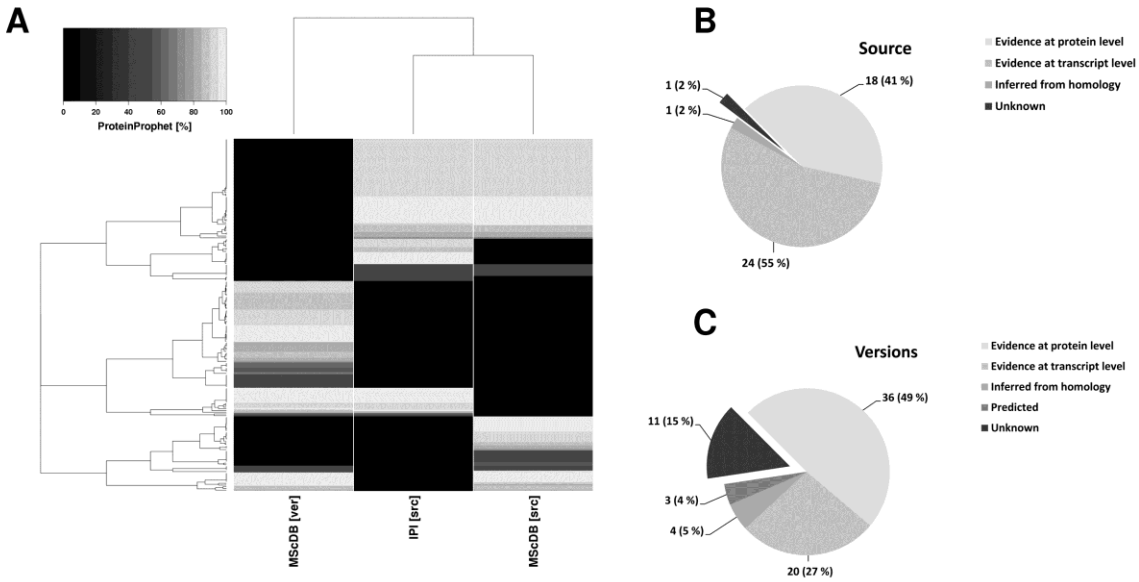

Figure 12. Complement identifications and classifications against UniProtKB. (A) The heat map is a representation of the protein probabilities of complement peptides (not present in at least one database) over various protein sequence databases (Table 1) to show the overall confidence in peptide identifications (light grey indicates a high probability and black the absence of the peptide). Peptides clustering together are likely to origin from the same protein (left dendrogram). Groups in the databases point to similar source databases (top dendrogram). (B) Classifications based on the UniprotKB protein evidence scheme for complement identifications in MScDB [src] not present in IPI [src] against UniProtKB [all]. (C) Classifications based on the UniprotKB protein evidence scheme for complement identifications in MScDB [ver] not present in IPI [src] against UniProtKB [all].

To analyze the effect of the discontinuation of IPI, we compared the substitute database UniprotKB complete proteome set (UniProtKB [cpl]) against IPI [src] and MScDB [src]. The theoretical peptide to protein ratio (Fig. 10A) decreases in UniProtKB [cpl] to 13.9 (11 % more in IPI [src] and 29% in MScDB [src]). The MS analysis of the cell lines reveals for most complement identifications an association with high protein probabilities (Fig. 10A). We identify 132 IPI [src] and 193 MScDB [src] peptides not present in UniProtKB [cpl]. The non-redundant total are 208 peptides, where 76 (66 proteins) are present in MScDB [src], 15 (15 proteins) in IPI [src] and 117 (89 proteins) in both.
Using the UniProtKB protein evidence and sequence annotation classification scheme on the non-redundant set against all available UniProtKB sequences (UniProtKB [all]; placenta classification in Fig. 12B) results in first time protein evidence for 76% of the identifications (Fig. 11B). 62% (103 proteins) have evidence on the transcript level, illustrating an overrepresentation in comparison to the 35% in UniProtKB [all] (Fig. 13). The remaining 14% are 21 automatic annotated proteins (12%) and 4 unknown proteins to UniProtKB [all] (2%). 24% (46 proteins) have protein evidence, due to the absence of natural variants and sequence conflicts in the downloadable UniProtKB FASTA files.
Over the protein evidence categories we provide evidence for 4 unknown peptides and 15 putative single amino acid polymorphisms (SAPs).

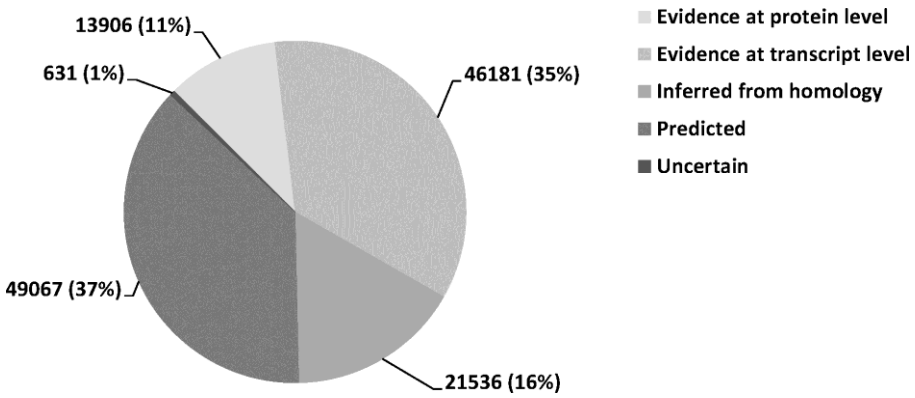

Figure 13. Protein evidence in UniProtKB. The UniProtKB protein evidence categories and distribution for the proteins in UniProtKB [all].

Comprehensive protein space

The content of sequence databases is in constant flux because of multiple reasons including altered gene models or experimental evidence (e.g. cDNA, protein level). When analyzing IPI v3.72 for the version numbers of IPI sequence identifiers, we found that about one third of IPI entries had undergone changes in their sequences (28915 of 86392 sequences have a version number greater than 1). This carries the risk that valid sequences are removed from a database version despite the possibility that the corresponding protein may physically exist. We therefore constructed an MScDB version from IPI Human versions 3.00 to v3.72 (92,934 entries) and compared it to IPI v3.72 (86,392 entries). This also allowed us to generate a peptide centric history for each IPI entry.

Searching the aforementioned human cell lines and placenta data set against both databases (Table 2) showed that >99% of all peptides are identical but that MScDB [ver] identified 132 peptides (115 protein groups; 87 peptides (68 protein groups) in placenta) not present in IPI [src] (Fig. 11A and Fig. 12A). 97 peptides (90 protein groups; 84 peptides (80 protein groups) in placenta) are in IPI [src] but not in MScDB [ver]. This is also reflected by an increase in the peptide to protein ratio from 15.46 (IPI [src]) to 17.88 (MScDB [ver]) (Fig. 9A and Fig. 10A).

The construction of a database (SC [ver]) using a basic sequence clustering approach (100% sequence identity) on the identical set of source databases (Table 1) results in 151 exclusive peptides (132 protein groups) in comparison to IPI [src] (20 exclusive peptides (20 protein groups), likely to be false positive identifications). The theoretical search space of MScDB [ver] is a subset of SC [ver]. The peptide centric clustering increases the proportion of unique peptides by 28% and the ratio of peptides to proteins from 10.20 to 17.89 (Fig. 10A). The 45% decrease in database size (Fig. 9A) consequently decreases the search time by 19% (509.08 to 414.89 s on a single file; Fig. 14B).

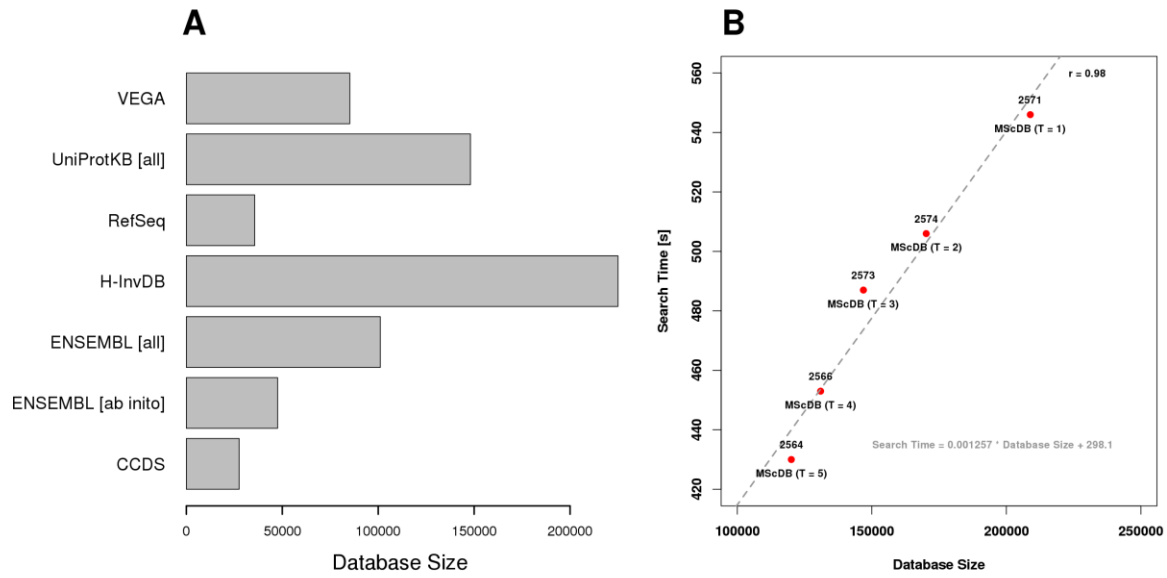

Figure 14. Common protein sequence databases. (A) Source databases of MScDB (T = 1, 2, 3, 4, 5). (B) Influence of the clustering parameter on the size of the database in linear correlation to the search time for one fraction (numbers indicate identified proteins over Mascot identity score).

In comparison to UniProtKB [cpl], MScDB [ver] results in 223 exclusive peptides (183 proteins) and SC [ver] in 250 (208 proteins) on the cell line data set (Fig. 11A). In total 256 peptides are not present in UniProtKB [cpl] (214 proteins). BLAST analysis of the 214 proteins (placenta analysis in Fig. 12C) found exclusively in MScDB [ver] and or SC [ver] showed that 11 are outdated IPI entries (12 peptides) but valid protein groups (Fig. 11C). Furthermore our data provides protein evidence for 116 protein groups (146 peptides) for the first time and we identified 3 alternative sequences and 37 (35 peptides) putative single amino acid polymorphisms. The 87 known protein groups (98 peptides) relate to sequence conflicts and natural variants.

The Mass Spectrometry centric Protein Sequence Database

The prior use cases of MScDB indicate a beneficial perspective on the sequence space using peptide instead of sequence clustering. Therefore we constructed an MScDB version of the latest and most popular proteome databases (Fig. 14A), to create a consensus of the theoretical human proteome. Additionally, we further investigated the influence of the clustering parameter (T) in constructing versions with T = {1, 2, 3, 4, 5}. MScDB (T = 2; 170,218 proteins) is almost comparable in size to UniProtKB [all] (148,079 proteins; Fig. 9A) but provides 50% more tryptic peptides even over the varying cluster parameter (Fig. 15C). The peptide to protein ratio is 14.55 in MScDB (T = 2) to 8.90 in UniProtKB [all] (Fig. 10B) and the ratio of unique peptides (once in database) is three to one.

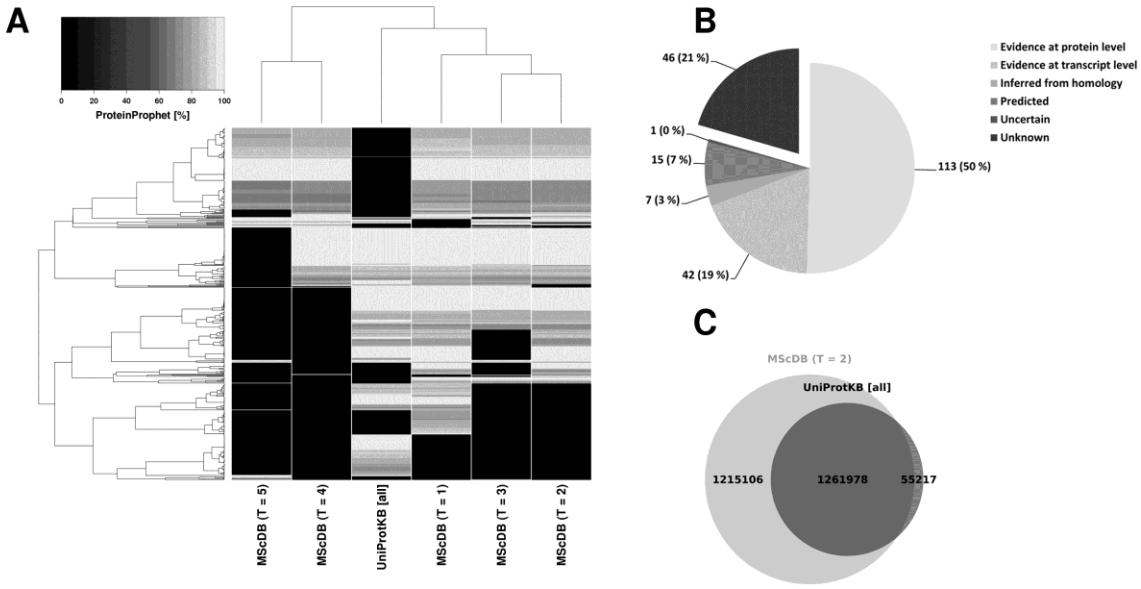

Figure 15. MScDB complement identifications and classifications based on the latest proteome databases against UniProtKB. (A) The heat map is a representation of the protein probabilities of complement peptides (not present in at least one database) over various protein sequence databases (Table 1), to show the overall confidence in peptide identifications. Peptides clustering together are likely to origin from the same protein (left dendrogram). Groups in the databases point to similar source databases or parameter settings (top dendrogram). (B) Classifications based on the UniprotKB protein evidence scheme for complement identifications in MScDB (T = 1, 2, 3, 4, 5) not present in UniProtKB [cpl] against UniProtKB [all]. (C) Number of in-silico peptides in MScDB (T = 2) in comparison to UniProtKB [all].

We were able to identify in the cell line data set 224 peptides complement (Fig. 15A; more robust identifications due to database size) to MScDB and 72 to the reference database UniProtKB [all] (Table 2). Out of these 72, only 8 were not present in MScDB, pointing to false positive identifications. Peptides with protein evidence (50%) are sequence conflicts or natural variants (Fig. 15B). 21% of our exclusive identifications are unknown in UniProtKB. An increasing cluster parameter decreases the number of exclusive identifications, the size of the database and the search speed (Fig. 14B). The parameters result in 207 (190 proteins) peptides in MScDB (T = 1), 176 (160 proteins) in MScDB (T = 2), 157 (143 proteins) in MScDB (T = 3), 150 (137 proteins) in MScDB (T = 4) and 142 (128 proteins) in MScDB (T = 5). Over all protein evidence categories, we identified 2 alternative sequences and 46 sequence conflicts. Most interestingly we discovered 41 (46 peptides) hitherto unknown proteins.

Cross-species protein identification

As illustrated above, MScDB clusters together proteins that are indistinguishable by mass spectrometry. This is not only useful for protein identification from a single organism but can also be used across organisms. For example, mouse is a common model organism to investigate basic biology as well as human diseases. In particular, human cancer derived cell lines are often grafted into mice to study tumor biology in-vivo. In such systems, protein identification is complicated by the fact that the grafts are comprised of mixed human and mouse proteomes and that many protein sequences of mice and man are very similar. We constructed an MScDB version from human IPI and

mouse IPI (v3.72 for both) resulting in a combined database with approximately 120,000 clusters (mismatch threshold zero, Fig. 16A). It is evident that most human and mouse proteins should be distinguishable by MS but also that the cluster composition in terms of taxonomy becomes more heterogeneous with increasing the mismatch parameter, leading to more and more orthologous proteins contained in the same cluster. We next used experimental data generated by Wei et al. [45] from a lung cancer xenograft study to identify proteins from the human/mouse MScDB.
From the total of 1,412 identified proteins, 80% were unambiguously assigned either to mouse (19%) or human (61%) and 20% of all proteins were not distinguishable between these organisms (Fig. 16B).

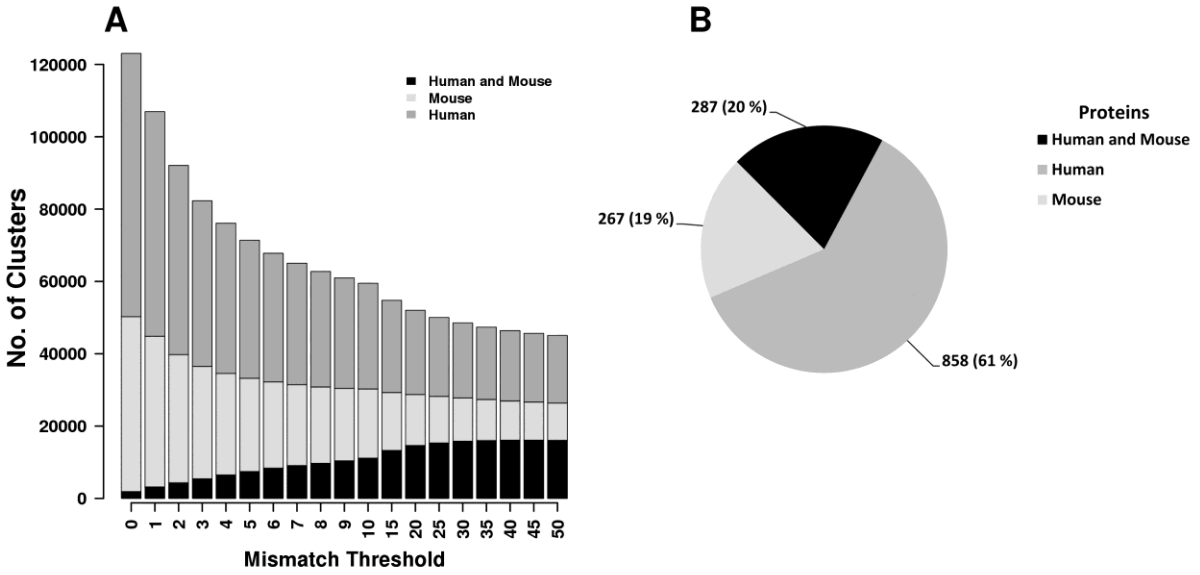

Figure 16. Cross-species protein identification. (A) Influence of the mismatch threshold on the taxonomic composition of MScDB clusters generated from human and mouse protein sequences. For increasing values of MM, more and more mouse and human proteins are clustered but the vast majority of proteins can be distinguished at the default parameter of T=2. (B) Experimental protein identification data from a human cancer xenograft model in the mouse showing that 80% of all proteins can be unambiguously assigned to one species.

Discussion

The common protein sequence databases used for mass spectrometry based proteomics strike a good balance between completeness and redundancy of the available protein sequence space. They all have in common that sequence similarity clustering is used to remove redundancy introduced by the vast number of sequences used for their construction (genomes, cDNA collections etc.). Redundancy reduction is important because redundancy aggravates the protein inference problem inherent to all bottom-up proteomics approaches. In our approach, we explored peptide sequence clustering as an alternative to protein sequence clustering to construct MScDB as a database for

protein identification. In clustering proteins on the basis of peptides that are distinguishable by mass spectrometry, MScDB is conceptually more aligned with the peptide centric procedure in which bottom up proteomics data is generated.

The results show that MScDB is at least as useful for protein identification compared to any of the other common database used for this purpose. In all use cases MScDB improves the peptide to protein ratio to one of the common reference databases (Fig. 10), such as IPI or UniProtKB, in a comparable protein sequence space (Fig. 9). In addition, peptide centric clustering removes much more redundancy from the source sequence collection than the traditional sequence clustering approaches. This not only enables faster database searching (a convenience) but also reduces the extent of the protein inference problem although we stress that it does not remove the issue altogether and that a posteriori methods to generate minimal protein lists are still required. MScDB also provides a convenient mechanism to harmonize legacy protein identification data with new experiments as MScDB provides a history of all its sequences (accession history for each protein, cluster affiliation and the original source annotation). Importantly, it further enables the recovery of outdated but valid protein sequences and our example of searching an MScDB version built from the combined IPIs v.3.00 – 3.72 using a complex human proteome digest shows that many such proteins genuinely exist.

As for any clustering approach, some sequences are ‘lost’ in the process of building MScDB. Although such sequences will be members of a cluster, they do not represent all the proteins in a cluster and are therefore not included in the FASTA file used for searching MS data. The extent to which this occurs obviously depends on the clustering parameters. For the default parameters used in this study, our comparisons indicate that the gains are larger than the losses and the MScDB code we provide allows scientists to choose the parameters at will (mass interval, mismatch, peptide sieve score). Hence, versions of the database can be generated to suit a particular purpose.

Over all human use cases we could identify 69 peptides (64 unknown and 5 alternative sequences) not in UniProtKB and 79 putative SAPs. We note that most of these identifications are from the consensus MScDB of the latest proteome databases. Our approach will gain more and more unknown evidence to date over experiments, regarding the efforts to describe the complete human proteome and as shown in Fig. 17 for the placenta and cell line data set which illustrates a minor intersection between these samples. We can therefore argue that MScDB is perhaps the better option for the discovery process of the complete human proteome than the well annotated UniProtKB.

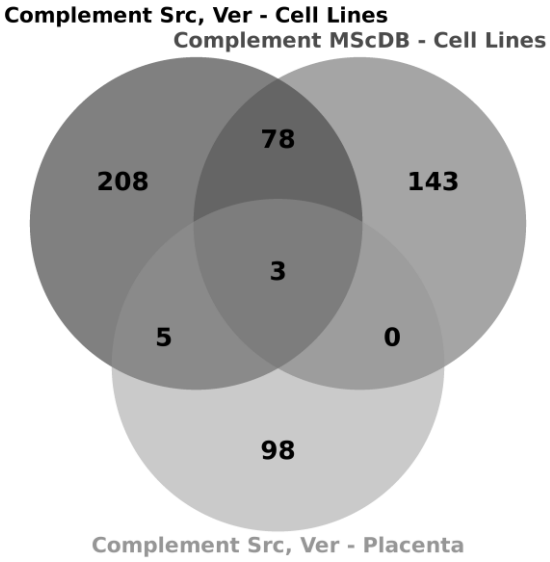

Figure 17. Intersection of complement peptides from MScDB use cases over the cell line and placenta data set.

MScDB focuses on the restructuring of the available sequence space from a MS centric perspective. In the process, MScDB addresses the protein inference problem in reducing the number of shared peptides but does not eliminate redundancy to the same extent as suffix arrays [50]. Furthermore, suffix arrays comprise a set of non-redundant peptides to search against, whereas MScDB uses the set of non-redundant peptides to construct a peptide-centric protein database and therefore does not reduce the peptide space to a pre-set in-silico digest parameter interval.

Perhaps the most interesting use case of MScDB that we have explored is the identification of proteins from systems of mixed species proteomes. Such experiments are not only relevant for the illustrated case of a xenograft cancer model, but also for the growing field of metaproteomics in general (e.g. the human microbiome) [51, 52]. We were somewhat surprised to learn that the level of uncertainty in assigning identified proteins to one species or the other is only about 20% indicating that experiments may not be as difficult to interpret as previously anticipated [53]. As far as we are aware, MScDB is the first demonstration of the merits of a fully peptide-centric view of the protein sequence space for proteomics. In light of the conceptual advantages, the presented data and the fact that the very popular IPI database has been discontinued [42], we consider MScDB to be a viable, indeed attractive alternative for bottom-up proteomics. To facilitate its use in the community, ready to use FASTA files for MScDB for human and mouse are available from our web site and we will continue to expand the list of available species in the future.

Acknowledgements

This research was in part funded by the DFG International Research Training Group "Regulation and Evolution of Cellular Systems" (GRK 1563). HM acknowledges the support of the TUM Graduate

School at the Technische Universität München, Germany and the authors thank Hannes Hahne for stimulating discussions.

Author contributions

Marx H., Rattei T. and Kuster B. designed the study. Lemeer S. and Klaeger S. performed experiments. Marx H. analyzed data. Marx H. and Kuster B. wrote manuscript.

Abbreviations

BHI	biological homogeneity index
CCC	connected component cluster
CCDS	consensus coding sequence
DAO	data access object
EBI	European bioinformatics institute
HCD	higher energy collisional dissociation
INSDC	International nucleotide sequence collaboration
IPI	International protein index
mgf	mascot generic format
MS	mass spectrometry
MS/MS	tandem mass spectra
MScDB	mass spectrometry centric protein sequence database
RC	representative cluster
RefSeq	reference sequence collection
SAP	single amino acid polymorphisms
SAX	strong anion exchange
SC	sequence clustering
SNP	single nucleotide polymorphism
UCSC	university of California Santa Cruz genome browser
UniProtKB	UniProt knowledgebase
UniRef	UniProt reference clusters
Vega	Wellcome Trust Sanger institute vertebrate genome annotation

References

1. Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol 2004, 5 (9), 699-711.
2. Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: an integrated database for proteomics experiments. Proteomics 2004, 4 (7), 1985-1988.
3. Magrane, M.; Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011.
4. Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Fingerman, I. M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Krasnov, S.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Karsch-Mizrachi, I.; Ostell, J.; Panchenko, A.; Phan, L.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; Wilbur, W. J.; Yaschenko, E.; Ye, J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2012, 40 (Database issue), D13--D25.
5. Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005, 33 (Database issue), D501--D504.
6. Karsch-Mizrachi, I.; Nakamura, Y.; Cochrane, G.; International, N. S. D. C. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2012, 40 (Database issue), D33--D37.
7. Leinonen, R.; Akhtar, R.; Birney, E.; Bower, L.; Cerdeno-Tárraga, A.; Cheng, Y.; Cleland, I.; Faruque, N.; Goodgame, N.; Gibson, R.; Hoad, G.; Jang, M.; Pakseresht, N.; Plaister, S.; Radhakrishnan, R.; Reddy, K.; Sobhany, S.; Ten Hoopen, P.; Vaughan, R.; Zalunin, V.; Cochrane, G. The European Nucleotide Archive. Nucleic Acids Res 2011, 39 (Database issue), D28--D31.
8. Tateno, Y.; Imanishi, T.; Miyazaki, S.; Fukami-Kobayashi, K.; Saitou, N.; Sugawara, H.; Gojobori, T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 2002, 30 (1), 27-30.
9. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. GenBank: update. Nucleic Acids Res 2004, 32 (Database issue), D23--D26.
10. Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Brent, S.; Chen, Y.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; Gordon, L.; Hendrix, M.; Hourlier, T.; Johnson, N.; Kähäri, A.; Keefe, D.; Keenan, S.; Kinsella, R.; Kokocinski, F.; Kulesha, E.; Larsson, P.; Longden, I.; McLaren, W.; Overduin, B.; Pritchard, B.; Riat, H. S.; Rios, D.; Ritchie, G. R. S.; Ruffier, M.; Schuster, M.; Sobral, D.; Spudich, G.; Tang, Y. A.; Trevanion, S.; Vandrovcova, J.; Vilella, A. J.; White, S.; Wilder, S. P.; Zadissa, A.; Zamora, J.; Aken, B. L.; Birney, E.; Cunningham, F.; Dunham, I.; Durbin, R.; Fernández-Suarez, X. M.; Herrero, J.; Hubbard, T. J. P.; Parker, A.; Proctor, G.; Vogel, J.; Searle, S. M. J. Ensembl 2011. Nucleic Acids Res 2011, 39 (Database issue), D800--D806.
11. Yamasaki, C.; Murakami, K.; Takeda, J.-i.; Sato, Y.; Noda, A.; Sakate, R.; Habara, T.; Nakaoka, H.; Todokoro, F.; Matsuya, A.; Imanishi, T.; Gojobori, T. H-InvDB in 2009: extended database

and data mining resources for human genes and transcripts. Nucleic Acids Res 2010, 38 (Database issue), D626--D632.

12. Wilming, L. G.; Gilbert, J. G. R.; Howe, K.; Trevanion, S.; Hubbard, T.; Harrow, J. L. The vertebrate genome annotation (Vega) database. Nucleic Acids Res 2008, 36 (Database issue), D753--D760.
13. Lamesch, P.; Berardini, T. Z.; Li, D.; Swarbreck, D.; Wilks, C.; Sasidharan, R.; Muller, R.; Dreher, K.; Alexander, D. L.; Garcia-Hernandez, M.; Karthikeyan, A. S.; Lee, C. H.; Nelson, W. D.; Ploetz, L.; Singh, S.; Wensel, A.; Huala, E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 2012, 40 (Database issue), D1202--D1210.
14. Pruitt, K. D.; Harrow, J.; Harte, R. A.; Wallin, C.; Diekhans, M.; Maglott, D. R.; Searle, S.; Farrell, C. M.; Loveland, J. E.; Ruef, B. J.; Hart, E.; Suner, M.-M.; Landrum, M. J.; Aken, B.; Ayling, S.; Baertsch, R.; Fernandez-Banet, J.; Cherry, J. L.; Curwen, V.; Dicuccio, M.; Kellis, M.; Lee, J.; Lin, M. F.; Schuster, M.; Shkeda, A.; Amid, C.; Brown, G.; Dukhanina, O.; Frankish, A.; Hart, J.; Maidak, B. L.; Mudge, J.; Murphy, M. R.; Murphy, T.; Rajan, J.; Rajput, B.; Riddick, L. D.; Snow, C.; Steward, C.; Webb, D.; Weber, J. A.; Wilming, L.; Wu, W.; Birney, E.; Haussler, D.; Hubbard, T.; Ostell, J.; Durbin, R.; Lipman, D. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 2009, 19 (7), 1316-1323.
15. Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007, 23 (10), 1282-1288.
16. Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22 (13), 1658-1659.
17. Enright, A. J.; Ouzounis, C. A. GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 2000, 16 (5), 451-457.
18. Mika, S.; Rost, B. UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res 2003, 31 (13), 3789-3791.
19. Paccanaro, A.; Casbon, J. A.; Saqi, M. A. S. Spectral clustering of protein sequences. Nucleic Acids Res 2006, 34 (5), 1571-1580.
20. Pipenbacher, P.; Schliep, A.; Schneckener, S.; Schönhuth, A.; Schomburg, D.; Schrader, R. ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics 2002, 18 Suppl 2, S182--S191.
21. Enright, A. J.; Van Dongen, S.; Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002, 30 (7), 1575-1584.
22. Li, W.; Jaroszewski, L.; Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001, 17 (3), 282-283.
23. Holm, L.; Sander, C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998, 14 (5), 423-429.
24. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25 (17), 3389-3402.
25. Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 2005, 4 (10), 1419-1440.

26. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002, 74 (20), 5383-5392.
27. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003, 75 (17), 4646-4658.
28. Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res 2010, 9 (10), 5346-5357.
29. Li, Y. F.; Arnold, R. J.; Li, Y.; Radivojac, P.; Sheng, Q.; Tang, H. A bayesian approach to protein inference problem in shotgun proteomics. J Comput Biol 2009, 16 (8), 1183-1193.
30. Ma, Z.-Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res 2009, 8 (8), 3872-3881.
31. Grobei, M. A.; Qeli, E.; Brunner, E.; Rehrauer, H.; Zhang, R.; Roschitzki, B.; Basler, K.; Ahrens, C. H.; Grossniklaus, U. Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Res 2009, 19 (10), 1786-1800.
32. Qeli, E.; Ahrens, C. H. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol 2010, 28 (7), 647-650.
33. Li, J.; Zimmerman, L. J.; Park, B.-H.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Network-assisted protein identification and data interpretation in shotgun proteomics. Mol Syst Biol 2009, 5, 303.
34. Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res 2007, 6 (9), 3549-3557.
35. Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 2007, 25 (1), 125-131.
36. Tang, H.; Arnold, R. J.; Alves, P.; Xun, Z.; Clemmer, D. E.; Novotny, M. V.; Reilly, J. P.; Radivojac, P. A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics 2006, 22 (14), e481--e488.
37. Alves, P.; Arnold, R. J.; Novotny, M. V.; Radivojac, P.; Reilly, J. P.; Tang, H. Advancement in protein inference from shotgun proteomics using peptide detectability. Pac Symp Biocomput 2007, 409-420.
38. Meyer-Arendt, K.; Old, W. M.; Houel, S.; Renganathan, K.; Eichelberger, B.; Resing, K. A.; Ahn, N. G. IsoformResolver: A peptide-centric algorithm for protein inference. J Proteome Res 2011, 10 (7), 3060-3075.
39. Zhou, A.; Zhang, F.; Chen, J. Y. PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. BMC Bioinformatics 2010, 11 Suppl 6, S7.
40. Schandorff, S.; Olsen, J. V.; Bunkenborg, J.; Blagoev, B.; Zhang, Y.; Andersen, J. S.; Mann, M. A mass spectrometry-friendly database for cSNP identification. Nat Methods 2007, 4 (6), 465-466.
41. de Souza, G. A.; Arntzen, M. Ø.; Fortuin, S.; Schürch, A. C.; Målen, H.; McEvoy, C. R. E.; van Soolingen, D.; Thiede, B.; Warren, R. M.; Wiker, H. G. Proteogenomic analysis of

polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. Mol Cell Proteomics 2011, 10 (1), M110.002527.

42. Griss, J.; Martín, M.; O'Donovan, C.; Apweiler, R.; Hermjakob, H.; Vizcaíno, J. A. Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. Proteomics 2011, 11 (22), 4434-4438.
43. Bantscheff, M.; Eberhard, D.; Abraham, Y.; Bastuck, S.; Boesche, M.; Hobson, S.; Mathieson, T.; Perrin, J.; Raida, M.; Rau, C.; Reader, V.; Sweetman, G.; Bauer, A.; Bouwmeester, T.; Hopf, C.; Kruse, U.; Neubauer, G.; Ramsden, N.; Rick, J.; Kuster, B.; Drewes, G. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol 2007, 25 (9), 1035-1044.
44. Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. Nat Methods 2009, 6 (5), 359-362.
45. Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem 2003, 75 (3), 663-670.
46. Wei, Y.; Tong, J.; Taylor, P.; Strumpf, D.; Ignatchenko, V.; Pham, N.-A.; Yanagawa, N.; Liu, G.; Jurisica, I.; Shepherd, F. A.; Tsao, M.-S.; Kislinger, T.; Moran, M. F. Primary tumor xenografts of human lung adeno and squamous cell carcinoma express distinct proteomic signatures. J Proteome Res 2011, 10 (1), 161-174.
47. Datta, S.; Datta, S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics 2006, 7, 397.
48. Team, R. D. C. R: A Language and Environment for Statistical Computing; Vienna, Austria, 2008.
49. Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011, 27 (3), 431-432.
50. Edwards, N.; Lippert, R.; Edwards, N. Generating Peptide Candidates from Amino-Acid Sequence Databases for Protein Identification via Mass Spectrometry. In Proceedings of the Second International Workshop on Algorithms in Bioinformatics, 2002; pp 68-81.
51. Wilmes, P.; Bond, P. L. Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 2006, 14 (2), 92-97.
52. Verberkmoes, N. C.; Russell, A. L.; Shah, M.; Godzik, A.; Rosenquist, M.; Halfvarson, J.; Lefsrud, M. G.; Apajalahti, J.; Tysk, C.; Hettich, R. L.; Jansson, J. K. Shotgun metaproteomics of the human distal gut microbiota. ISME J 2009, 3 (2), 179-189.
53. Mallick, P.; Kuster, B. Proteomics: a pragmatic perspective. Nat Biotechnol 2010, 28 (7), 695 709.

Chapter 3

A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics

Abstract

We present a peptide library and data resource of >100,000 synthetic, unmodified peptides and their phosphorylated counterparts with known sequences and phosphorylation sites. Analysis of the library by mass spectrometry yielded a data set that we used to evaluate the merits of different search engines (Mascot and Andromeda) and fragmentation methods (beam-type collision-induced dissociation (HCD) and electron transfer dissociation (ETD)) for peptide identification. We also compared the sensitivities and accuracies of phosphorylation-site localization tools (Mascot Delta Score, PTM score and phosphoRS), and we characterized the chromatographic behavior of peptides in the library. We found that HCD identified more peptides and phosphopeptides than did ETD, that phosphopeptides generally eluted later from reversed-phase columns and were easier to identify than unmodified peptides and that current computational tools for proteomics can still be substantially improved. These peptides and spectra will facilitate the development, evaluation and improvement of experimental and computational proteomic strategies, such as separation techniques and the prediction of retention times and fragmentation patterns.

Introduction

The gold standard in molecular analytical sciences for the identification of a new substance is the synthesis of a reference standard and to compare its physico-chemical properties to that of the analyte in question. If both molecules are identical in all measurable parameters, they are considered the same. However, the vast number of possible peptides derived from a proteome together with the phenomenal speed at which peptide identification data can be generated by state-of-the-art mass spectrometers has largely precluded the systematic validation of peptide identities by synthetic standards beyond relatively few examples with limited scope [1-6]. Instead, mass spectrometry based proteomics has come to rely on computational tools that match an experimental peptide tandem MS spectrum to in silico generated spectra derived from protein sequence databases using statistical models and empirical knowledge about peptide fragmentation behavior inside a mass spectrometer(reviewed in 7, 8). Albeit proven powerful, one can never be certain if the computational models correctly identify a given peptide (and thus a protein). Therefore, a number of probabalistic and decoy count models have been devised in order to estimate the false discovery rate (FDR) of peptide and protein identifications in a proteomic experiment [9-11]. Surprisingly, and to the best of our knowledge, none of the protein identification algorithms in popular use today have ever been rigorously validated on a significant number of synthetic peptide standards or controlled digests of known proteins leaving considerable room for uncertainty with respect to their performance [12]. However, considerable effort has gone into the use of several thousand synthetic peptides to characterize LC-MS/MS instrumentation and the physical properties that govern peptide identification [13] and at least a few approaches based on synthetic phosphopeptide libraries [14, 15] or individual phosphopeptides [16-18] have been taken to develop computer models for phosphorylation site localization within peptides. Still, concerns have been voiced that the numerically small and possibly biased sets of synthetic peptides used in these studies

may be insufficient to arrive at firm conclusions. To address the above shortcomings, we have synthesized 96 peptide libraries representing >100,000 peptides and their phosphorylated counterparts and analyzed these by high-performance liquid chromatograph tandem mass spectrometry (LC-MS/MS). We show that the physical library and the generated data can be used in numerous ways to develop, evaluate and improve proteomic experiments and data evaluation strategies. We are making the physical library and the mass spectrometric data publically available so that the research community can further explore and refine methodologies.

Material and methods

Data from public repositories

Tryptic phosphopeptides were selected from five large scale human phosphorylation studies [19-23] to compile a representative set of 96 peptides that formed the basis for the synthesis of peptide libraries (Fig. 1). Briefly, the initial step in the selection process created a subset of 851 peptides present in at least three studies. Subsequently, 96 peptides residing in a 5th to 95th percentile interval of length and hydrophobicity as estimated by the GRAVY Score 47 were picked evenly across this area and also balancing the number of C-Terminal Arginine (n=47) and Lysine (n=49) residues as well as a similar distribution of the site of phosphorylation. In case the sample peptides contained multiple sites of phosphorylation, one of them was picked at random.

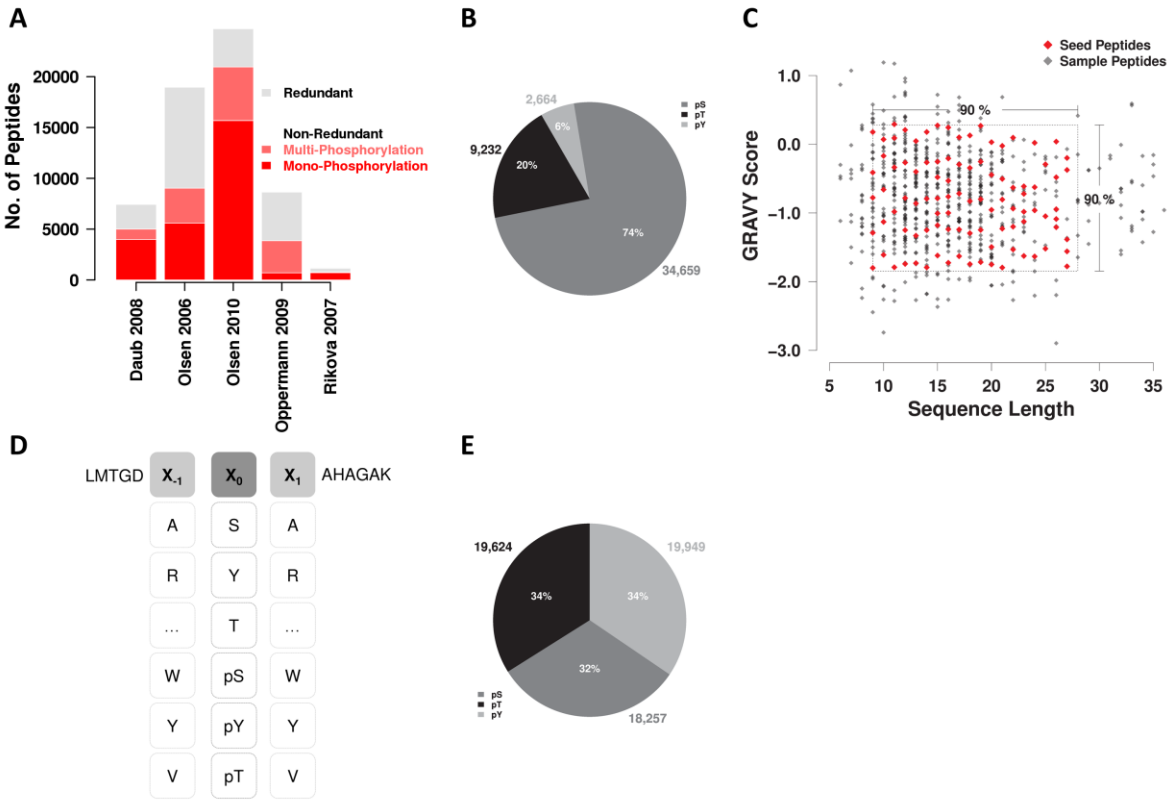

Figure 1. Library design and synthesis. (A) Number and distribution of identified mono-and multi phosphorylated peptides from published large scale datasets, that were used for the sample peptide and seed peptide selection in this study. Reported datasets contained many redundant peptides (gray) that were removed prior to sample selection. (B) Number and relative abundance of non-redundant serine, threonine and tyrosine phosphorylation sites identified in five large-scale human phosphoproteomic datasets [19-23]. (C) Hydrophobicity (GRAVY score) plotted against sequence length for the 851 peptides (black diamonds) identified in three out of the five large-scale datasets. The 96 representative 'seed' peptides in the 5–95% percentile interval (dashed box) were selected manually for subsequent library synthesis. Selected peptides are depicted in red, showing a representative distribution of both length and hydrophobicity. The selection of peptides also contains a representative distribution of phosphorylation sites of the sequence and a representative distribution of lysine or arginine residues at the C terminus. (D) Schematic representation of the peptide library design in which position x_0 of a seed peptide represents the site of phosphorylation and is synthesized with either S, T or Y or their phosphorylated forms pS, pT or pY. Both positions x_{-1} and x_{+1} are permutated with all 20 natural occurring amino acids during synthesis, creating up to 2,400 different (phospho)peptides for each library (E) Number of serine (dark grey), threonine (black) and tyrosine (light grey) phosphorylated peptides and their relative abundance, identified from LC-MS/MS analysis of the library using both HCD and ETD fragmentation. In total 57,830 phosphopeptides with equal representation of all phosphate acceptor amino acids were identified.

Library design and synthesis

Each of the so called 96 seed peptides represents a permutation template for the generation of 96 peptide libraries. The applied permutation scheme incorporates (S, T, Y, pS, pT, pY) at the position of the phosphosite (X_0)(in the original peptide from the literature) and the 20 standard amino acids in the direct vicinity (X_{-1} and X_{+1}) resulting in libraries of size 2,400 peptides (n=84) or 120 peptides (n=12) in case the phosphorylation site is either at the peptide N-terminus or directly N-terminal to the C-terminal Lys and Arg residues of the tryptic peptides. The total number of theoretical peptides across the 96 combinatorial libraries amounts to 203,040. The combinatorial libraries were synthesized at 2 µmol scale by standard solid-phase synthesis following the Fmoc strategy on a parallel peptide synthesizer (Intavis, Cologne). Fmoc protected amino acids were obtained from Intavis. Specifically, Fmoc-Ser(PO(OBzl)OH)-OH, Fmoc-Thr(PO(OBzl)OH)-OH and Fmoc-Tyr(PO(OBzl)OH)-OH were used as building blocks for the synthesis of the phosphopeptides. Briefly, the synthesis started with a C-Terminal tryptic amino acid (Arginine, Lysine) and then proceeds sequentially to concatenate single amino acids, except at the permutation site(s), where isokinetic mixtures of amino acids (20 or 6) were incorporated to attempt to create a discrete uniform distribution. An acetylation step was added after each synthesis cycle to block any remaining free amino group in order to prevent the synthesis of mixed sequences. Following completion of synthesis, peptides were deprotected and released from the solid phase using 92.5% trifluoro acetic acid, 5% tri-isopropyl silane , 2.5% water. Crude synthetic peptide libraries were subjected to Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) without further work up or purification.

Liquid Chromatography Tandem Mass Spectrometry

Peptide library HCD data were acquired by coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent, Dublin, CA) to an Oribtrap Velos (Thermo Scientific, Bremen, Germany). Peptides were delivered to a trap

column (100 µm i.d. × 2 cm, packed with 5µm C18 resin, Reprosil PUR AQ, Dr. Maisch, Ammerbuch, Germany) at a flow rate of 5 µL/minute in 100% buffer A (0.1% FA in HPLC grade water). After 10 minutes of loading and washing, peptides were transferred to an analytical column (75µmx40 cm C18 column Reprosil PUR AQ, 3µm, Dr. Maisch, Ammerbuch, Germany) and separated using a 110 minute gradient from 2% to 35% of buffer B (0.1% FA in acetonitrile) at 300 nL/minute flow rate. Full scan MS spectra were acquired in the Orbitrap at 30,000 resolution. The 5 most intense precursors were selected for HCD fragmentation (isolation width 2.0 Th) with normalized collision energy of 40% at an AGC target setting of 50,000. HCD spectra were acquired in the Orbitrap at 7,500 resolution. Dynamic exlusion was enabled for a 10 s repeat duration and a 10 s exclusion duration with a repeat count of 1.

Peptide Library ETD data (with Orbitrap readout, ETD-FT) were acquired on an ETD enabled Orbitrap Velos instrument (Thermo Fisher Scientific, Bremen) connected to a UHPLC Proxeon EASY-nLC 1000 (Thermo Scientific). Peptides were trapped on a double-fritted trap column (Dr. Maisch ReproSil C18, 3 µm, 2 cm x 100 µm) and separated on an analytical column (Agilent Zorbax SB-C18, 1.8 µm, 40 cm x 75 µm). Solvent A consisted of 0.1 M acetic acid (Merck), solvent B of 0.1 M in 80 % acetonitrile (Biosolve). Samples were first loaded at a maximum pressure of 980 bar with 100 % solvent A. Peptides were separated by a 110 min gradient from 10% to 40% of buffer B at a flow rate of 200 nL/minute. Full scan MS spectra were acquired at 30,000 resolution. The 5 most intense precursors were selected for ETF FT fragmentation (isolation width 1.5 Th) at 7,500 resolution and a target setting of 100,000. Supplemental activation was enabled, the activation time was 50 ms and target setting was 300,000 for the ETD reagent. Dynamic exlusion was enabled for a 30 s repeat duration and a 30 s exlusion duration with a repeat count of 1.

As an example for a biological sample, phosphopeptides from a total of 2 mg desalted K562 digest were enriched using Ti4+-IMAC as described (Zhou et al. (2013) Nat Protocols, 8, 1-22, in press). Briefly, the enriched sample was analyzed on a nanoLC-MS/MS platform consisting of UHPLC instrument (EASY-nLC 1000, Thermo) connected to a Q Exactive quadrupole orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen), in a 3 hour run. Mass spectra were acquired with an automatic switch between a full scan (target value 3,000,000, resolution 35,000) and up to 20 most intense ions were sequentially isolated and accumulated to a target value of 50,000 with a maximum injection time of 120 ms and were fragmented by HCD at a normalized collision energy of 25%. The spectra of the fragmented ions were acquired in the Orbitrap analyzer at a resolution of 17,500.

MS data processing and database search for Mascot

Raw MS data files were converted into Mascot generic format files (MGF) using Mascot Distiller (2.4.2.0, www.matrixscience.com). Important parameters included: i) signal to noise ratio of 20 for MS/MS and ii) time domain off (no merging of spectra of the same precursor). The MGF files were searched against human IPI v3.72 including the sequences of all 96 libraries, each comprising concatenations of all theoretically possible peptides within a synthesized library) using the Mascot search engine (2.3.1, www.matrixscience.com). Important parameters for ETD and HCD: decoy search using a randomized version of the human IPI v3.72 including the sequences of all 96 libraries was enabled; monoisotopic peptide mass (considering up to two 13C isotopes); trypsin/P as protease; a maximum of four missed cleavages; peptide charge +2 and +3; peptide tol. +/- 5 ppm; MS/MS tol. +/- 0.02 Da; instrument type ESI-Trap (for HCD data) or ETD-Trap (for ETD data) respectively; Cys residues were unmodified; variable modifications: oxidation (M), phospho (ST),

phospho (Y). The result files were exported to pepXML and Mascot XML (www.matrixscience.com) with default options provided by Mascot.

MS data processing and database search for PhosphoRS

PhosphoRS site localization was performed using the PhosphoRS 2.0 embedded in the Proteome Discoverer 1.3 software (Thermo Fisher Scientific, Bremen, Germany). Raw MS data files were converted into Mascot generic format files (MGF) using Mascot Distiller (2.4.2.0) as described above. The MGF files were searched against human IPI v3.72 (supplemented with additional 96 entries, each comprising concatenations of all theoretically possible peptides within a synthesized library) using the Mascot search engine (2.3.1) embedded in the Proteome Discoverer 1.3 software. In the spectrum selector node, the unrecognized mass analyzer replacements was set as FTMS and the unrecognized activation type replacements as HCD or ETD, respectively. Search settings for Mascot were identical to described above. Phosphorylation site localization was performed on the Mascot results using PhosphoRS 2.0. The result files were exported to csv format. Note that the authors had no opportunity to influence extra data processing steps potentially performed by Proteome Discoverer such as spectral filtering or grouping. We suspect that some data processing not transparent to the authors is performed in Proteome Discoverer because the number of spectra/PSMs in the PhosphoRS result files are not the same as for our Mascot analysis despite the fact that both analysis were performed on the same input files (i.e. mgf produced by Mascot Distiller, see above).

MS data processing and database search for Andromeda

MaxQuant, version 1.3.0.3 was used to generate peak lists from the MS/MS spectra for database searching. High-resolution profile MS/MS data was deconvoluted before extraction of the ten most abundant peaks per 100 Th. All statistical filters in MaxQuant such as peptide and protein false discovery rates and mass deviation filters were disabled in order to score all submitted MS/MS spectra. Peptide masses were recalibrated by MaxQuant prior to Andromeda searches. Peak lists were searched against human IPI v3.72 (supplemented with additional 96 entries, each comprising concatenations of all theoretically possible peptides within a synthesized library). Oxidation (M), phosphorylation (STY) were used as variable modifications. A mass tolerance of 5 ppm was used for the peptide mass. Both HCD and ETD data were searched with a 0.02 Da tolerance window. Trypsin/P was set as proteolytic enzyme and a maximum of four miss cleavages were allowed. The MS/MS.txt output file of the software was used for further data analysis

Data analysis

The pepXML (Mascot) and MS/MS.txt (Andromeda) files provide detailed information about the database search results including precursor intensity, retention time and charge state, etc. as well as up to 10 peptide spectrum matches (PSMs; Mascot; 15 for Andromeda). The PSMs were classified into true positive (TP: PSM with the highest score that matches to a sequence that is present in the library) and false positive (FP: PSM with the highest score that matches to a sequence that is NOT present in the library, i.e. it matches to a sequence in the IPI human database). To get an overview of the global detection of synthesized peptides (Fig. 3), we counted all TPs as defined above. For all subsequent data analysis steps, we counted TPs and FPs as follows: for single PSMs with highest ranking score, we count the highest ranking PSM as one FP (or TP if it is not a library peptide) and

ignore all PSMs with lower scores. For multiple highest ranking PSMs (i.e. identical score), three scenarios are possible: i) all PSMs are TPs in which case we count one TP to avoid inflating the number of TPs; ii) all PSMs are FPs in which case we count one FP to avoid inflating the number of FPs; and iii) PSMs are a mix of TPs and FPs in which case we count one TP and one FP. For the retention time analysis, we used the retention time value associated with the highest intensity of a particular precursor ion at the time it was picked for MS/MS. The average LC peak width across the LC separation was 10-25s full width at half maximum (FWHM) which is why we considered peptide isomers eluting within 25s to be indistinguishable by retention time. All statistical analysis was performed using the R Statistical Programming Language (www.r-project.org). The False Discovery Rate (FDR) was calculated as the ratio of FP to the sum of FP and TP. To calculate the local FDR, the score distributions were divided in bins of width one (Mascot score, Mascot delta score, Andromeda score, PTM-score). In case a bin contained less than 100 classifications, the width was extended by one as often as necessary to meet the minimum criterion. A similar criterion applies for computing global FDRs, calculating from highest to lowest score until at least 100 classifications are accumulated. The determination of local and global False Localization Rates (FLR) for MD-score, phosphoRS and Andromeda followed the same scheme as the FDR calculation described above. The Mascot-Delta Score 17 was calculated by subtracting the highest ranking PSM from the subsequent PSM. Curve fitting was performed using the Non-Least Square function in the stats library of R with the prot algorithm and the formula FDR (or FLR) = A * exp-B * Score + C * exp-D * Score. Initial parameters were estimated for the nonlinear modeling using the optim function in the stats library. Curve fitting was confined to the meaningful intervals of the FDR/FLR plots.

Results

Peptide library design and synthesis

The guiding principles in the design of the library were (i) to represent the typical peptide sequence and composition space of bottom-up proteomics, (ii) to be sufficiently large to enable rigorous statistical treatment, (iii) to contain an equal representation of unmodified and S, T, Y phosphorylated peptides and (iv) to contain a large number of isomeric sequences (by amino acid sequence/composition) to enable the simultaneous assessment of a range of analytical parameters. We reviewed five large-scale human phosphoproteomic studies [19-23] representing some 40,000 phosphopeptide identifications and 46,000 phosphorylation sites dominated by pS and pT phosphorylation (Fig. 1A, B and Fig. 2). We selected the 851 sample peptides common to at least three of the five studies and analysed them for hydrophobicity vs. length (Fig. 1C). From this plot, we selected 96 representative 'seed' peptides that cover 90% of this area and also contain the phosphorylation sites in representative positions along the peptide sequence. Each of the 96 seed peptides formed the basis for the synthesis of a library (Fig. 1D) in which one amino acid position x_0 was permuted to contain the six amino acids S, T, Y, pS, pT and pY. The flanking positions before (x_{-1}) and after (x_{+1}) were permuted by all 20 naturally occurring amino acids. This design resulted in 84 libraries with 2,400 members each and a further 12 libraries with 120 members each, totalling 203,040 theoretical unmodified and corresponding phosphorylated peptides.

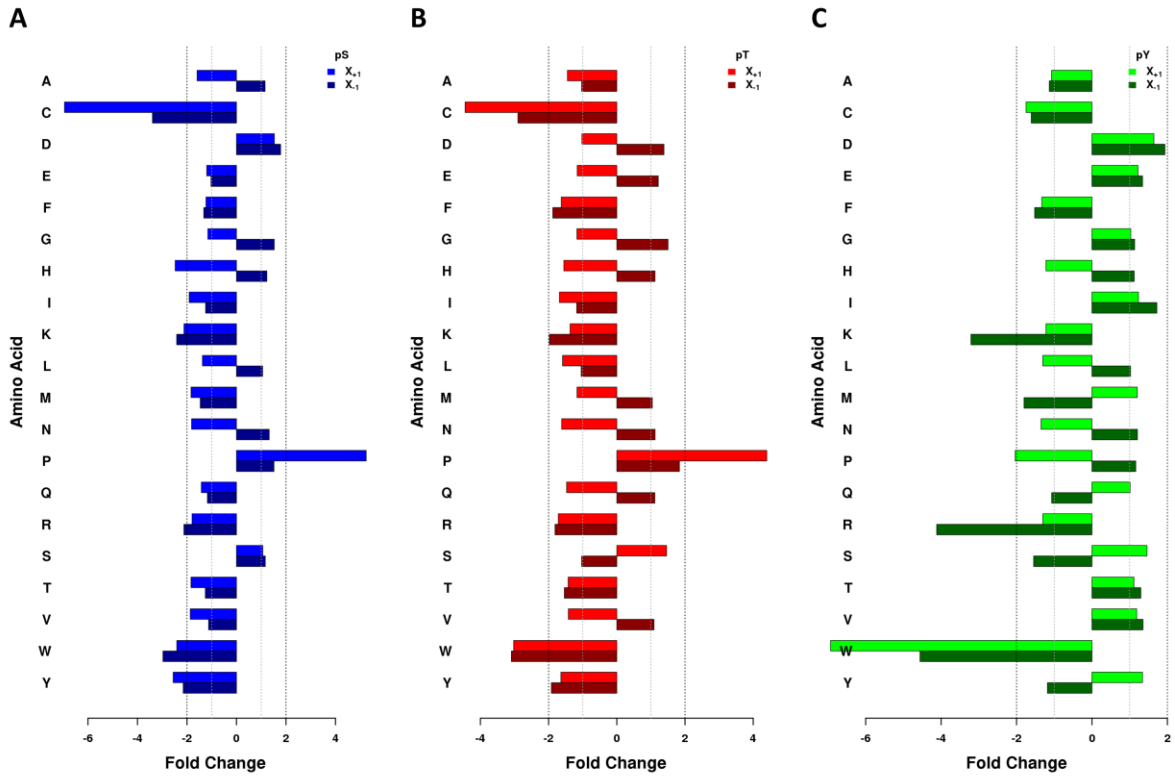

Figure 2. Amino acid composition analysis at the x_{+1} and x_{-1} position of the phosphorylation site in (A) serine, (B) threonine and (C) tyrosine phosphorylated peptides that were identified in the 5 large scale phosphopeptide datasets (Fig. 1A). Fold changes were calculated on the basis of the amino acid composition of the IPI protein database. Proline is clearly overrepresented at the x_{+1} position for pS/pT peptides reflecting the abundance of SP, TP phosphorylation motifs in these data sets. Cysteine residues are clearly underrepresented at position x_{+1} and x_{-1} in both pS/pT containing, but not in pY containing peptides. Tryptophan is underrepresented at the x_{+1} and x_{-1} in pY containing peptides. The reasons for these over- and under-representations have not been investigated further.

Peptide and phosphopeptide identification by LC-MS/MS

Each of the synthesized libraries was subjected to LC-MS/MS analysis on an Orbitrap Velos instrument using either beam-type collision induced dissociation or electron transfer dissociation with Orbitrap readout of fragment ion spectra. Database searching using Mascot revealed that all but three libraries were synthesized successfully with an average detection efficiency of 63% (Fig. 3a).

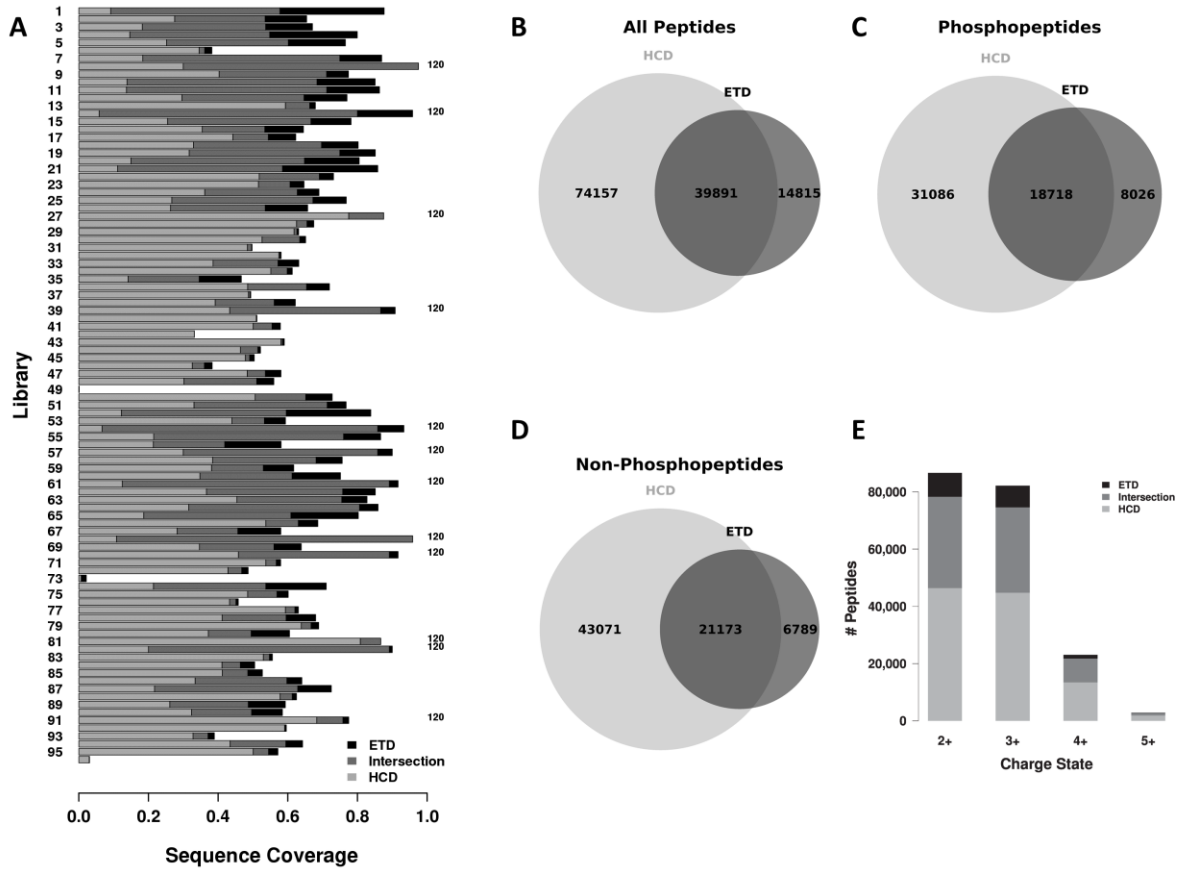

Figure 3. Peptide library identification rate. (A) Total sequence coverage of each peptide library showing an overall identification rate of 63% by Mascot (70% for unmodified peptides, 57% for phosphopeptides, not FDR adjusted). Libraries that were based on seed peptides with C- or N-terminal phosphorylation sites only contain a maximum of 120 (phospho) peptides and are therefore marked as such. (B-D) Venn diagrams showing the overlap between peptides (b), phosphopeptides (C) and non-phosphorylated peptides (D) identified by HCD and ETD fragmentation. (E) Number of peptides identified from each precursor charge state (2+, 3+, 4+ and 5+) by HCD only (light grey), ETD only (black) or by both methods (dark grey). Further information can be found in Figure 4.

The HCD and ETD data collectively provided raw analytical evidence (i.e. not FDR adjusted but requiring a correct library sequence and the correct phosphorylation site where applicable, Fig. 4) for 128,863 library peptides comprising 71,033 non-redundant unmodified peptides (70% of all possible) and 57,830 non-redundant phosphopeptides (57% of all possible, Fig. 3B-D) which formed the basis for all subsequent analysis.

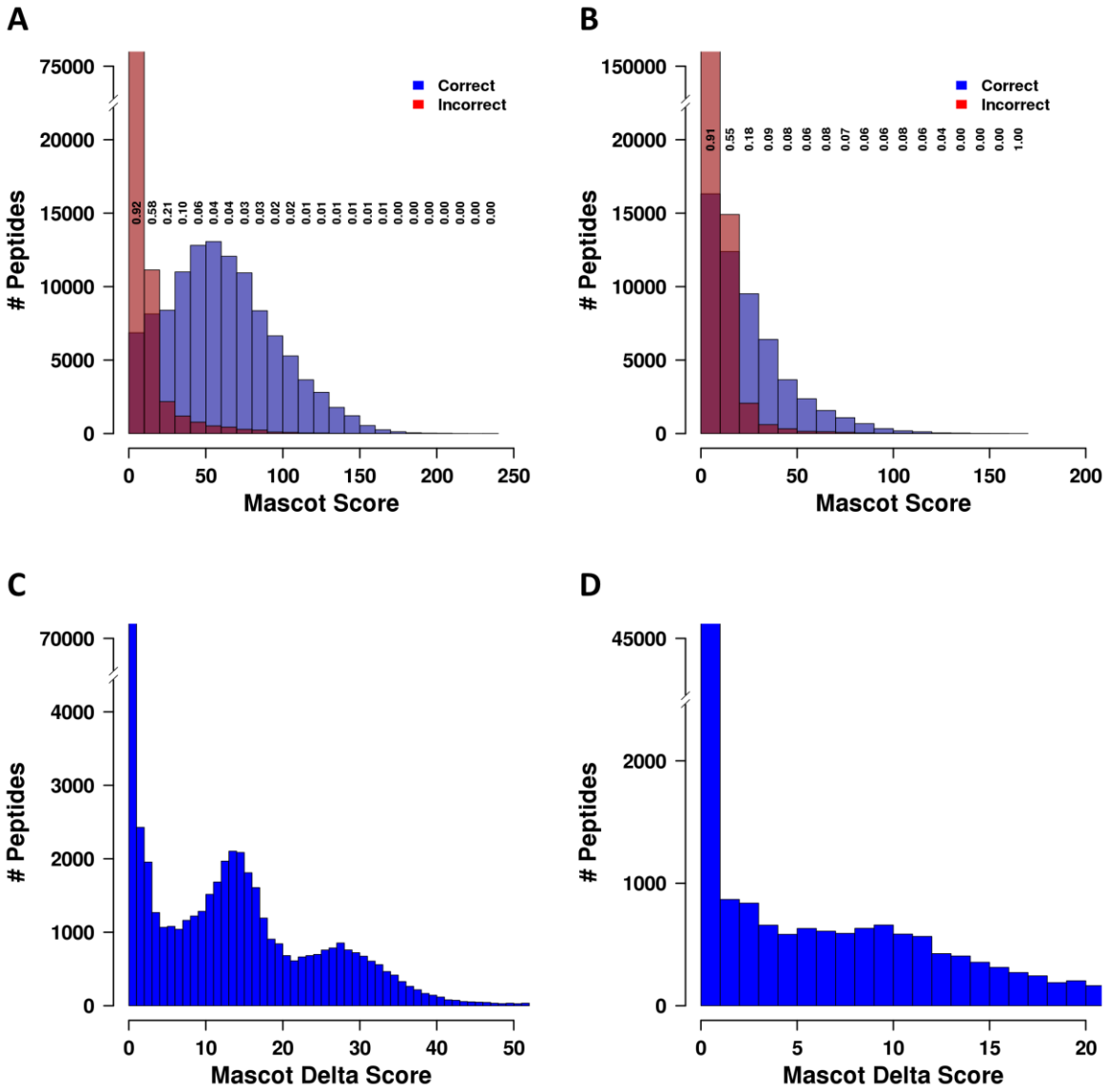

Figure 4. Peptide classifications. (A) Distribution of correct (true positive, TP) and incorrect (false positive, FP) peptides as a function of the Mascot Score (HCD). FDR values are indicated within each score bin (bin width 10 score points) (B) analogous to (A) for ETD data. (C) MDscore distribution of TP peptides. 43,672 out of 114,048 peptides have an MDscore different from zero (HCD). (D) analogous to (C) for ETD data; 12,771 out of 54,706 peptides have an MDscore different from zero.

The detected peptides show equal representation of pY, pS and pT and no significant compositional or phosphorylation motif bias (Fig. 1E). This is both notable and important as it can now be shown that the pS or pT peptides are not substantially more difficult to identify by (high resolution) mass spectrometry than pY peptides and that the dominance of SP and TP phosphorylation motifs in large-scale data sets from biological sources are not owing to bias in the LC-MS/MS readout of such studies. This very large library of synthetic peptides with known sequences and modification sites along with the associated mass spectrometric data enables a multitude of investigations relevant for proteomics, some of which are outlined below. Many peptides were identified by both HCD and ETD

fragmentation and Orbitrap readout but HCD identified substantially more peptides and with better Mascot scores (Fig. 3A-D and Fig. 5). The observed peptide charge state distribution follows expectations in light of the synthesis design (Fig. 3E). But contrary to commonly accepted notions in the field, ETD with Orbitrap readout provided no appreciable advantage for peptide identification at any charge state, and HCD was also substantially more successful for the analysis of phosphopeptides (Fig. 3C).

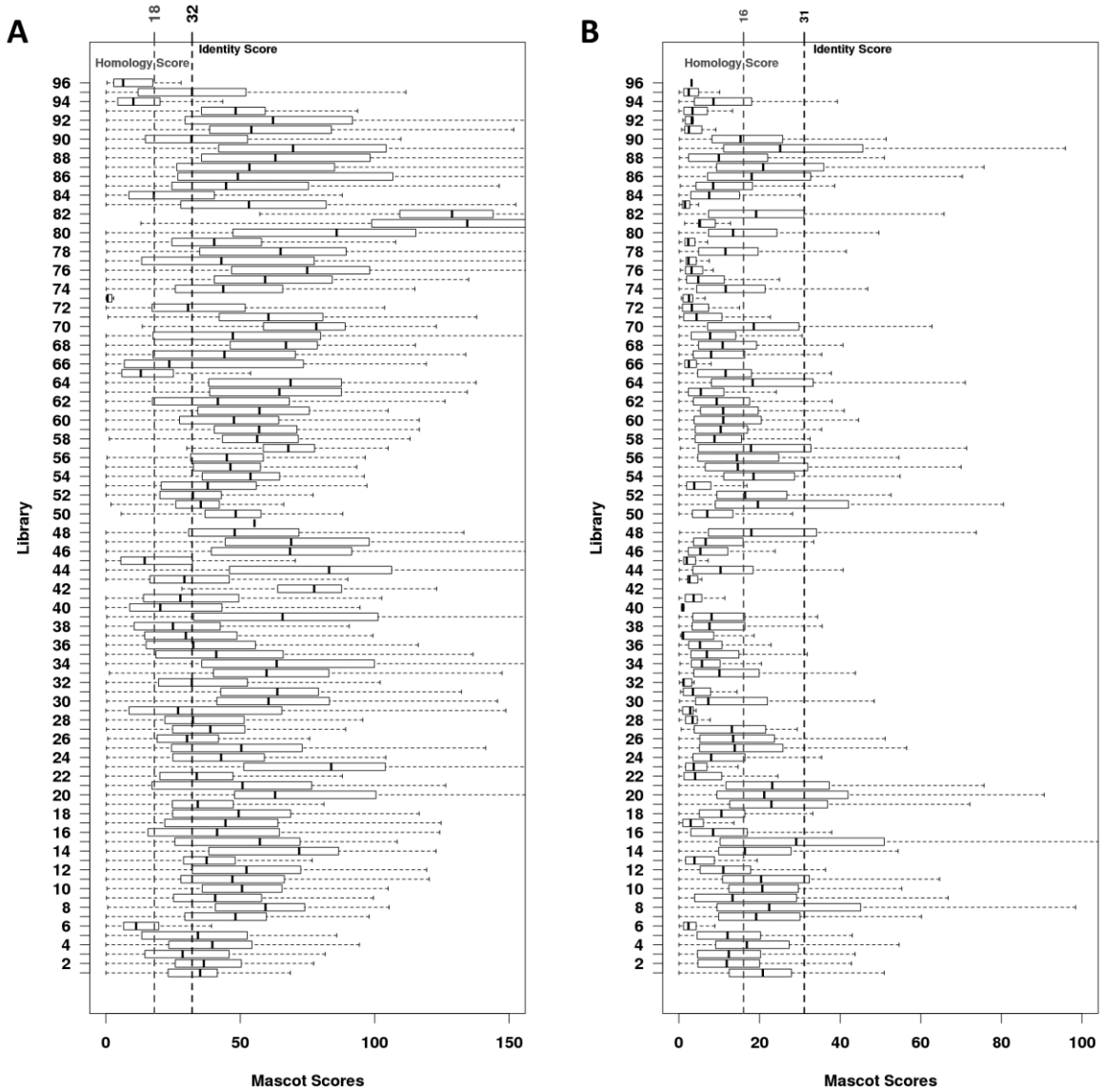

Figure 5. Score distributions over libraries. Mascot score distributions of peptides identified in each library by either (A) HCD or (B) ETD as fragmentation method. Medians are marked with black lines and boxes indicate the middle two quartiles of the score distributions. For comparison, Mascot identity and homology scores are indicated by dotted lines. For HCD, most of the peptides are identified with scores higher than the Mascot Identity score. As depicted in panel b, Mascot scores are overall significantly lower when ETD is used as fragmentation technique.

Evaluation of peptide identification algorithms

Owing to the fact that our peptide library is very large and the sequences of the synthetic peptide standards are known, we were able to address a fundamental issue in proteomics—namely, the merits of peptide identification by database searching. Using the popular search engines Mascot and Andromeda, we counted the number of correct and incorrect identifications to derive models for local and global false discovery rates as a function of the search engine score (Figs. 6-8). As expected, both score distributions show a rapid drop of FDRs as the search engine score increases. Notably, phosphorylated peptides appear to be easier to identify than the corresponding unmodified peptides as phosphopeptides show significantly lower FDR values at any search engine score.

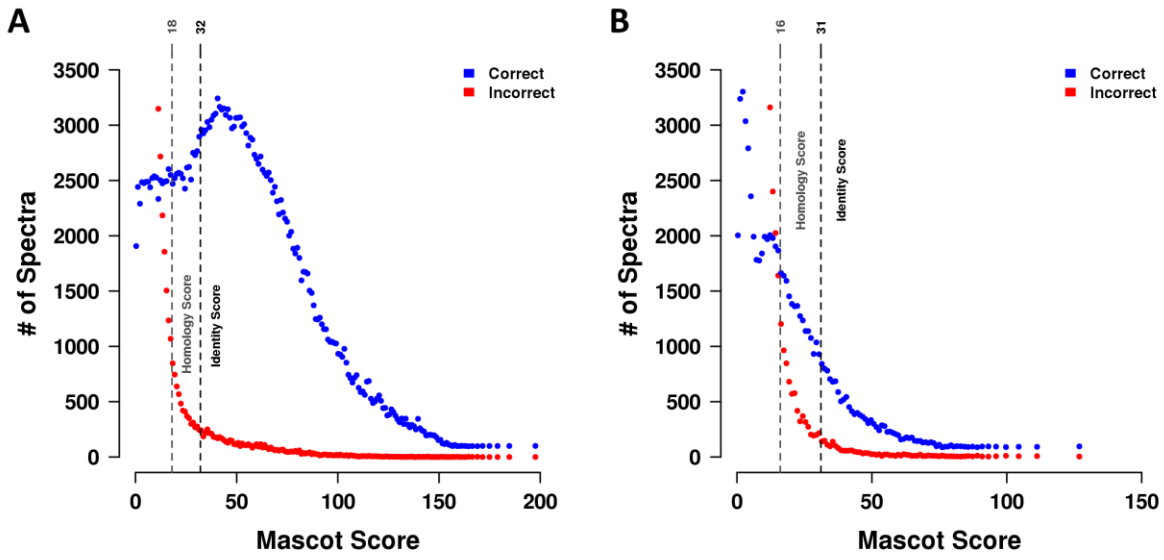

Figure 6. Spectra classification. Number of correct (blue) and incorrect (red) assigned spectra for each Mascot score, using either HCD (a) or ETD (b) as fragmentation technique. As can be clearly seen in both figures, for low Mascot scores, i.e. below Identity threshold, the number of incorrect assigned spectra rapidly increases. For Mascot scores higher then the Identity threshold, incorrect assignments rapidly decrease.

An important consequence of this analysis is that actual global or local FDR values can now be easily computed from fit functions representing a sum of two exponentials of the form FDR=A*exp(–C*Score)+B*exp(–D*Score) [17] for any peptide identification made by these two search. With the appropriate adaptations, such calculations can likely be applied to scores from any other search engines as well. A comparison of global FDR values (computed as above) with the decoy count approach used by Mascot shows that the decoy approach underestimates the genuine global FDR for most of the 96 libraries analyzed by a factor of 1.5 to 3. This raises questions about the general validity of a decoy count approach or at least about how decoy count methods should be devised and 'calibrated' to reflect the true FDR. The resource we provide in this work enables such valuable future investigations.

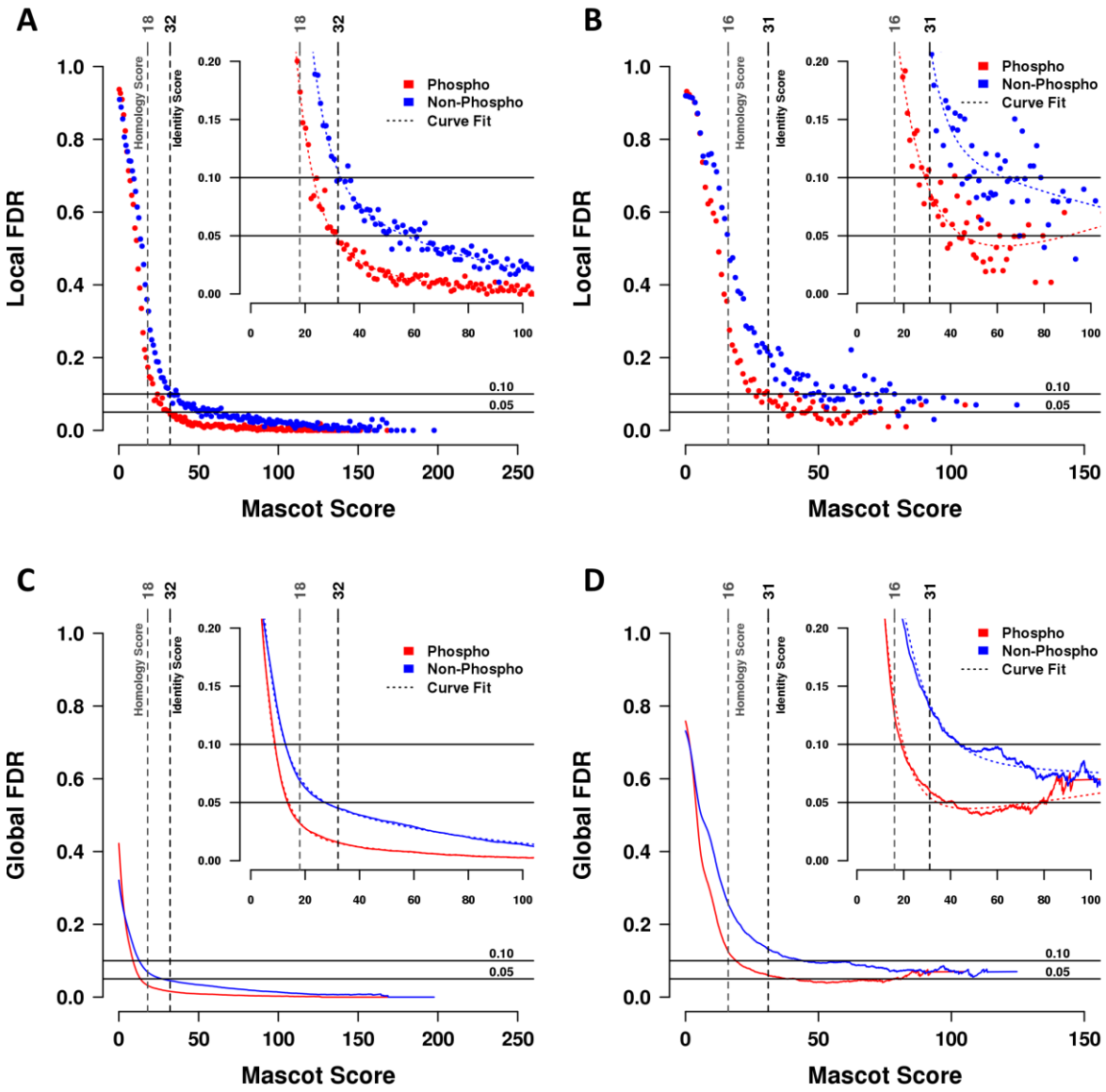

Figure 7. Local and global false discovery rate (FDR) analysis as a function of the Mascot score. Peptide identifications using (A, C) HCD and (B, D) ETD as fragmentation technique. Phosphorylated peptides are marked in red and non phosphorylated peptides are marked in blue. For HCD, the Mascot identity score nicely correlates with a 5% local FDR for phosphopeptides and a 10% local FDR for non phosphopeptides. The ETD data follows the same trend but with generally higher FDR values at a given Mascot score. Insets show expanded regions of the same plots. The score distributions can be approximated with a fit function of the form: FDR= A * exp(-C * Score) + B * exp(-D * Score). This allows the calculation of a local and global FDR for any identification at a particular Mascot score.

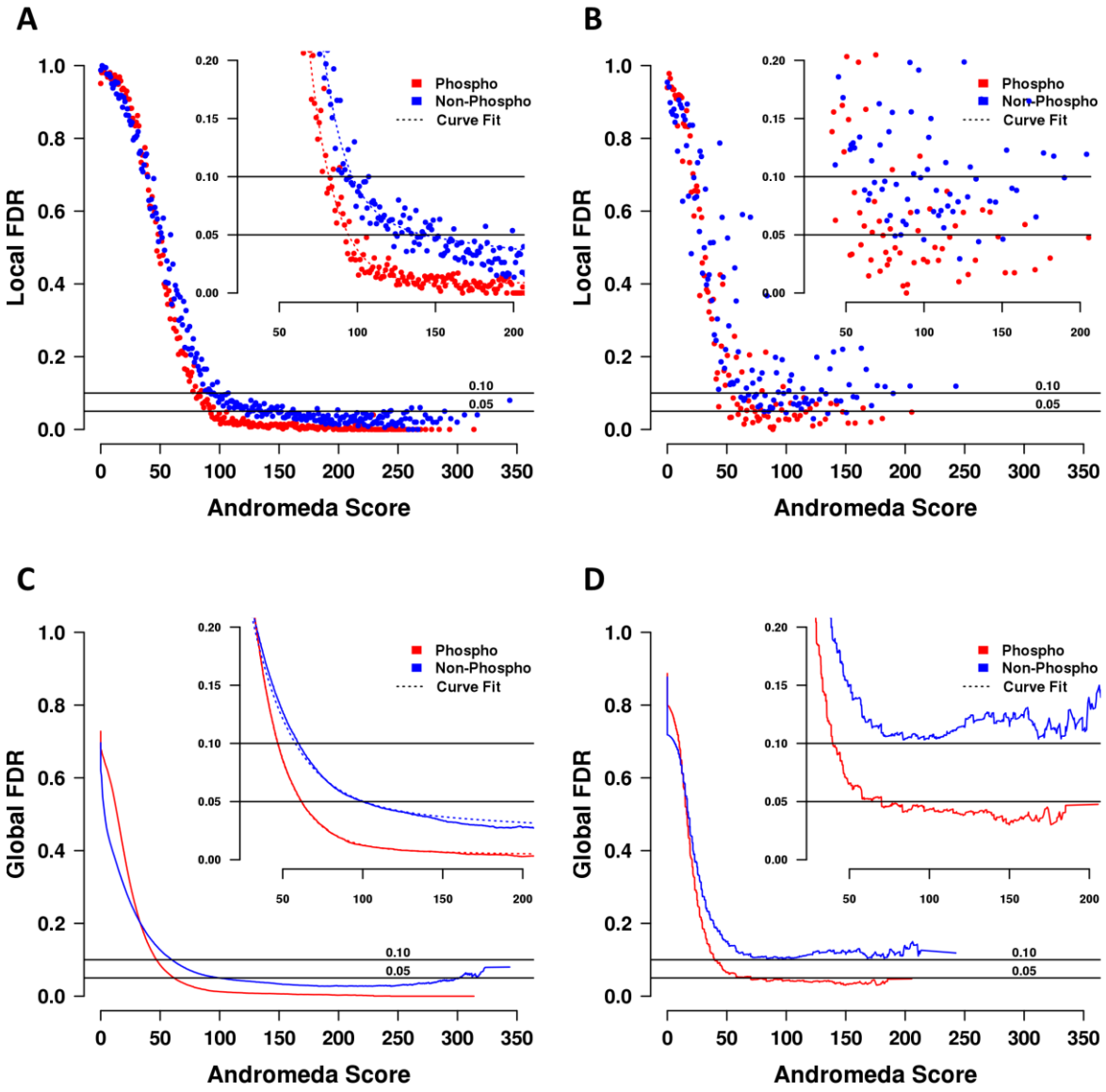

Figure 8. Local and global false discovery rate (FDR) analysis as a function of the Andromeda score. Peptide identifications using (A, C) HCD and (B, D) ETD as fragmentation technique. Phosphorylated peptides are marked in red and non phosphorylated peptides are marked in blue. For HCD, the Mascot identity score nicely correlates with a 5% local FDR for phosphopeptides and a 10% local FDR for non phosphopeptides. The ETD data follows the same trend but with generally higher FDR values at a given Mascot score. Insets show expanded regions of the same plots. The score distributions can be approximated with a fit function of the form: FDR= A * exp(-C * Score) + B * exp(-D * Score). This allows the calculation of a local and global FDR for any identification at a particular Mascot score.

Phosphopeptide site localization

Previous work has shown that synthetic phosphopeptides can be used to generate phosphorylation site localization scores[14, 15, 17]. Although these tools have proven to be very useful, the relatively small number of synthetic peptides used in these studies confines the scores to global assessments, thus leading to uncertainty with respect to their accuracy for individual phosphopeptides. The volume of

data generated in this study allowed us to refine the Mascot Delta Score (MD-score) for phosphorylation site localization in a number of ways.

It is now possible to compute false localization rates (FLR) for global data sets as well as individual peptides using the same general equation as the one used for FDR calculations (Fig. 9). The results of the analysis of this bigger data set largely confirm the earlier studies [17] and extended its scope to ETD with Orbitrap readout. Furthermore, although ETD overall identifies fewer phosphopeptides than HCD (Fig. 3), the site localization for peptides that are identified by ETD tends to be more accurate (Fig. 9A, B). Notably, there are, albeit relatively small, differences in localization performance for the different phosphorylated amino acids (Fig. 9C, D), which should be investigated in more detail in the future. We noted before [17] that confident phosphorylation site localization becomes more difficult when two possible acceptor amino acids are directly adjacent. The larger data basis in the current study confirmed this for HCD data and also showed that no appreciable such effect is observed for ETD.

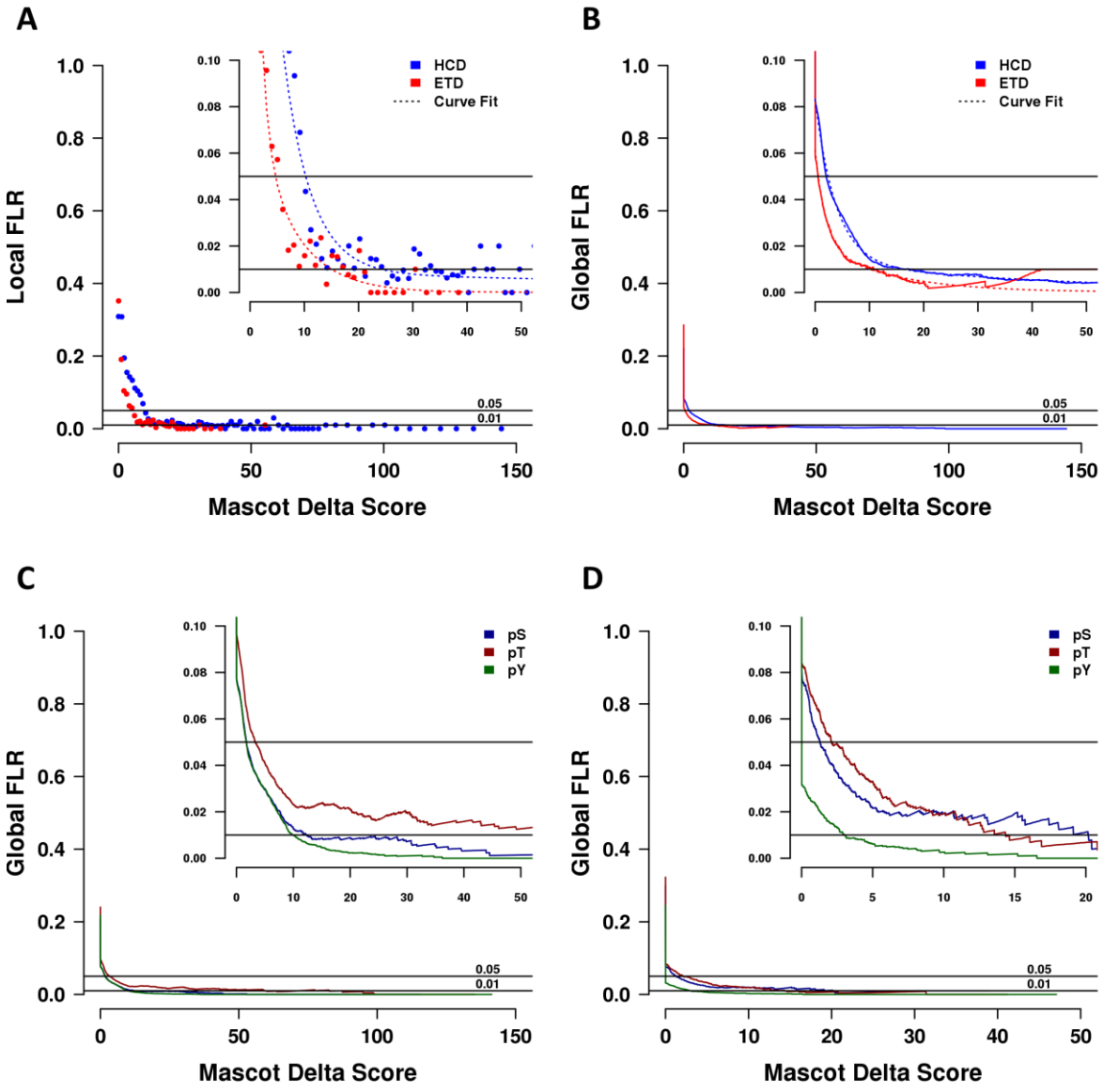

Figure 9. False phosphorylation site localization rate analysis for HCD and ETD data as a function of Mascot Delta score. (A) Plot showing Local FLRs in each MDscore bin and (B) plot showing global FLRs. ETD performs slightly better on site localization than HCD. For example, for a local FLR of 1%, ETD requires a MDscore of 15 whereas HCD requires a MDscore of 25. Insets show expanded regions of the same plots. The fit functions are the same as in Supplemental Figure 8 but with different coefficients. Insets show expanded regions of the same plots. (C, D) Global false localization rates as a function of Mascot delta score for serine (blue), threonine (red) and tyrosine (green) phosphorylated peptides. For both HCD (C) and ETD (D). Insets show expanded regions of the same plots. Slight differences are observed for the different phospho amino acids with pY localization being more easily correctly localized than serine (2nd best) and threonine phosphorylation sites, i.e. lower MDscores are required for pY peptides compared to pS and pT peptides to maintain the same FLR.

We next applied our library data to benchmark three phosphorylation site localization tools (the MD-score, the PTM-score of MaxQuant and its associated search engine Andromeda and the phosphoRS score embedded in Proteome Discoverer). Analysis of PSMs classified by all three localization tools shows that at 1% FLR all three tools cover 90–95% of the correct spectra (Fig. 10A, B). We then examined how well the probability reported by a given localization tool correlates with the FLR determined by counting correct and incorrect spectra (Fig. 10C). Of note, all localization tools underestimate the true FLR within a probability bin, suggesting that they can be improved. Reassuringly, however, the absolute error is small for the vast majority of the data. PhosphoRS appears to be a special case as the probability distribution is particularly narrow. As we used a commercial implementation of PhosphoRS (in Proteome Discoverer), the reasons for this unexpected behavior are unclear at present. Apart from the reliability with which a phosphorylation site can be called, it is obviously also important to know how many phosphopeptides can be correctly assigned by each of the tools. As can be seen in the inset of Figure 10A, the PTM-score and phosphoRS are more sensitive than the MD-score as they achieve the highest number of localized library peptides at 1% FDR/FLR. Notably, there is substantial complementarity between all localization tools, suggesting that their interpretation of the underlying tandem mass spectra follows different rules and that they can be improved. As a result, and at least for the time being, using more than one localization tool may lead to a more comprehensive analysis of a given phosphoproteome.
We also applied the FDR/FLR models derived for the two search engines and three localization tools to a complex phosphoproteomic sample generated by Ti-IMAC enrichment from human K562 cancer cells. And similar to what was obtained for the library peptides, the different identification and localization tools showed substantial complementarity. The data also suggests that when applying very stringent requirements (e.g., 1% peptide identification FDR and 1% phosphorylation site FLR), the overall number of identified and localized phosphopeptides is comparable between the three tools (Fig. 10D).

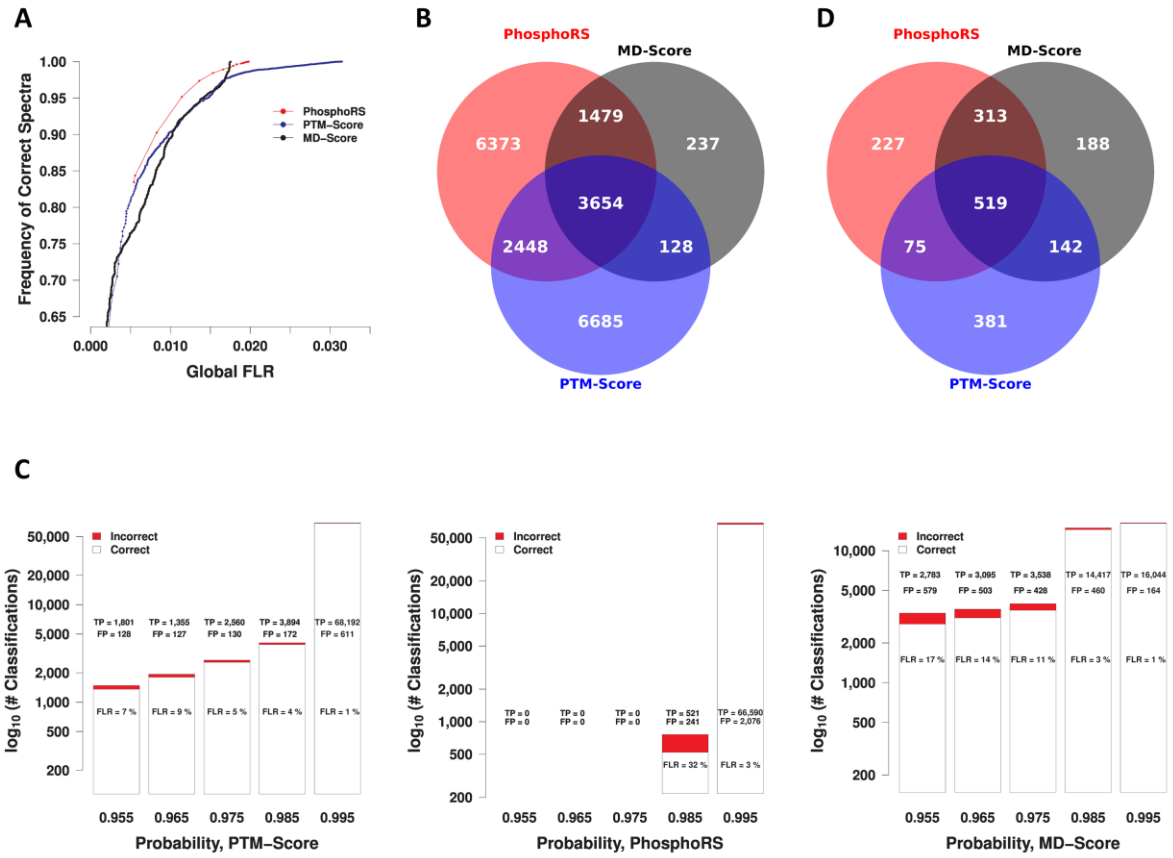

Figure 10. False localization rate determination for phosphorylated peptides. Because the modification site of all library peptides is known, false localisation rates (FLRs) can be determined by counting the number of correct and incorrect matches. (A) Qualitative and quantitative comparison of PTM-score, MD-score and PhosphoRS for phosphorylation site localization (HCD data). Although all three scores exhibit comparable overall accuracy (using data from the intersection of all three tools), (B) the Venn diagram in the inset shows the complementarity of the different tools at 1% FDR and 1% FLR. (C) Histogram plot of the number of correctly and incorrectly assigned spectra within probability bins provided by the three localization tools. The graphs show that all localization tools underestimate the true FLR within most bins but also indicate that this error is small for the vast majority of the data (see Supplementary Fig. 20 for further information). (D) Application of the FDR and FLR models derived from library spectra to the analysis of a phosphoproteomic sample generated by Ti-IMAC enrichment from human K562 cells. The results confirm the complementarity of the different localization scores at the level of 1% FDR and 1% FLR shown here.

Peptide separation by liquid chromatography

Peptide separation by high performance liquid chromatography is an integral part of mass spectrometry based proteomic workflows. The design of our peptide library also facilitates the study of a number of topics in this area.

First, the library contains a matched collection of peptides and their phosphorylated counterparts. More than 95% of the 11,381 peptide pairs (5% global FDR) from the HCD data set shown in Figure 11A and 11B for ETD data were separated by the LC system (depending on gradient length and chromatographic resolution). As noted before [24], the majority of all phosphopeptides elute

substantially later than their unmodified counterpart when using formic acid or acetic acid in the LC mobile phase , but the extent of increased retention is not correlated with peptide length (Fig. 11A). However, we did observe that phosphorylation on Ser and Thr increases the retention time substantially more than does phosphorylation on Tyr (Fig. 11C), possibly because Tyr is the most hydrophobic amino acid of the three.

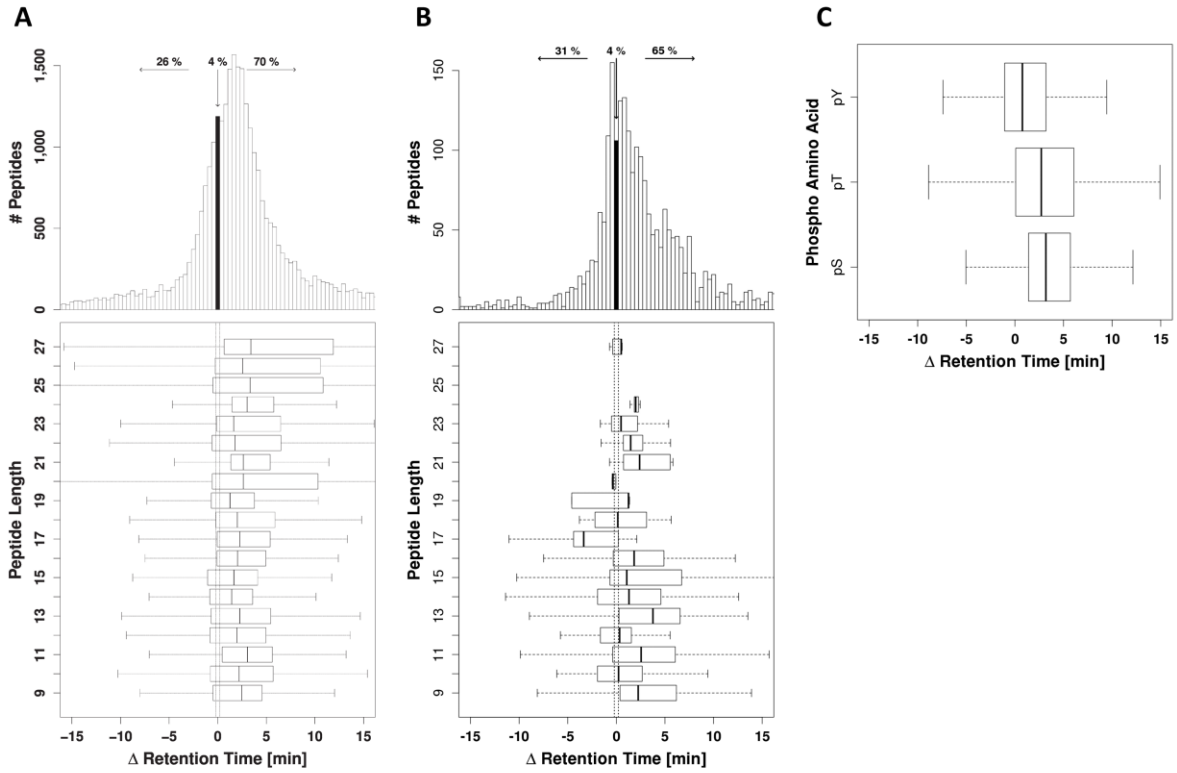

Figure 11. Retention time analysis. (A) Distribution of the retention time shift introduced by the addition of a phosphate group to a peptide (HCD data, delta RT values were calculated by subtracting the retention time of the unmodified peptide from that of the corresponding phosphorylated peptide). The majority of phosphopeptides (70%) elute later than their non-phosphorylated counterparts, 4% elute within the same time window and 26% elute earlier (25 second window, corresponding to the width of an average LC peak at half maximum). Retention time shifts appear to be independent of peptide length (B) Distribution of retention time differences between an unphosphorylated peptide and its phosphorylated counterpart for ETD data collected on an Orbitrap Velos instrument and using acetic acid as the ion pairing agent in the LC solvent. Also for acetic acid, the vast majority of the peptides can be separated by the LC system (96%) and most phosphopeptides elute significantly later than their unphosphorylated counterparts. Only 4% of all peptides cannot be separated by the LC system (LC resolution as measured by the width of an LC peak at half maximum is ~25 seconds. (C) Contribution of the phosphate group to peptide retention. As can be seen, the contribution of the phosphate group to the overall retention of a peptide is small for pY peptides but significantly larger for pS and pT peptides.

Second, the library contains tens of thousands of paired sequence isomers because the x_{-1} and x_{+1} positions were systematically permuted over all 20 amino acids. Isomeric peptides represent a challenge for mass spectrometric identification because their precursor masses are identical and the tandem mass spectra may also be rather similar if the discriminating b- and y-type fragment ions are missing, which can lead to ambiguity (Figs. 4, 12A and 12B). However, sequence isomers of the same amino acid composition may be separated by reversed phase chromatography. The one-dimensional LC system employed in this study allowed the separation of about 40% of all isomers (Fig. 12C; 45,516 isomeric pairs, 5% global FDR, HCD data) using a conservative 25 s retention time window, within which eluting isomers were considered indistinguishable by chromatography (average LC peak widths were generally between 10 and 25 seconds; full width at half maximum, FWHM). Notably, we observed a massive under-representation of glycine in position x_{-1} and x_{+1} for these peptides, suggesting that an individual glycine residue exerts little or no influence on peptide retention on the stationary phase. About 60% of isomeric pairs could not be fully separated by the one dimensional LC system (Fig. 12C) but the information contained in the tandem mass spectrum allowed the assignment of 30% of these peptides (judged from non-identical Mascot scores). This is either because isomeric peptides did not exactly co-elute so that the respective tandem mass spectra are dominated by one species or because one isomer was more abundant than the other (i.e. higher synthesis yield). Isobaric, or near isobaric peptides may also arise from two amino acid combinations having the same or similar mass (e.g. AS/GT, same mass or DG/AT, similar mass). Because of the high mass accuracy afforded by the instrument used in this study for both precursor and fragment ions, most of the ‘similar mass’ cases can be readily distinguished by their precursor or fragment ion masses alone. Given the differences in contribution of individual amino acids to the retention of a peptide (see below), many such near-isobaric peptides can also be separated by a one dimensional, and likely much better by a two dimensional LC system.

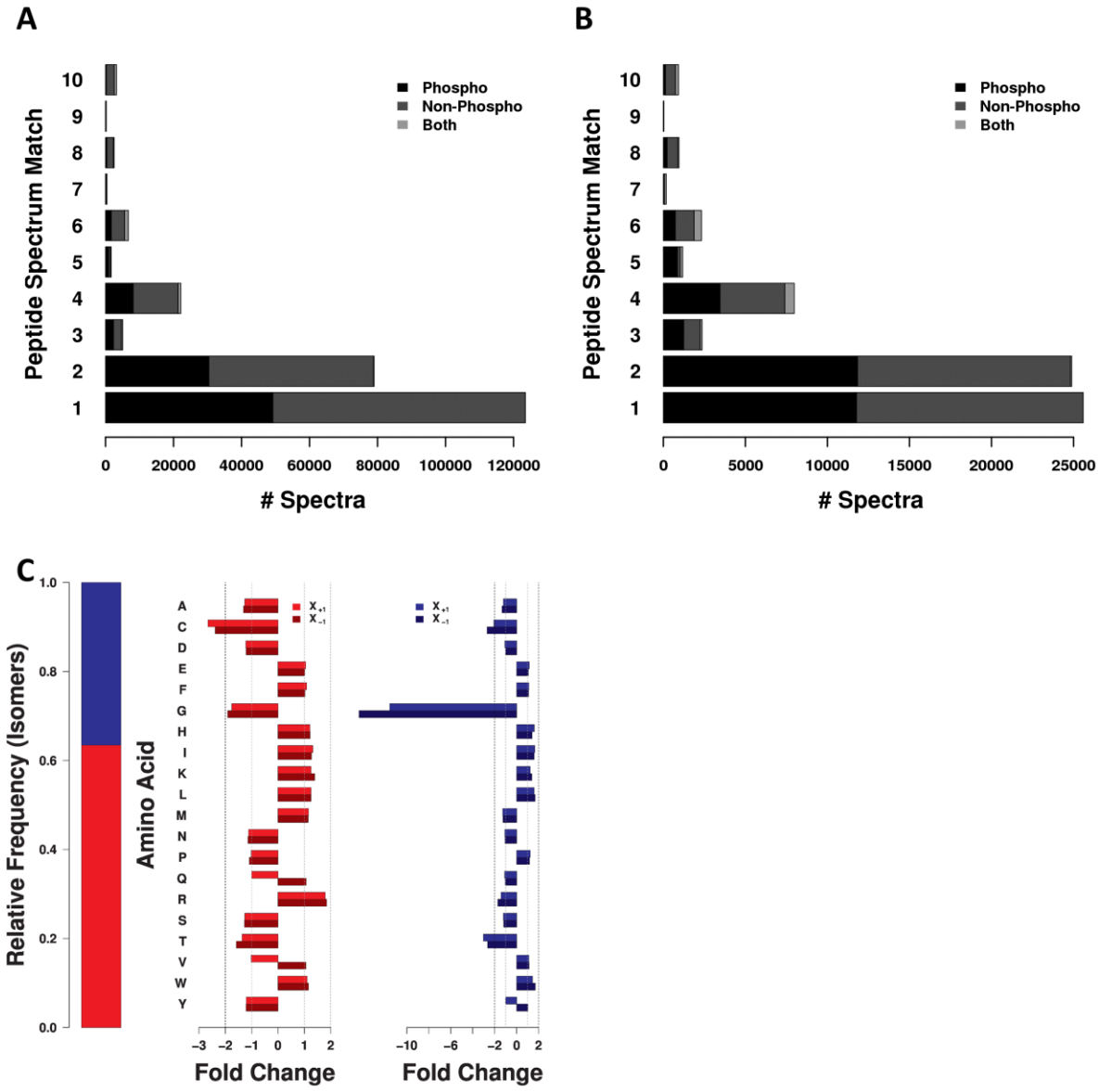

Figure 12. Isomer analysis. (A) Analysis of the number of peptide spectrum matches (PSMs) for tandem mass spectra (assigned by Mascot) for both HCD and (B) ETD. It is apparent that a large number of spectra only result in a single PSM (i.e. matching peptide sequence). Spectra generating 2 PSMs (with identical score) originate primarily from isomeric peptides generated by the permutation scheme used for the library synthesis. Spectra generating more than 2 PSMs are rare. The relative over-representation of 4, 6, 8, 10 PSMs also arise from the permutation scheme as there are a small number of amino acid compositions of identical mass (e.g. AS/GT) which may not always be distinguishable by the information contained in a tandem mass spectrum. For the FDR/FLR calculations used in this work, these cases are treated as one TP and one FP in order to avoid artificially inflating the number of TPs or FPs. (C) The library contains a large number of paired sequence isomers (x_{-1} and x_{+1} position around the phosphorylation site; n=45,516 pairs, 5% global FDR). Sixty-two percent of these positional isomers cannot be distinguished by a difference in retention time (red) whereas the remaining 38% can (blue, left panel. For the indistinguishable positional isomers, no clear over or underrepresentation of amino acids at the x_{+1} and X_{-1} positions are detectable (middle panel). In contrast, for the paired positional isomers that can be distinguished by retention time (right panel), a strong underrepresentation of glycine (G) at the x_{+1} and x_{-1} positions is observed.

Third, we asked the question how much retention to the reversed phase material of the LC column is added by any of the individual amino acids. Based on 156 complete permutation sets (i.e. peptides in which all of the x_{-1} and x_{+1} amino acid permutations were observed), the trend closely follows the described hydrophobicity scales of amino acids during RP-HPLC published by Krokhin [25] with tryptophan and phenylalanine adding the most and histidine and lysine adding the least retention to a peptide (Fig. 13). It is likely that more factors influencing chromatographic behavior of tryptic peptides can be extracted from the data, which may be subject to future investigation.

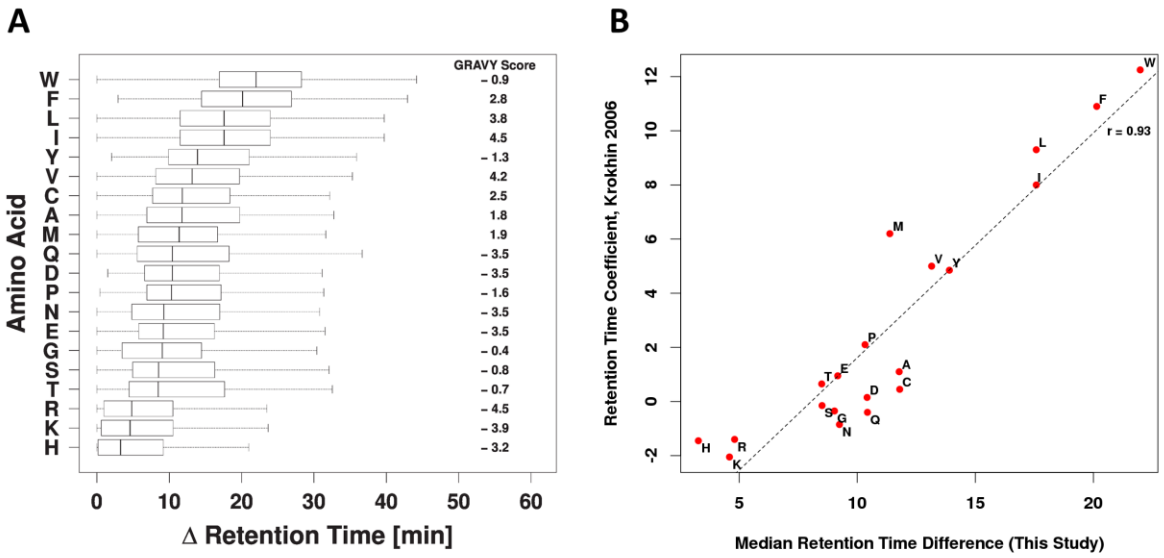

Figure 13. Amino acid retention time behavior. (A) The influence of individual amino acids on retention behavior was studied using peptides for which all 156 possible amino acid permutations in the x_{+1} and x_{-1} positions were observed by HCD. The observed trend clearly follows the general hydrophobicity and charge properties of the amino acids as approximated by the GRAVY score and reported by others (see Supplementary Fig. 23 for further details). (B) Contribution of individual amino acids on retention time shifts observed during reversed phase chromatography. Data derived in this study correlates very well with data derived by Krokhin (Anal Chem, 2006) (r=0.93).

Discussion

We have created the largest physical synthetic peptide library (>200,000 theoretical members) and associated mass spectrometric data set (>50,000 identified phosphopeptides, >70,000 identified unmodified peptides) available for proteomic research to date. We are making this resource available to the scientific community with the aim to provide a reference standard for the evaluation of proteomic data acquisition and analysis platforms and to foster the development of novel experimental and computational approaches. Many applications of the library and data can be envisaged. Within the space and scope constraints of this article, we could only outline a few that

are of particular relevance for the field and all of which were facilitated by the design and size of the library. Collectively, the data obtained confirms many earlier findings but also holds some rather unanticipated surprises.

As for mass spectrometric data acquisition, the merits of ETD, CID and HCD for the analysis of peptides in general and phosphopeptides in particular are subject to much debate in the community [26-32]. The data presented in this study show that HCD (i.e. beam type CID with Orbitrap read out) is generally more successful than is ETD with Orbitrap readout, both for the identification of ordinary and phosphopeptides. Inspection of the raw MS data suggests that this is largely due to the fact that the ETD spectra contain fewer and less intense signals (owing to factors such as the generally lower duty cycle of the instrument in ETD mode, the low mass cut-off of ion traps, the general tendency to produce fewer signals in the low m/z region, inefficient fragmentation of precursor ions and ion losses during the ETD experiment). Notably, phosphopeptides turned out to be more readily identified than unmodified ones. We attribute this observation primarily to the addition of a highly mass-deficient phosphate group, which shifts the precursor and part of the fragment ion mass distributions away from that of unmodified peptides, creating a discriminating feature over unmodified peptides in a database search that considers variable modifications. Clearly, HCD and ETD are complementary fragmentation techniques. Still, we were surprised to find that ETD (with Orbitrap readout) did not perform appreciably better than HCD for peptide and phosphopeptide identification at higher charge states because prior work using ETD with ion trap read out had demonstrated superior performance on mis-cleaved tryptic peptides (typically having higher charge than fully cleaved tryptic peptides) [32]. For now, we are taking these observations at face value, but they are consistent with general past experience that results obtained by one type of fragmentation technique on one mass analyzer may not translate easily to a different mass analyzer. Nevertheless, our physical peptide library provides a means to address issues as the ones described above in a systematic manner for any mass spectrometric platform in the future.

It is well established, that different search engines generate somewhat different peptide identification results from the same mass spectrometric data [33]. Based on the example of the popular search engines Mascot and Andromeda, the MS data generated from the peptide library can be used to assess the merits of a search engine in an objective fashion and without the need for applying a decoy count or probabilistic error rate model as a proxy for false discovery rate computation. In general, the data shows that the Mascot score performs quite well, as the 5% global FDR value obtained by counting correct and incorrect matches for unmodified peptides is very close to the Mascot identity threshold, which corresponds to a 5% probability of a match to be a random event. For phosphopeptides however, the Mascot score model appears to be rather too conservative. Analogously, it was also possible to establish objective FDR criteria for peptides and phosphopeptides for Andromeda, which will aid in the further development of this search engine. In addition, the observed differences in FDR values generated by counting true and false positives (our approach) and by decoy counting (generally used by database search engines) suggests that the decoy approach often underestimates the true FDR and these results indeed raise questions as to the general validity of estimating FDR values from decoy counts. Hence, one notable outcome of the current study is the documentation that different and more refined computational models for peptide identification by database searching are needed depending on which fragmentation method is used, which peptide modification is considered and how FDRs should be estimated. The value of the data provided from this study is that these models can now be built using the MS data for HCD and ETD directly or using the physical library for those instrument platforms not covered here [34, 35].

In light of the rapidly increasing volume of phosphoproteomic data [36], we and others have devised computation approaches for the localization of phosphorylation sites within peptides [14, 15, 17, 18, 20, 37-39]. Again, the peptides and data generated in the current study can be used to assess and improve existing methods. The increase in available peptides has allowed us to refine the MD-score for phosphorylation site localization at several levels (fragmentation technique, type of phosphorylated amino acid, effect of distance between two possible phosphorylation sites, global and local FLR). Similarly, the comparison of the MD-score, phosphoRS and PTM-score showed that despite their similar overall accuracy, there is substantial complementarity, suggesting that all of the methods can be improved. Again, our data and peptides should facilitate such improvements or perhaps spur the development of entirely new ideas, which may be needed to address the considerable challenge of scoring multiply phosphorylated peptides having many potential acceptor sites.

Lastly, we demonstrated the utility of the synthetic library for analyzing peptide-retention behavior on reversed phase liquid chromatography. Our analyses confirmed previous empirical observations and the library can be similarly applied to analyze multidimensional separations. Although this is beyond the scope of the current study, we anticipate that the diversity and size of the library may stimulate further research, for example, into how retention time prediction models can be refined, and how such refinements may be used to improve peptide identification algorithms, targeted peptide quantification or modification site analysis [40-43], particularly with a view to distinguishing isomeric (phospho) peptides. Future work could also address topics such as (i) ion mobility measurements and their value for peptide identification and modification analysis [44, 45], particularly for isomers, (ii) understanding more fundamentally how the presence of a phosphate group influences fragmentation behavior and how that may be used for better computational tools (see Fig. 3) and (iii) the comparison of upstream peptide separation and enrichment methods [46] as well as alternative fragmentation techniques and instrument platforms [17, 31].

Despite all the described useful features, the library is not perfect. The libraries are mixtures of unpurified, synthesized peptides. The presence of the inevitable synthesis artifacts in the data may render the statistical analysis more conservative than necessary because tandem mass spectra from synthesis artifacts are more likely to contribute more to false positives than true positives. The permutation scheme purposely generated many isomeric peptides, which represents a challenge for applications requiring the analysis of single species. The large number of isomers in our library may not exist in nature to a similar extent and which, again, may make it harder for database search algorithms to identify the correct one. However, the data may also facilitate the development of computational tools to deal with mixed tandem mass spectra resulting from the co-elution and co-fragmentation of (isomeric) peptides. The library does not contain any positional phosphorylation isomers. We had addressed this topic in previous work [17] using individually synthesized peptides, but a broader collection of these would certainly be desirable to refine further localization scores and models of retention time. Another important future step would be the extension of the library to multiply phosphorylated peptides to learn if and how experimental and computational approaches have to be adapted to enable the analysis of these potentially important peptides in a more comprehensive fashion than currently possible.

In conclusion, we believe that the synthetic peptide library and derived mass spectrometric data described in this manuscript will serve as a valuable resource for the experimental and computational research community. We are confident that its use will generate many new ideas and multiple lines of follow-up investigation including the generation of further physical libraries tailored to particular questions and new computational tools that collectively improve proteomic technology.

Author contributions

Marx H., Schliep J.E. and Kuster B. designed the study. Lemeer S., Schliep J.E.., Matheron L. and Shabaz M. performed experiments. Marx H., Lemeer S, Matheron L. and Cox J. analyzed data. Marx H., Lemeer S., Heck A.J.R., Mann M. and Kuster B. wrote manuscript.

Abbreviations

ETD	electron transfer dissociation
FDR	false discovery rate
FLR	false localization rate
FP	false positive
FWHM	full width at half maximum
HCD	beam-type collision-induced dissociation
LC-MS/MS	liquid chromatograph tandem mass spectrometry
MD score	mascot delta score
mgf	mascot generic format
PSM	peptide spectrum match
TP	true positive

References

1. Chen, Y., Kwon, S.W., Kim, S.C. & Zhao, Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. Journal of proteome research 4, 998-1005 (2005).
2. Chen, Y., Zhang, J., Xing, G. & Zhao, Y. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. Journal of proteome research 8, 3141-3147 (2009).
3. Keller, A. et al. Experimental protein mixture for validating tandem mass spectral analysis. Omics : a journal of integrative biology 6, 207-212 (2002).
4. Rudnick, P.A., Wang, Y., Evans, E., Lee, C.S. & Balgley, B.M. Large scale analysis of MASCOT results using a Mass Accuracy-based THreshold (MATH) effectively improves data interpretation. Journal of proteome research 4, 1353-1360 (2005).
5. Klimek, J. et al. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. Journal of proteome research 7, 96-103 (2008).
6. Mallick, P. et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nature biotechnology 25, 125-131 (2007).
7. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature methods 4, 787-797 (2007).
8. Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. Nature biotechnology 28, 695-709 (2010).
9. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature methods 4, 207-214 (2007).
10. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical chemistry 74, 5383-5392 (2002).
11. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. Analytical chemistry 75, 4646-4658 (2003).
12. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Molecular & cellular proteomics : MCP 10, M111 007690 (2011).
13. Bohrer, B.C. et al. Combinatorial libraries of synthetic peptides as a model for shotgun proteomics. Analytical chemistry 82, 6559-6568 (2010).
14. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nature biotechnology 24, 1285-1292 (2006).
15. Bailey, C.M. et al. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. Journal of proteome research 8, 1965-1971 (2009).
16. Lemeer, S. et al. Phosphorylation site localization in peptides by MALDI MS/MS and the Mascot Delta Score. Analytical and bioanalytical chemistry 402, 249-260 (2012).
17. Savitski, M.M. et al. Confident phosphorylation site localization using the Mascot Delta Score. Molecular & cellular proteomics : MCP 10, M110 003830 (2011).

18. Taus, T. et al. Universal and confident phosphorylation site localization using phosphoRS. Journal of proteome research 10, 5354-5362 (2011).
19. Daub, H. et al. Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. Molecular cell 31, 438-448 (2008).
20. Olsen, J.V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635-648 (2006).
21. Olsen, J.V. et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Science signaling 3, ra3 (2010).
22. Oppermann, F.S. et al. Large-scale proteomics analysis of the human kinome. Molecular & cellular proteomics : MCP 8, 1751-1764 (2009).
23. Rikova, K. et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 131, 1190-1203 (2007).
24. Steen, H., Jebanathirajah, J.A., Rush, J., Morrice, N. & Kirschner, M.W. Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. Molecular & cellular proteomics : MCP 5, 172-181 (2006).
25. Krokhin, O.V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. Analytical chemistry 78, 7785-7795 (2006).
26. Jedrychowski, M.P. et al. Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. Molecular & cellular proteomics : MCP 10, M111 009910 (2011).
27. Nagaraj, N., D'Souza, R.C., Cox, J., Olsen, J.V. & Mann, M. Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation. Journal of proteome research 9, 6786-6794 (2010).
28. Swaney, D.L., McAlister, G.C. & Coon, J.J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nature methods 5, 959-964 (2008).
29. Swaney, D.L., Wenger, C.D., Thomson, J.A. & Coon, J.J. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 106, 995-1000 (2009).
30. Boersema, P.J., Mohammed, S. & Heck, A.J. Phosphopeptide fragmentation and analysis by mass spectrometry. Journal of mass spectrometry : JMS 44, 861-878 (2009).
31. Frese, C.K. et al. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. Journal of proteome research 10, 2377-2388 (2011).
32. Zhou, H. et al. Enhancing the identification of phosphopeptides from putative basophilic kinase substrates using Ti (IV) based IMAC enrichment. Molecular & cellular proteomics : MCP 10, M110 006452 (2011).
33. Kapp, E.A. et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5, 3475-3490 (2005).
34. Frank, A.M. et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nature methods 8, 587-591 (2011).
35. Huang, Y. et al. A data-mining scheme for identifying peptide structural motifs responsible for different MS/MS fragmentation intensity patterns. Journal of proteome research 7, 70-79 (2008).

36. Lemeer, S. & Heck, A.J. The phosphoproteomics data explosion. Current opinion in chemical biology 13, 414-420 (2009).
37. Baker, P.R., Trinidad, J.C. & Chalkley, R.J. Modification site localization scoring integrated into a search engine. Molecular & cellular proteomics : MCP 10, M111 008078 (2011).
38. Chalkley, R.J. & Clauser, K.R. Modification site localization scoring: strategies and performance. Molecular & cellular proteomics : MCP 11, 3-14 (2012).
39. Kelstrup, C.D., Hekmat, O., Francavilla, C. & Olsen, J.V. Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. Journal of proteome research 10, 2937-2948 (2011).
40. Krokhin, O.V. & Spicer, V. Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. Analytical chemistry 81, 9522-9530 (2009).
41. Conrads, T.P., Anderson, G.A., Veenstra, T.D., Pasa-Tolic, L. & Smith, R.D. Utility of accurate mass tags for proteome-wide protein identification. Analytical chemistry 72, 3349-3354 (2000).
42. Moruz, L. et al. Chromatographic retention time prediction for posttranslationally modified peptides. Proteomics 12, 1151-1159 (2012).
43. Moruz, L., Tomazela, D. & Kall, L. Training, selection, and robust calibration of retention time models for targeted proteomics. Journal of proteome research 9, 5209-5216 (2010).
44. Geromanos, S.J. et al. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. Proteomics 9, 1683-1695 (2009).
45. Hoaglund-Hyzer, C.S., Li, J. & Clemmer, D.E. Mobility labeling for parallel CID of ion mixtures. Analytical chemistry 72, 2737-2740 (2000).
46. Thingholm, T.E., Jensen, O.N. & Larsen, M.R. Analytical strategies for phosphoproteomics. Proteomics 9, 1451-1468 (2009).
47. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology 157, 105-132 (1982).

Chapter 4

Annotation of the pig genome by mass spectrometry-based proteomics

Abstract

We present the first draft of the porcine proteome to date. The recent sequencing of the Sus Scrofa (pig) genome and its importance as a potential mammalian model organism and livestock species necessitate the structural and functional genome annotation. Mass spectrometry based proteogenomics facilitates the identification of novel genes as well as the validation and refinement of gene and transcript models.
We searched nine juvenile organs and six embryonic stages against one of the latest genome assemblies, transcript entities (EST, cDNA, mRNA) and protein sequences. We also proposed a tailored strategy to combine inference-prone data from multiple proteogenomics experiments into a consensus set of peptide and protein identifications. Using this approach we identified 7,108 proteins originating from 5,968 known, 690 novel and 176 refined gene models.

Introduction

The morphological and physiological resemblance of *Sus Scrofa* to *Homo Sapiens* facilitates the study of human diseases in a potential mammalian model organism [1]. The recent genome sequencing [2] and prior physical mapping [3] revealed also a close phylogenetic relationship on a molecular level. Furthermore is the domestic pig an important livestock species to produce various commodities. It is therefore critical to improve and extend the structural and functional genome annotation to advance medical and biological research [4].
A popular example for a genome annotation pipeline is Ensembl [5] comprising ab initio gene predictions, the projection of orthologous information and integration of available gene, transcript and protein data. Mass spectrometry (MS) based proteogenomics is able to provide unprecedented protein level evidence to supplement and improve the annotation [6, 7]. The merits of proteogenomics were highlighted for various model organisms including plants such as *Arabidopsis thaliana* [8, 9], *Medicago truncatula* [10], *Zea mays* [11] and eukaryotes namely *Saccharomyces cerevisiae* [12], *Mus musculus* [13, 14], *Rattus norvegicus* [15] as well as *Homo Sapiens* [16, 17] uncovering novel gene models, alternative splice products and refining, validating existing gene and transcript models.
Even though proof of concept for proteogenomics has been shown over a variety of taxa, analysis of higher eukaryotes continues to be a computational challenge [18], due to the complexity of alternative splicing, the large amount and varying sources of sequences and the respective data processing and validation. The technological advance in high resolution mass spectrometry envisages a comprehensive and deep proteome coverage generating more and more data not interpretable with common protein sequence databases necessitating the addition of other sequence sources. Therefore a common approach in proteogenomics is to extend the protein sequence search space by including genome or transcriptome data. In doing so, the genome is translated in all six reading frames [19] and valid exon information is used to construct exon graphs [20] to cover all possible alternative splicing events. Alternative splice variants are also investigated by transcript data, such as ESTs [21, 22], cDNA and lately RNA-Seq [23, 24]. The ready access to RNA-Seq data drove recent proteogenomic studies [25, 26, 27] to cope with the extensive search space of exon graphs by pruning the

exons to the ones covered in RNA-Seq data. Even though beneficial, RNA-Seq solely covers a set of transcripts in a cell state, type or tissue and the complementarity of each platform (RNA-Seq, MS-based proteomics) is also able to provide unique insights [11, 28, 29].
The ultimate goal of a proteogenomic experiment is to predict novel and refined gene models, e.g. with Augustus [30] supplementing the gene predictor with extrinsic information such as transcript information and interpreted data from MS-based proteomics [31].
In this study, we provide the first draft of the porcine proteome by MS-based proteomics. Our biological samples comprise most of the juvenile organs and early embryonic stages after gestation. The resulting protein level evidence facilitates the annotation of the corresponding genome.
As introduced, we pursue the idea of extending the search space and hence search the peptide fragment spectra against the latest genome assembly (Ensembl, 10.2.70), transcript entities (EST, cDNA, mRNA) and protein sequences [32] from the Ensembl, NCBI (RefSeq, GenBank) and UniProt consortia. To this end, the genome was translated to a six-frame translation and exons to a compact representation of the exon graph as a peptide centric version [33]. The peptide centric exon graph (PEPcEX) covers the theoretical splice search space based on Ensembl predicted and valid exons. To minimize the accumulation of false positive peptide spectrum matches (PSM) in multi database searches, we introduced a naive PSM grouping approach and also investigate the influence of statistical measures [34] to validate peptide and protein identifications. The strategy could be used as a guideline to standard data processing in a proteogenomic context, stating not to take identifications at face value.

Material and methods

Animal welfare and keeping

The juvenile domestic German landrace pigs (gilts), age 5.5 to 6 months, were kept in compliance with the animal welfare for pigs of the EU (directive 2008/120/EC) at a livestock breeding in Thalhausen at one of the agricultural experimental stations of the Technische Universitaet Muenchen (Center of Life and Food Sciences Weihenstephan, Germany).

Sample extraction

Embryos

To initiate synchronous estrous, the gilts were treated as standard ovulation synchronization schedule with Altrenogest® (progesteron) for 18 days followed by 750 IE Intergonan® (PMSG, Gonadotropin) 24h after the last Altrenogest and 750 IE Ovogest® (hCG, Chorionic gonadotropin) 80h after Intergonan application. Twenty-four hours later they are artificially inseminated with the semen of the same boar, which is repeated again 12 h later. The day after the last insemination was determined as day one. The gilts were slaughtered at days 18, 22, 25, 28, 32 and 39, respectively, at the local slaughterhouse. The uterus was an animal by-product of the slaughtering and is collected immediately after slaughtering. The implanted embryos were excised from the endometrium, washed with PBS and stored on dry ice.

Gilt organs

A female gilt was sedated and bled out (euthanized). The organs were excised shortly after, washed with PBS and stored on dry ice for the transportation. In total nine organs were extracted, namely diaphragm, spleen, biliary, kidney, liver, lung, brain, pancreas and heart.

Sample preparation

The juvenile organs and embryos were washed multiple times with cold PBS. 10g sample (random location) were lysed using Tris-HCl buffer containing 4% SDS and a Miccra D-9 homogenizer (ART Labortechnik, Germany). The lysate was ultracentrifuged for 1 h at 20 °C and 52000× g. The protein extract ~ 1.3 - 15.5 µg / µl (Bradford Protein Assay) was reduced with 10 mM dithiothreitol and alkylated with 55 mM iodoacetamide. Proteins were separated by 1D SDS gel electrophoresis and each lane was cut into 12 regions. The regions were digested with trypsin [35].

LC-MS/MS analysis

Nanoflow LC-MS/MS was performed by coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent, Dublin, CA) to a Velos-Oribtrap Elite (Thermo Scientific, Bremen, Germany). Peptides were delivered to a trap column (100 µm i.d. × 2 cm, packed with 5µm C18 resin, Reprosil PUR AQ, Dr. Maisch, Ammerbuch, Germany) at a flow rate of 5 µL/minute in 100% buffer A (0.1% FA in HPLC grade water). After 10 minutes of loading and washing, peptides were transferred to an analytical column (75µmx40 cm C18 column Reprosil PUR AQ, 3µm, Dr. Maisch, Ammerbuch, Germany) and separated using a 55 minute gradient from 2% to 35% of buffer B (0.1% FA in acetonitrile) at 300 nL/minute flow rate. The Velos-Orbitrap Elite was operated in data dependent mode, automatically switching between MS and MS2. Full scan MS spectra were acquired in the Orbitrap at 30,000 resolution. Internal calibration was performed using the ion signal (Si (CH3)2O) 6 H + at m/z 445.120025 present in ambient laboratory air. Tandem mass spectra were generated for up to 15 peptide precursors in the linear ion trap for fragment by using Higher energy collisional dissociation (HCD).

Search databases

The Ensembl genome (DNA) assembly (Sscrofa10.2.70 build) and the respective gene, exon, transcript models and protein sequences (Ensembl - all) were the reference to compare or rather classify against. All other databases cover EST, cDNA, mRNA and Protein (PEP) sequences from the Ensembl, NCBI (RefSeq, GenBank) and UniProt consortium (Table 1). The [cDNA] Ensembl - all sequences comprise EST, cDNA and RNA-Seq data (see supplementary information, Groenen et al., Nature, 2012) [2].

Name	Molecule Type	Size (Entries)	Version
Ensembl - all	PEP	25,883	10.2.70
Ensembl - ab initio	PEP	52,372	10.2.70
RefSeq	PEP	59,991	01.03.2013
UniProtKB	PEP	33,205	28.02.2013
Ensembl - all	cDNA	82,635	10.2.70
Ensembl - ab initio	cDNA	157,116	10.2.70
RefSeq	mRNA	195,291	28.03.2013
GenBank	EST	10,034,796	28.03.2013
PEPcEX	DNA	2,936,004	10.2.70

Ensembl	DNA	85,738,772	10.2.70

Table 1. Databases meta information. Database names, size, molecule type (DNA, EST, cDNA, mRNA, PEP) and version.

Search database construction - Frame translation

The pre-masked genome sequence, i.e. low complexity regions were masked with RepeatMasker [36], was translated in all six reading frames (+1, +2, +3, -1, -2, -3), where each FASTA entry in the resulting database is an open reading frame (ORF). Transcript sequences (cDNA, mRNA and ESTs) were translated in all three forward frames (+1, +2, +3).

Search database construction - Peptide centric Exon graph (PEPcEX)

In brief, each node of the exon graph represents an exon from Ensembl (validated and predicted) and the edge the splice site. In higher eukaryotes exist for each locus theoretical 2^n - 1 linear exon (n) combinations [37]. Exon phases restricted the number of valid combinations. To reduce the size of the resulting database, we perform an on-the-fly in silico digest to construct a peptide centric database (PEPcEX). The most notable digest parameter is the mass interval required to omit undetectable peptides. The monoisotopic precursor mass interval was derived from the peak lists based on [min. -5 ppm, max. + 5 ppm] mass resulting in [299.0049, 8373.878] [Da].

Search database construction - Consensus

To derive a consensus database of known and novel transcripts (proteins), we used a peptide centric clustering algorithm [32]. The input databases comprised all identified transcript models (PEP, cDNA, mRNA) and the Augustus gene predictions. Notable parameters are T = 0, Length = true, Length_Min = 7, Length_Max = 52 and Miss_Cleavages = 2.

False discovery rate (FDR) models

In a recent study [38], we were able to derive (objective) false discovery rate (FDR) models as a function of a search engine score (Mascot, MaxQuant) based on a synthetic (phospho-) peptide library. To optimize the local confined models we applied four new models (m_1 (Score) = A exp (B Score), m_2 (Score) = A exp (B Score) + C, m_3 (Score) = A exp (B Score) + exp (C Score), m_4 (Score) = A exp (B Score) + D exp (C Score)). The respective coefficients are available for unmodified (Table 2) and phosphorylated (Table 3) peptides. We rank the models based on ANOVA p-value and the Residual Sum of Squares (RSS). If applicable we illustrate beside the i) best fit also the ii) original or a more iii) asymptotic model. To smooth the distribution of data points for the local FDR models we omit FDRs following the condition $FDR_{n-1} < FDR_n$ and fit a model according to Jones et al. [39] (Fig. 1).

Fragmentation	Search engine	FDR	Model	Score	A	B	C	D
HCD	Mascot	Local	m_3	[0.0, ∞]	-0.10312	-0.94020	-0.05769	-

			m_4	[21.60, 100.38]	0.14802	-0.01921	-0.16158	4.54646
		Global	m_4	[0.0, ∞]	0.25637	-0.13345	-0.01574	0.06859
ETD		Local	m_3	[0.0, ∞]	-0.08910	-1.36322	-0.04568	-
			m_4	[31.88, 124.65]	1.95789	-0.07292	-0.14787	-1.04657
		Global	m_3	[0.0, ∞]	-0.31956	-0.04985	-0.04985	-
			m_4	[0.0, ∞]	0.085207	-0.001163	-0.082414	0.687492
HCD	Andromeda	Local	m_1	[0.0, ∞]	1.20968	-0.02002	-	-
			m_4	[75.84, 155.76]	0.085097	-0.004016	-0.066623	26.825405
		Global	m_3	[0.0, ∞]	-0.42677	-0.02988	-0.02988	-
			m_4	[0.0, ∞]	0.048624	-0.002116	-0.038946	0.542661
ETD		Local	m_1	[0.0, ∞]	0.99025	-0.03332	-	-
		Global	m_2	[0.0, ∞]	0.70867	-0.03743	0.08729	-
			m_4	[0.0, ∞]	0.765448	-0.032880	0.008588	0.026013

Table 2. False Discovery Rate (FDR)-Models and coefficients for unmodified peptides. Non-linear least square regression using four base models: m_1 (Score) = A exp $^{(B \cdot Score)}$, m_2 (Score) = A exp $^{(B\ Score)}$ + C,m_3 (Score) = A exp $^{(B\ Score)}$ + exp $^{(C \cdot Score)}$,m_4 (Score) = A exp $^{(B\ Score)}$ + D exp $^{(C\ Score)}$ and deriving the respective coefficients. The models are functions of the score and applicable for the fragmentation methods HCD and ETD, search engines Mascot and Andromeda, FDR methods Local and Global.

Fragmentation	Search engine	FDR	Model	Score	A	B	C	D
HCD	Mascot	Local	m_3	[0.0, ∞]	-0.17458	-5.67732	-0.07704	-
			m_4	[16.12, 100.97]	0.09376	-0.0339	-0.15771	2.13304
		Global	m_4	[0.0, ∞]	0.38796	-0.19020	-0.02567	0.03268
			m_4	[3.99, 100.10]	0.39446	-0.19036	-0.0249	0.03166
ETD		Local	m_1	[0.0, ∞]	1.11225	-0.07547	-	-
			m_4	[18.57, 105.16]	0.01828	0.01105	-0.08226	0.81877
		Global	m_3	[0.0, ∞]	-0.2228	-0.1035	-0.1035	-
			m_4	[0.0, ∞]	0.78917	-0.12750	0.00534	0.03336
HCD	Andromeda	Local	m_1	[0.0, ∞]	1.42325	-0.02339	-	-
			m_4	[69.55, 180.49]	0.003722	0.004204	-0.060589	13.43191
		Global	m_4	[0.0, ∞]	-0.70316	-0.10613	-0.05393	1.37562
			m_4	[29.61, 149.41]	0.01111	-0.00384	-0.05423	1.16356
ETD		Local	m_1	[0.0, ∞]	1.12449	-0.03691	-	-
		Global	m_4	[0.0, ∞]	1.82587	-0.06963	-0.14468	-1.04418

Table 3. False Discovery Rate (FDR)-Models and coefficients for phosphorylated peptides. Non-linear least square regression using four base models: m_1 (Score) = A exp $^{(B \cdot Score)}$, m_2 (Score) = A exp $^{(B\ Score)}$ + C,m_3 (Score) = A exp $^{(B\ Score)}$ + exp $^{(C \cdot Score)}$,m_4 (Score) = A exp $^{(B\ Score)}$ + D exp $^{(C\ Score)}$ and deriving the respective coefficients. The models are functions of the score and applicable for the fragmentation methods HCD and ETD, search engines Mascot and Andromeda, FDR methods Local and Global.

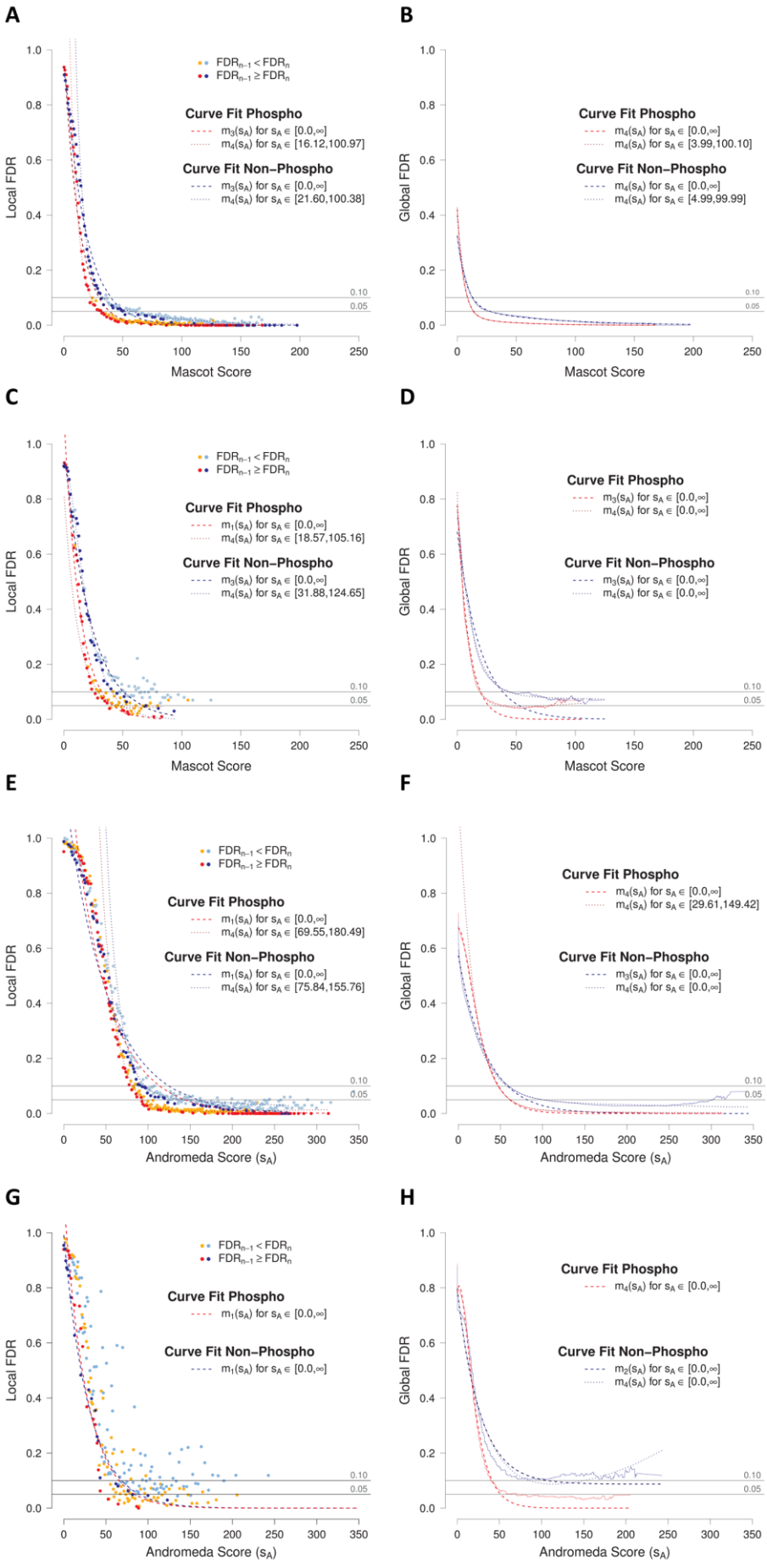
A
B
C
D
E
F
G
H
Local FDR
Global FDR
Mascot Score
Andromeda Score (sA)
FDRn−1 < FDRn
FDRn−1 ≥ FDRn
Curve Fit Phospho
Curve Fit Non−Phospho
m3(sA) for sA ∈ [0.0,∞]
m4(sA) for sA ∈ [16.12,100.97]
m4(sA) for sA ∈ [21.60,100.38]
m4(sA) for sA ∈ [3.99,100.10]
m4(sA) for sA ∈ [4.99,99.99]
m1(sA) for sA ∈ [0.0,∞]
m4(sA) for sA ∈ [18.57,105.16]
m4(sA) for sA ∈ [31.88,124.65]
m4(sA) for sA ∈ [0.0,∞]
m4(sA) for sA ∈ [69.55,180.49]
m4(sA) for sA ∈ [75.84,155.76]
m4(sA) for sA ∈ [29.61,149.42]
m2(sA) for sA ∈ [0.0,∞]
0.10
0.05

Figure 1. Novel ETD and HCD False Discovery Rate (FDR) models based on a synthetic library. Non-linear least square regression using four base models: m_1 (Score) = A exp $^{(B\ Score)}$, m_2 (Score) = A exp $^{(B\ Score)}$ + exp $^{(Score)}$,m_3 (Score) = A exp $^{(B\ Score)}$ + exp $^{(C\ Score)}$,m_4 (Score) = A exp $^{(B\ Score)}$ + D exp $^{(C\ Score)}$ and deriving the respective coefficients.(a) Local HCD, Mascot. (b) Global HCD, Mascot. (c) Local ETD, Mascot. (d) Global ETD, Mascot. (e) Local HCD, Andromeda. (f) Global HCD, Andromeda. (g) Local ETD, Andromeda. (h) Global ETD, Andromeda.

MS Data processing

Peak picking and database searching

MaxQuant version 1.3.0.3 was used to generate peak lists (apl files) from the raw MS files for subsequent database searching (single search against each database). Notable parameters for the search were: Oxidation (M) and Acetyl (Protein N-Term) as variable modifications, a mass tolerance window of 5 ppm for MS1 and 20 ppm for MS2, trypsin as enzyme, up to 2 max. missed cleavages and enabled reverse decoy database option. The resulting peptide spectrum matches (PSMs) were used from evidence.txt for each search.

Peptide spectrum match grouping

We introduce a naive spectrum clustering approach to reduce false positive peptide spectrum match (PSM) assignments over multiple database searches. First, we extracted the spectrum meta information, i.e. raw file name and respective scan number. Second, we assigned the PSMs of each individual database search to the distinct spectrum meta information. The trivial case was a 1 : 1 relation of a spectrum to a PSM (identical sequence and score over all searches) and hence classified as unique. In case of a 1 : n relation, were the PSMs classified into representatives, member and chimeric (second peptides). A representative PSM had the highest score in comparison to the other assigned PSMs (member) except for chimeric PSMs. In case of identical scores for multiple representative PSMs were all counted as valid. In general was the highest scoring PSM for each spectrum subject of further analysis as well as chimeric PSMs.

Score and FDR Filtering

The filtering of the PSMs was at a 0.01 peptide FDR and protein FDR (MaxQuant). Additional Andromeda Score (S_A) filtering ($S_A \geq 58.44$, FDR (Library) ≤ 0.10) removed low (quality) scoring spectra in the qualitative analysis based on the Non-Phospho global FDR HCD Andromeda model (see FDR models).

To reduce the chance of false positives in an increasing search space additional criteria beyond standard proteomics approaches are required in proteogenomics. To this end were all peptide classifications (not matching in reference annotations) subject to additional local FDR (MaxQuant posterior error probabilities) filtering < 0.01 in contrast to the prior global FDR filters [40, 41].

Gene prediction

To train the gene predictor Augustus, we built a valid GenBank file comprising 11,636 Ensembl (Sscrofa10.2.70 build) genes (forward strand) and randomly select a set of 1,100 entries (see Augustus Online Tutorial). The set is split in a training set (1,000 entries) and a test set (100 entries).

The ab initio training resulted in a gene level sensitivity of 0.16 and specificity of 0.109. Subsequent parameter optimization did not improve the outcome.

Extrinsic sources, i.e. EST, cDNA, mRNA, PEP were subjected to Augustus as hints. To prepare the hints, pre-processing of the information was necessary. To derive valid exons of alternative splice variants on the protein evidence of the various databases we used 7,841 exon branches (PEPcEX) and 12,990 protein groups (no inference) in conjunction with exonerate (--model protein2genome --showtargetgff T).

We used the BLAST-like Alignment tool (BLAT) (-noHead -minIdentity=92) on the masked (RepeatMasker) genome to map the EST, cDNA and mRNA sequences to the genome. And post-processed these with pslCDnaFilter (-minId=0.9 –localNearBest=0.005 –ignoreNs -bestOverlap) to find the best match. We ran exonerate (--model protein2genome --showtargetgff T) on the BLAT output and merge all the exonerate results in a single file (extrinsic information).

Augustus was run in parallel for each chromosome (--protein=on,--introns=on,--start=on, --stop=on,--cds=on,--codingseq=on,--alternatives-from-evidence=true,--alternatives-from-sampling=false,--sample=100,--extrinsicCfgFile=extrinsic.MPE.cfg). The gff output was parsed to retrieve exon, transcript, protein coordinates and sequences.

Peptide coordinates

Mapping

To derive peptide coordinates relative to the genome, we searched against the i) six-frame translation of the genome, ii) the peptide centric exon graph and the remainders iii) with BLAT (-out=pslx, -t=dnax, -q=prot) and iv) with BLAST (-word_size 2 -matrix PAM30 -seg "no" -evalue 20000 -comp_based_stats 0). Criteria for the best BLAST and BLAT matches were to allow a single amino acid polymorphism (SAP) in the alignment and in case of splice events, i.e. spanning the N-Terminal and C-Terminal subsequences over separate alignments, no SAP. Splice peptides, reaching over alignments would generate a tree structure, therefore we limit the depth = 2 and the occurrence of a consecutive alignment in a genomic interval of 9,369 bp (median gene size, Fig. 2A).

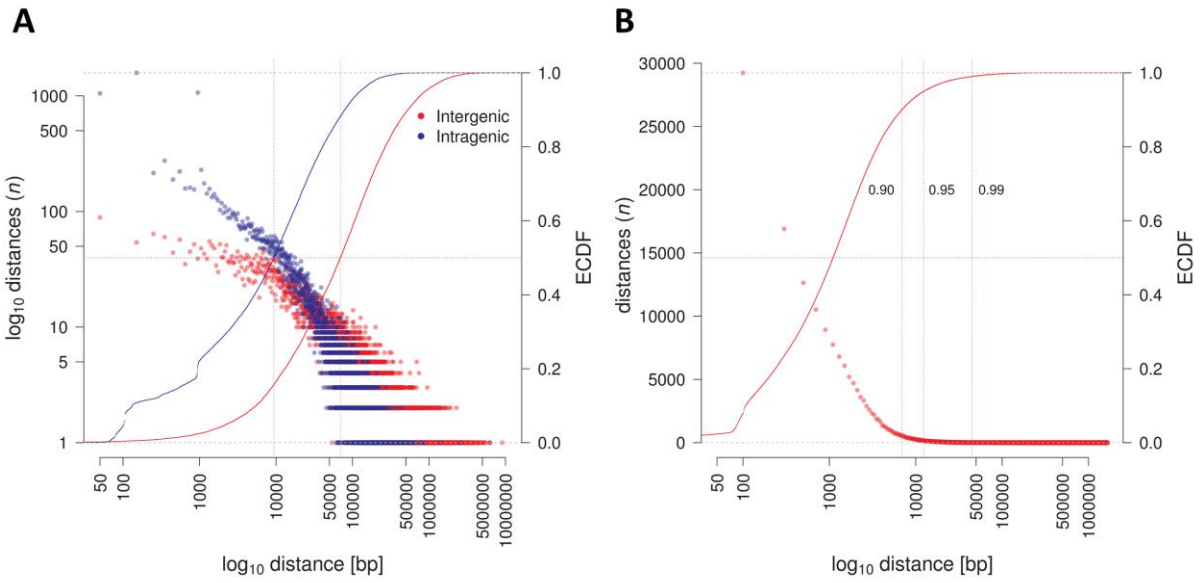

Figure 2. Inter- and intragenic length. (A) Number of inter (between genes)- and intragenic (in genes) distance as a function of the distance in [bp]. (B) Number of introns [exon end, exon start] as a function of the distance in [bp]

Single linkage clustering

To estimate the correctness of known gene model boundaries or identification of novel gene models we applied single linkage clustering.

In brief, each peptide with a genomic coordinate was linked to its nearest neighbours, considering a threshold T, following

$d(x,y) \leq T,$

where x, y are peptide coordinates and d the distance. The threshold was set to 12,373 bp corresponding to the 0.95 quantile intron length (Fig. 2B).

Annotations

Ensembl gene

The reference gene annotation includes the information of the Ensembl gene biotype and status. The top-tier biotype categories are Non-Coding (misc_RNA, snRNA, antisense, miRNA, processed_transcript, snoRNA, non_coding, lincRNA), Pseudogene (IG_V_pseudogene, pseudogene) and IG gene (IG_C_gene, IG_J_gene, IG_V_gene). The gene status is Known, Known by Projection, Novel and Merged (www.gencodegenes.org/gencode_biotypes.html).

Peptide classification

The top-tier peptide classification were intra- and intergenic events. We differentiated classifications over the genome, transcriptome and proteome (Fig. 3). Fusion (Genome, Proteome) classifications were peptides mapping between two genes, distances were defined over single linkage peptide cluster or the transcript boundaries (mapped protein). Exon skipping classifications were peptides matching to exon combinations in PEPcEX or BLAT, BLAST not present in [PEP] Ensembl - all. Exon boundary classifications were peptides reaching over the exon C-terminus indicating the splice site to be in a false position. UTR exons had no Ensembl phase (-1/-1) information and are by definition part of the UTR region. Phase shift classifications were peptides matching to another frame and consequently not matching to the assigned phases of the exon.

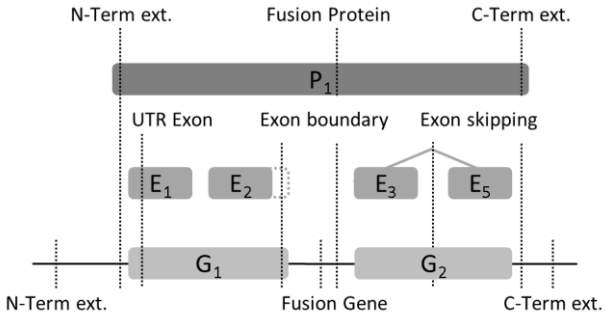

Figure 3. Peptide classification. Schematic of intergenic (N-Term ext., C-Term ext., Fusion Gene and Protein) and intragenic events (UTR-Exon, Exon boundary, Exon skipping) over the genome, transcriptome and proteome.

Inference problem

Genome and protein inference

In general, inference in MS-based proteomics distinguishes unique and shared peptides. In protein inference a peptide is shared or unique in the proteome [42], whereas in genome inference a peptide is distinct (unique) to a genomic location or to multiple (shared).

All gene, transcript models and proteins were subject to genome inference, requiring at least a single genomic unique peptide [43]. In addition proteins required at least a single unique peptide on the proteome level.

Model grouping

The gene model, transcript model and protein grouping is a naive approach to remove subsets and same-sets of peptide identifications, i.e. assigning to each model or protein all peptide identifications. In case of multiple sequences sharing all peptide identifications the longest sequence was selected as representative for the group.

Results

Workflow

In this study, we profiled the proteome of 15 porcine biological samples, nine juvenile organs and six embryonic stages using a conventional GeLC-MS/MS approach [44] in combination with a high resolution mass spectrometer (Fig. 4).

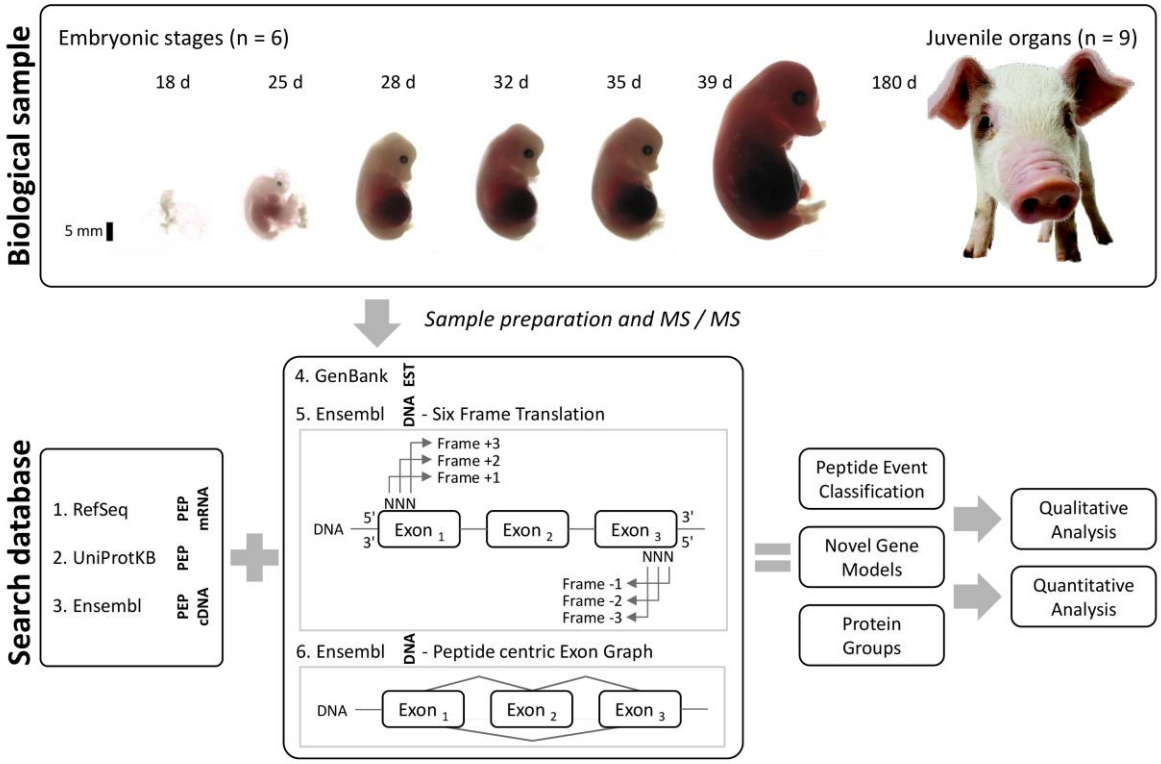

Figure 4. Workflow figure. The biological samples as a morphological timeline of Sus Scrofa embryonic stages 18d, 25d, 28d, 32d, 35d, 39d and a 180d female juvenile (adapted from Nature, Volume 491, 2012). In total nine juvenile organs were analyzed, namely Diaphragm, Spleen, Biliary, Kidney, Liver, Lung, Brain, Pancreas and Heart (top panel). The samples were lysed and subjected to LC-MS/MS analysis. Subsequent the data was searched against various databases covering the Genome, Transcriptome and Proteome including a six-frame translation and a peptide centric exon graph. The qualitative analysis comprises peptide event classification (intergenic, intragenic), novel gene model identification and refinement (bottom panel).

The subsequent database search of the acquired peptide fragment spectra covers a comprehensive search space including ten databases (Table 1) comprising EST, cDNA, mRNA and protein sequences (PEP). Additionally we constructed from the genome (DNA) a six frame translation database enabling the identification of exact peptide matches to the Ensembl reference genome and an exon graph resulting in a peptide centric database (PEPcEX) to cover splice matches and alternative starts. The exon graph construction runs with polynomial time complexity (Fig. 5A). Therefore we restricted the input to a maximum of 24 exons per transcript covering 0.95 of the transcripts in Ensembl all and ab initio (Fig. 5B) resulting in 2,936,004 peptide sequences.

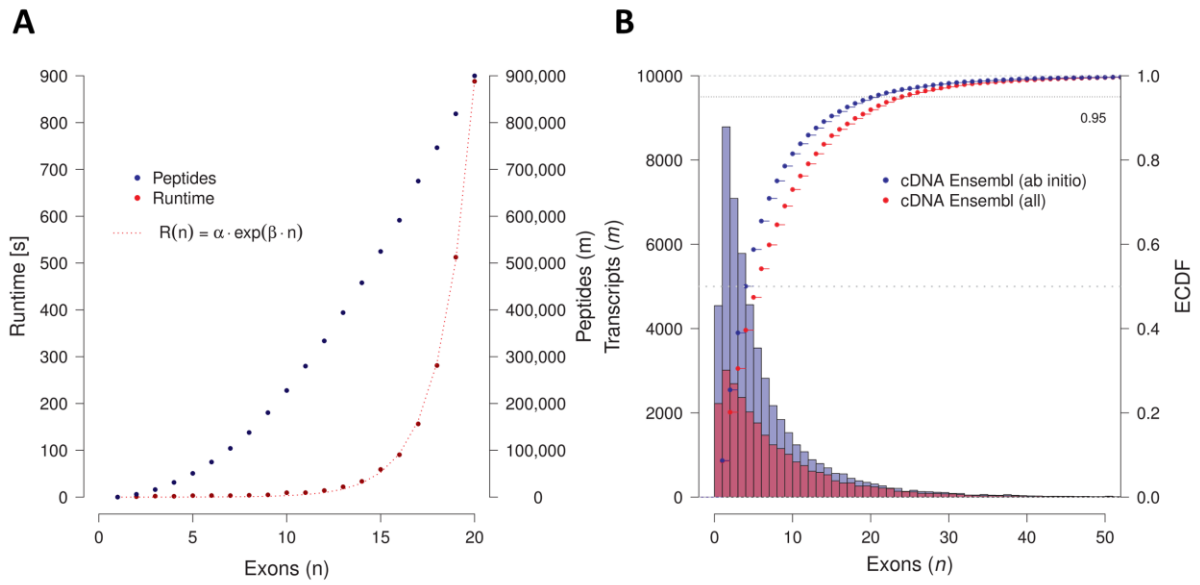

Figure 5. Exon graph construction. (A) Runtime analysis of the exon graph construction as an exponential function of the number of exons (n) per gene locus. (B) Number of transcripts (m) to exons(n). To cover 95% of all possible exon combinations all gene loci with up to 24 exons are considered.

The peptide fragment spectra were searched against all databases with MaxQuant resulting in 810,225 peptide spectrum matches (PSM) and 93,494 peptide identifications. To derive valid PSMs respectively peptide identifications in a proteogenomic context, we introduced a PSM inference grouping, an objective criteria to control the quality of the spectra based on the search engine score and the notion of genome inference (Fig. 6). The subsequent two-tier analysis, includes i) qualitative aspects, i.e peptide event classification, gene and transcript model identification and validation, whereas the ii) quantitative analysis will include expression profile analysis over embryonic stages and gene ontology (GO) enrichment.

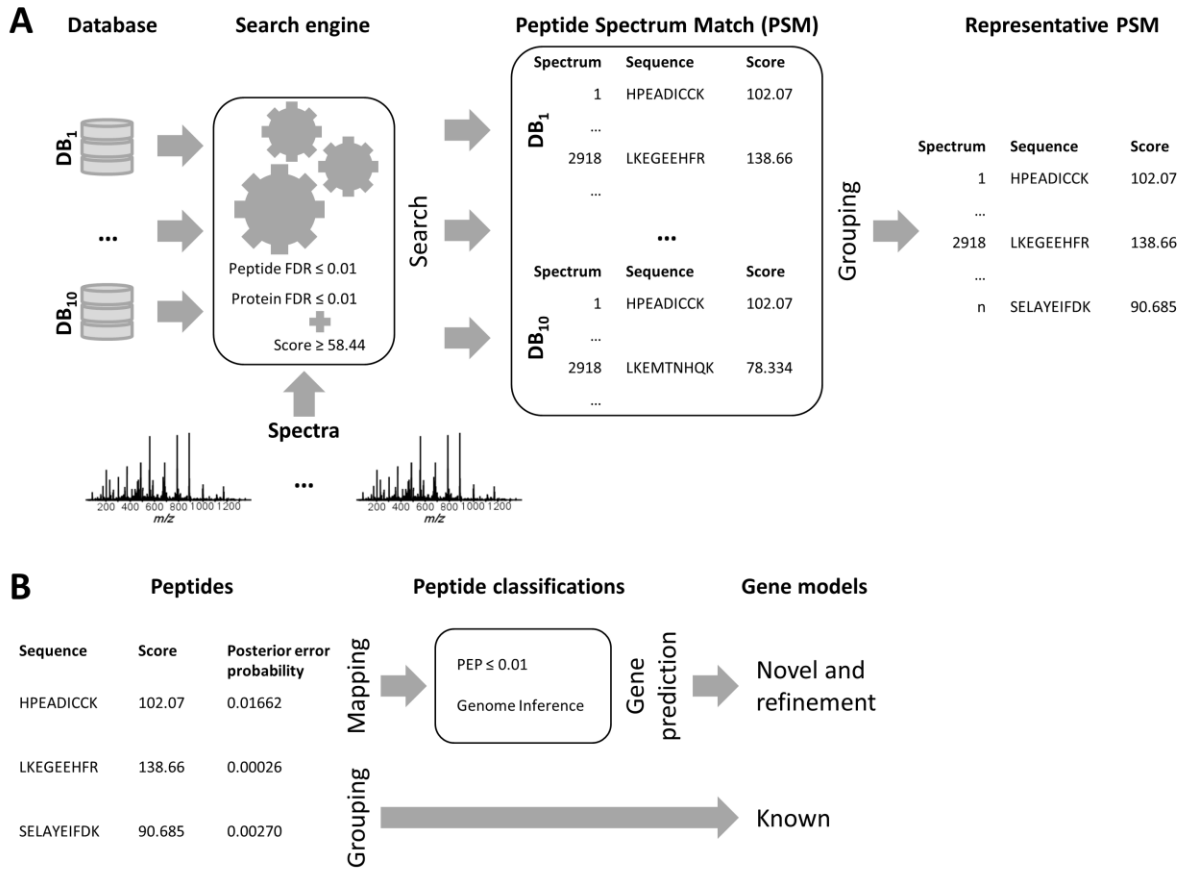

Figure 6. Data processing. (A) Ten database searches filtered with a 0.01 peptide and protein FDR. Additional all PSMs were subject to score filtering (score ≥ 58.44). Subsequent PSMs are grouped to derive a representative match for each spectrum. (B) The resulting peptide identifications are mapped to the genome. Peptides not matching to a known gene model were filtered with a posterior error probability of 0.01 and required a distinct genome coordinate (genome inference) before classification into inter- and intragenic events. Peptides matching to known genome models were grouped together, whereas classified peptides are input to a gene predictor to derive novel and refined gene models.

PSM grouping and peptide evidence validation

Searching the peptide fragment spectra separately against each database requires post processing steps to derive reliable identifications, i.e. conclusive PSMs and respective quality validation.

To omit ambiguities in the PSMs over search spaces, we introduced a naive PSM grouping approach to reduce the number of false positive assignments (varying peptide sequence matches to a spectrum) (Fig. 6). The majority (763,677) are singletons or rather unique (Fig 7A) assignments, but 46,548 of 810,225 PSMs were degenerate. We define three classes for degenerate spectra, namely chimeric, representative and member. The 12,540 chimeric classifications are not a result of the data processing and therefore valid. The remaining 34,008 (4.20 %) are subject to PSM grouping. We omit all members from further analysis due to higher scoring PSMs. In the case of [PEP] Ensembl - all , we see the highest number of PSMs sharing information (member) with other databases. Representative PSMs are overrepresented due to indistinguishable PSMs in case of identical scores (Fig 7A).

Surprisingly in 221 cases even though a 1 : 1 relation of peptide sequence to spectrum existed, were the Andromeda scores (S_A) differing by $\Delta = 7.19$ (median; Fig. 7 B). We were proceeding in this cases with the higher scoring PSM. In total we identify 800,195 conclusive and valid PSMs.

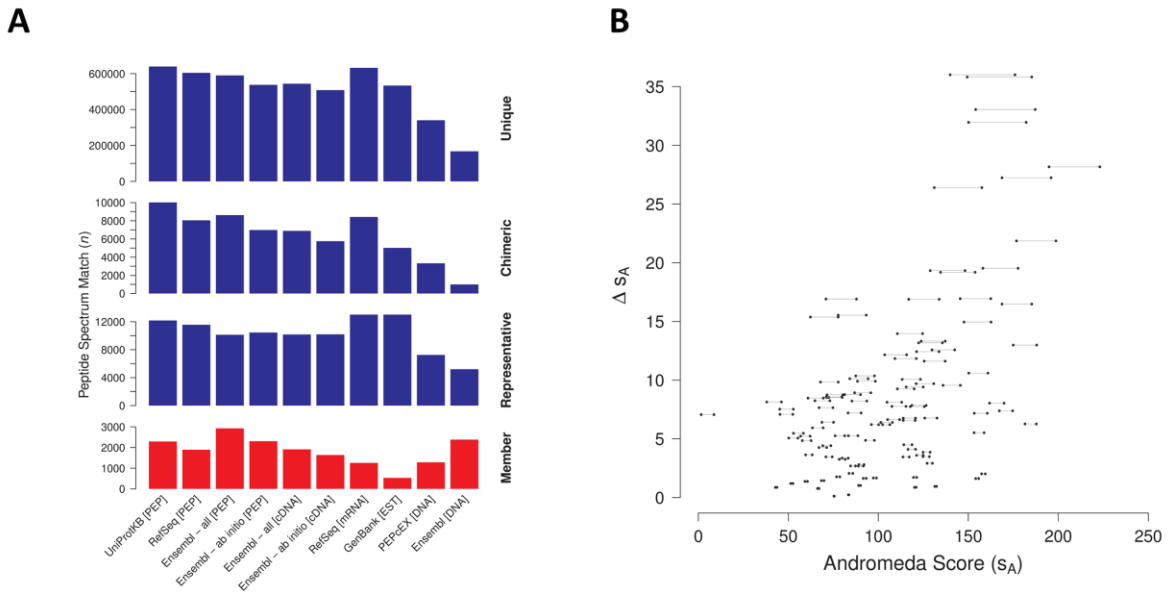

Figure 7. PSM grouping (A) The spectrum inference leads to false positive peptide spectrum assignments therefore we remove these with a naive spectrum grouping approach. The inference categories are unique and degenerate, where we define subcategories (representative, member, origin, chimeric) for degenerate PSMs. (B) Spectrum to PSM scores. 109 pairs and one triple.

In a second post processing step, we omit low (quality) scoring and insignificant peptide identifications based on the assigned search engine score and the posterior error probability.
In a recent study [38] we were able to derive objective false discovery rate (FDR) models as a function of the search engine score based on a large synthetic (phospho-) peptide library for ETD and HCD data. The optimization of the models was on a local scale, we refine these to be more comprehensive (Fig. 1) and use the HCD-Global model to process the data at hand (Fig. 1F). The derived Andromeda score criteria ($s_A \geq 58.4376$) allows us to omit 6,683 low (quality) scoring peptides out of 93,494 (Fig. 8A).
Even though the importance of statistical validity for single PSMs is critical in proteogenomics, did we apply the MaxQuant posterior error probability (< 0.01) only for peptide identifications not matching to the reference (Ensembl) annotation. Otherwise would be the criteria too conservative, filtering another 0.14 of all peptides (Fig. 8B). In total we identified 86,811 valid peptides and applied the posterior error probability in case of peptides not matching to a known gene model.

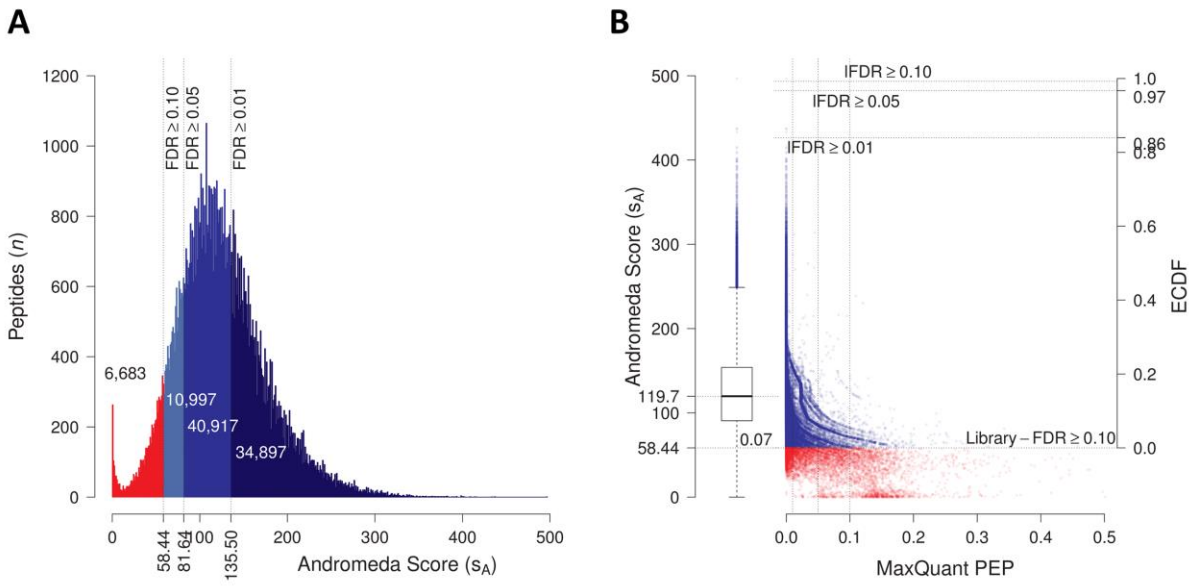

Figure 8. Peptide evidence validation. (A) Peptides as a function of the Andromeda score (s_A). FDR-thresholds (0.10, 0.05, 0.01) derived from the Global HCD, Andromeda FDR Model. To omit low quality spectra (red) we apply the 0.10 ($s_A \geq 58.44$) filter and use the subset for all subsequent data analysis (blue). (B) S_A as a function of the MaxQuant posterior error probability. To be more conservative about single PSM identifications and classifications, we use a posterior error probability threshold ≤ 0.01 covering 0.86 of the data (ECDF).

Initial database and sample characterization

To illustrate the merit of multi database searches, we characterize the 86,811 peptide identifications against the reference [PEP] Ensembl - all database (Fig. 9A, top histogram). The majority of peptides intersected (black bar) between reference and other databases. Of interest were the 19,008 complement non-reference peptide identifications, featuring prevalence in transcript databases (Fig. 9B).

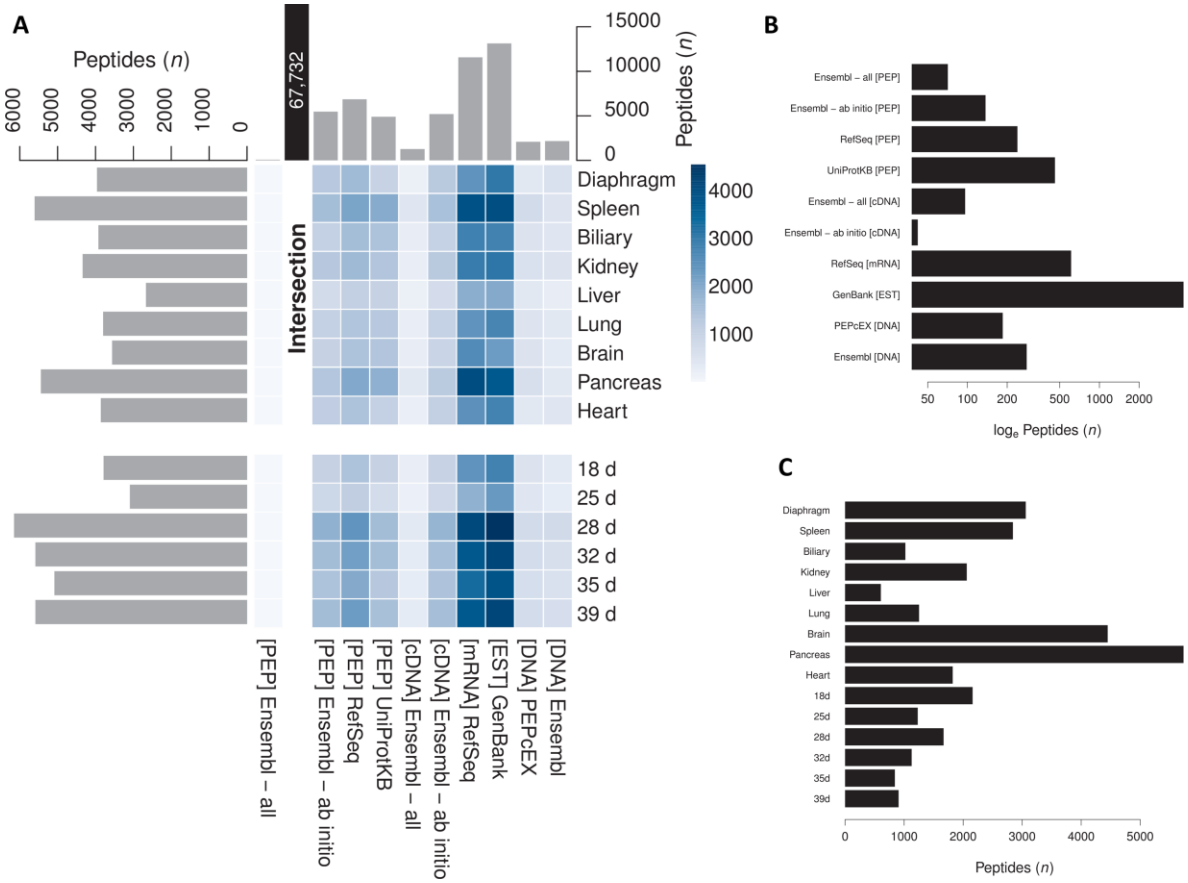

Figure 9. Peptide evidence characterization. (A) Sample to database characterization as a function of peptide abundance omitting peptides intersecting between the reference database (Ensembl [PEP] - all) and other databases. (B) Peptide identifications specific to a database and (C) biological sample.

The characterization over the biological samples resulted in 71,609 organ and 52,775 embryonic peptides (Fig. 9A, left histogram). The embryonic stages yield a gain of 15,202 specific peptide identifications over 34,036 in juvenile organs with most peptide identifications in brain and pancreas (Fig. 9C). In comparison are most peptides in embryonic stages part of early development (Fig. 9C).

The embryonic peptide identifications supported the functional annotation of proteins associated to fetal development. As an example we could provide evidence for a homologous protein (sp|Q9UEE9|CFDP1_HUMAN, UniProtKB, evidence at protein level) related to craniofacial development that was predicted as part of the [PEP] and [mRNA] RefSeq databases. The protein sequence is highly conserved with a 95% identity and 97.2% similarity between Homo Sapiens and Sus Scrofa. Additional we were able to associate proteins to a specific sample type, e.g. the RefSeq sequence gi|343790858|ref|NP_001230566.1 that is homologous to the uncharacterized proteins in Bos Taurus (UniProtKB, F1MC76_BOVIN) and Homo Sapiens (UniProtKB, CF211_HUMAN). The sequence was supported by EST and mRNA evidence and exclusively identified in embryonic stages 32d and 35d.

The initial characterization of peptide identifications illustrated the importance of RNA-Seq data in MS-based proteomics (Fig 9A, Fig. 9B, Fig. 10A). Additional were the embryonic stages a valuable and distinctive biological resource.

Peptide mapping

To distinguish reference (Ensembl) and non-reference matches required the assignment of genome coordinates to each peptide identification.

We could map 80,379 of 86,811 peptide identifications to 376,558 genomic locations. We distinguished two categories for the peptide mapping, exact (Six-Frame, Exon Graph) and approximate (BLAST, BLAT) matches, where approximate matches were supportive evidence, due to the mapping ambiguities resulting from nonsynonymous single nucleotide polymorphisms (SNPs). 67 % of the mapped peptides match to the six-frame translation, 22 % to the exon graph, 9 % are BLAT and 2 % BLAST hits. 6,432 peptides were unmappable (Fig. 10B), 61 % originating from transcript and 37 % protein databases. We assume, reasons were splice events not covered by the exon graph (> 24 exons per transcript) or SNPs with no BLAT, BLAST matches.

The peptide coordinates facilitate the notion of genome inference, resulting in 64,201 peptides with distinct genomic coordinates (unique) and 16,178 with multiple (shared). An example of an extreme shared peptide sequence is ILNPLSK, occurring in over 58,716 genome locations. Additional the peptide coordinates allowed to define 208,433 single linkage cluster [9].

The genome inference and the single linkage clustering facilitate the distinct mapping of peptide identifications relative to the reference annotation or rather coordinates.

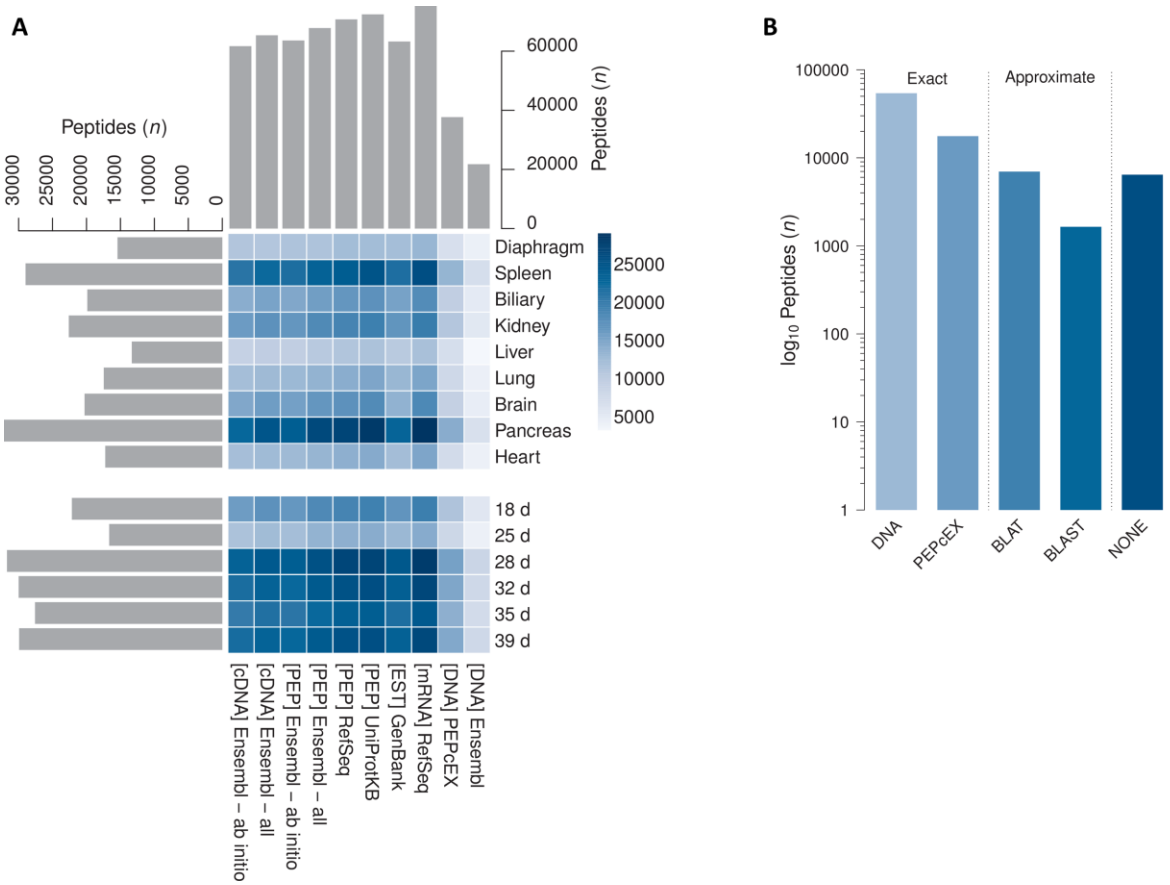

Figure 10. Peptide mapping. (A) Sample to database characterization as a function of peptide abundance over juvenile organs and embryonic stages. (B) Number of peptides to the mapping source. 6,432 peptides have no genome coordinates.

Genome annotation with proteogenomics

To annotate the genome, we defined events to classify the mapped peptides over the genome, transcriptome and proteome, referencing to the Ensembl gene, exon and transcript meta information (classification, coordinates). The top-tier classification discerns intergenic and intragenic events (Fig. 11).

In total we classify 9,565 non-redundant peptides over genome, transcriptome and proteome out of 19,008 peptides. 7,464 are intragenic and 6,266 intergenic, intersecting in 3,891 peptides due to ambiguities in classifications.

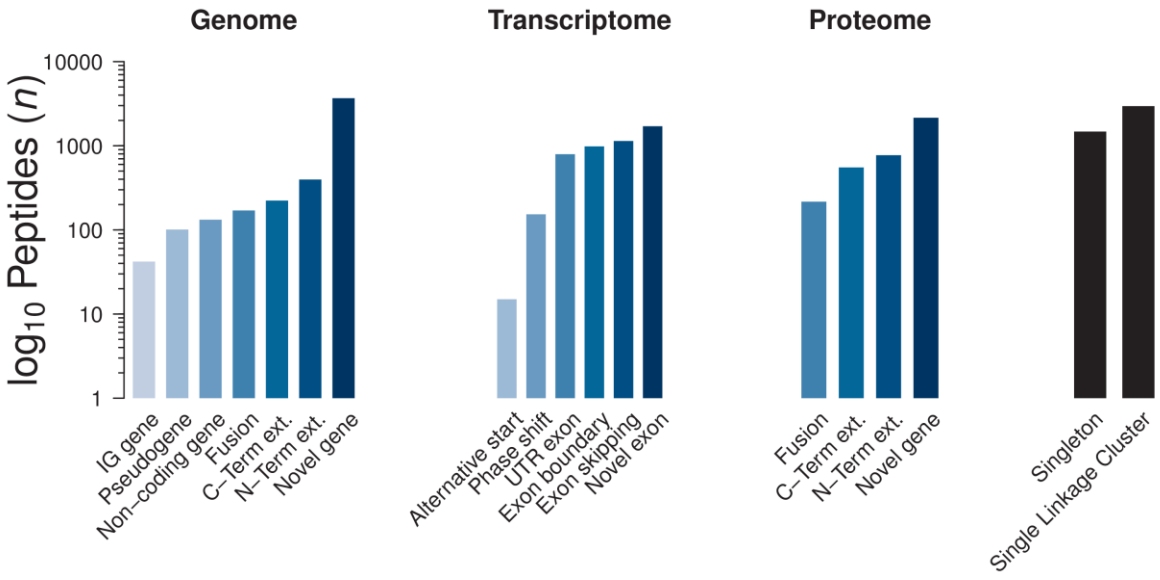

Figure 11. Peptide classification over omics databases. Complement identified and classified peptides excluding protein coding entities. Additional nonclassified peptides as singleton or single linkage cluster.

Peptide classifications - Genome

The genome intergenic events include positional classifications (Fig. 3) such as N-Term extension, C-Term extension and Fusion (between genes), as a result of the relative position of single linkage cluster members to the reference gene coordinates. Additional categories are IG-genes, pseudogenes, non-coding genes and novel genes (Fig. 11).

The gene immunoglobulin (IG) family is important for the (pre-) immune response in embryos and adult pigs. In a previous porcine study using transcript data of fetal piglets, three IGLV genes were identified to be critical for the pre-immune repertoire [45, 46]. Key players were the IGLV-3 and IGLV-8 family 20 days after gestation. We identified in total 10 IG-genes including the IGLV-3, IGLV-7 and IGLV-8 families (Fig. 11). And were able to identify and confirm the presence of the IGLV-3 and IGLV-8 gene exclusively in the 18d stage over the embryonic stages.

A gene class of much controversy are pseudogenes regarding the actual coding potential. We provide evidence matching distinct to 34 pseudogenes. The pseudogene ENSSSCG00000010221 has no protein product in Ensembl (UniProtKB, uncharacterized). The BLAST search in UniProtKB reveals close homology to other organisms with the general function “Heterogeneous nuclear ribonucleoprotein K”.
Another class of non-coding genes (no protein product) are processed transcripts not containing an open reading frame (ORF). The majority of our 18 identified non-coding genes constitute of processed transcripts comprising genes with high coverage, such as SIGLEC1 (ENSSSCG00000007146) with 42 peptides and TXLNA (ENSSSCG00000003617) with 10 peptides (Fig. 11).
The positional classifications are N-Term classifications indicating upstream open reading frames [47] or a misprediction of the translation start. Fusion classifications are likely intermediate evidence of a larger gene model including the previous models and C-Term extensions suggest a premature translation termination signal.
Novel gene classifications are evidence mapping to an Augustus gene model not present in Ensembl. We discern refined and novel gene models, i.e. refined share a classification type and therefore may overlap with an Ensembl gene. We identify 912 peptides with ambiguous classifications, comprising 99 non-coding, 211 C-Term, 65 pseudogene, 170 fusion and 367 N-Term. In total we could supplement 912 of the previous 1022 peptide classifications (0.89) with 216 respective refined gene models. The remaining 2,754 peptides are corresponding to 690 novel genes. As an example a novel gene (Augustus prediction) resides on chromosome 4, from 129,650,503 to 129,664,460 bp (c4.g2195.t1) with seven exons. A BLAST search in UniProtKB results in 95% identity with the “glycogen debranching enzyme” in Bos motus.

Peptide classification - Transcriptome

The transcriptome intragenic events include alternative start, phase shift, UTR-exon, exon boundary (exact DNA match reaching over the C-Terminus), exon skipping (alternative splicing) and novel exons (Fig. 3).
Ambiguities to genome classifications are expected for UTR-exons, exon skipping and novel exons to refined and novel gene models (N-Term, Fusion, C-Term).
We could identify 15 novel alternative start events (Fig. 11) not present in Ensembl [PEP] All, exclusive 6 in PEPcEX and the remaining with additional evidence (4 cDNA, 3 mRNA, 5 EST, 8 PEP). As an example the RefSeq [PEP] entry gi|311255064|ref|XP_003126064.1 with the predicted function “mitochondrial import receptor subunit TOM22 homolog” is uncharacterized in UniProtKB, but results in a close homolog after a BLAST search in Bos taurus with the function “Translocase of outer mitochondrial membrane 22”.
Phase shift events match to a exon model, but in a differing frame, indicating an issue with the phase assignment in the exon prediction.
The UTR-exons are similar to N-Term classifications, except the UTR-exon is part of a gene model. We identify 793 coding UTR exons, where 723 match to proteins. Ambiguities occur with exon skipping and exon boundary.
The 983 exon boundary events are related to exact matches to the genome, where the reference classification suggests a splice event, indicating false splice sites.
To count valid exon skipping events we omit all peptides present in Ensembl [PEP] all. Exon skipping is prevalent due to peptides originating from EST and mRNA sequences. 886 out of 1140 (0.78) match to (valid transcript models) proteins (cDNA, mRNA, PEP), remaining match to 193 EST and 61

PEPcEX. The majority of the mapping sources are 640 PEPcEX matches, additional 412 BLAT (high confidence) and 88 BLAST matches (low confidence). As a proof of concept for the construction of the exon graph with predicted exons are 441 PEPcEX (0.69) splice peptides originating from combinations of ab initio exons. 247 (101 Proteins, 118 Transcripts, 28 PEPcEX) of 441 ab initio are uniquely assigned to a database and 194 are in multiple databases. The exon skipping events match to 1,078 genes including 657 Augustus gene loci.
Out of 1,706 novel exon events 1,436 match to proteins (cDNA, mRNA, PEP) including 86 with an exon model but in the false frame (phase shift). The majority of the peptides intersecting between intergenic and intragenic events are due to 3,042 (3,891) peptides matching exon predictions in Augustus and Genscan, ergo not present in Ensembl.

Peptide classification - Proteome

The proteome events are identical to the genome events except peptides do not have to be a member of a single linkage cluster, allowing for distances > 12,373 bp (Fig. 11). Still 2,463 of the genome classifications intersect to a certain extend. All N-Term genome classifications (0.51 of total) are subset of the respective protein classifications (Fig. 12A). 0.33 of the fusion events (Fig. 12B), 0.25 of the C-Term events (Fig. 12C) and 0.45 novel events (Fig. 12D) intersect with genome classifications. The 1,815 of the 2,152 novel gene classifications intersect with the identical genome classification, representing 426 genes, the remaining are in 112 proteins (cDNA, mRNA, PEP). 715 (772) N-Term, 523 (552) C-Term and 213 (216) Fusion events intersect with novel genes.
An example of a novel gene exclusively present on the proteome level with no valid gene model is a Genscan transcript prediction, highly conserved in *Bos Taurus* related to the FAM107B family. An additional classification is slightly differing from fusion events, is the putative fusion, i.e. peptides matching to proteins reaching over multiple genes without evidence between genes. The 6,565 events indicate potential issues with the protein (transcript) model or the respective gene models. Another reason could be actual biological gene fusion.
In total 73,954 of 80,379 peptide identifications match in the boundaries of reference (Ensembl) genes (Fig. 12E) including the above classifications. The 24 exclusive peptides in the proteins set in comparison to the reference genes was due to peptides matching to non-coding genes.

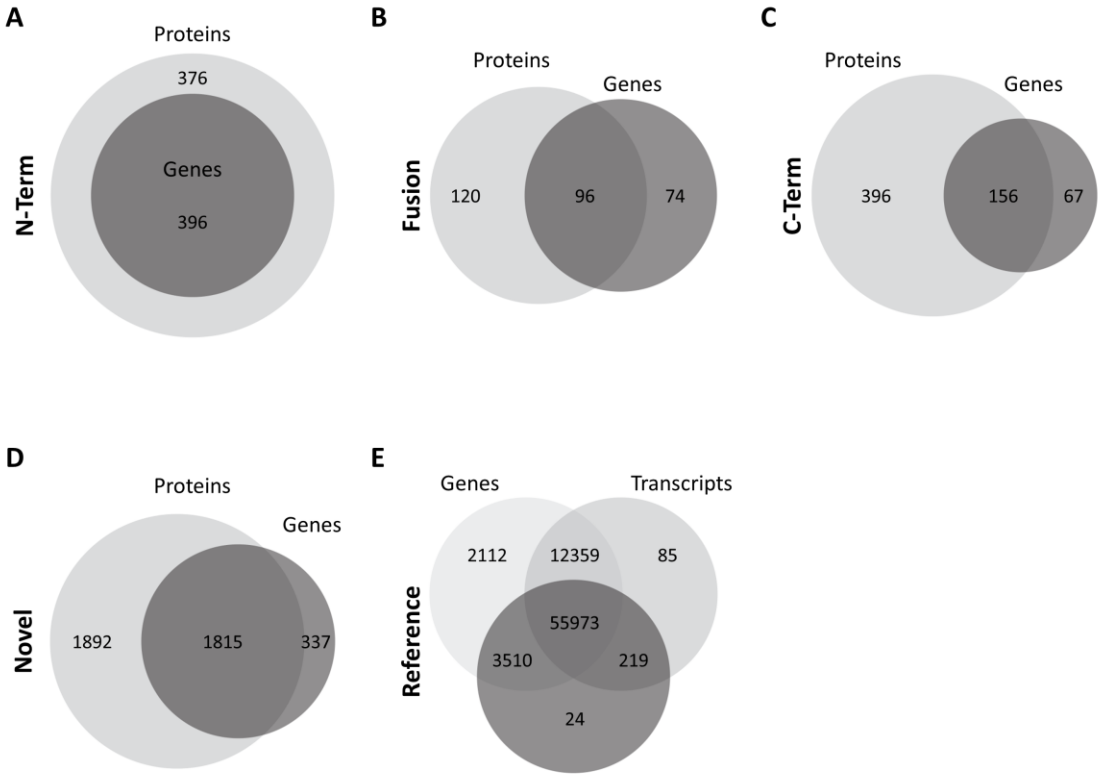

Figure 12. Peptide classification ambiguities. (A) Overlap of N-Term classification, (B) Fusion, (C) C-Term on genome and protein level. (D) Overlap of novel gene peptide evidence between genome and proteome. (E) Overlap peptide identifications matching to reference annotations.

Peptide classification - Single linkage cluster

The remaining unclassified peptides are in 2,220 single linkage cluster and products of DNA, PEPcEX, EST entities omitted in the protein classification, we identify 1,474 singletons (cluster with single evidence - 1,474 peptides) and 746 clusters (cluster with multiple evidence - 2,956 peptides). In total 437 peptides match to novel genes. As an example of such a matching cluster is the Augustus gene prediction of the small protein (73 amino acids) with close homology to human in UniProtKB (C9IZF9_HUMAN, Evidence at transcript level) with the function "Programmed cell death 6-interacting protein". Therefore we argue that single-linkage cluster can be potential coding small ORFs [48] not always detectable by gene predictors.

Inference in mass spectrometry based proteogenomics

Ambiguities in protein identification (inference) are a common issue in mass spectrometry based proteomics due to multiple isoforms originating from one gene locus but also of protein families consisting of a multitude of gene loci of known (e.g. gene duplication) and unknown origin (e.g. pseudogenes).

We investigate the inference phenomenon on the gene level in assigning to each Ensembl gene model the mapped peptide identifications. Of 7,702 identified genes, are 2,405 sharing information

in form of subsets and same sets, resulting in 1,713 links (Fig. 13). 36 % are connected in chromosomes (intra) and 64 % over chromosomes (inter). Links with a few peptides are putative noise, generating biological non meaningful links (light blue and red links). With increasing evidence are ambiguous genes an issue in the boundaries of a chromosome (84 % of all links ≥ 10 peptides) and not over chromosomes, most likely due to homologous recombination.

To further illustrate the inference issue, we analyze the Ensembl biotypes of the 2,405 genes. In total the set comprises 285 genes (11.85 %) associated with no coding potential, i.e. respective gene products are not occurring in the Ensembl protein sequence database [49].

We were able to unambiguously identify for example a processed transcript SLA-8 (ENSSSCG00000001396) representing the pseudogenes ENSSSCG00000030299 and ENSSSCG00000023113. Additional we provide distinct evidence for the expression of the processed transcript GPX3 (ENSSSCG00000017092), representing the protein coding gene loci GPX5 (ENSSSCG00000001214) and GPX6 (ENSSSCG00000001213).

As a result, we suggest to supplement the protein inference with distinct genome (gene) information to discriminate *bona fide* protein coding gene loci.

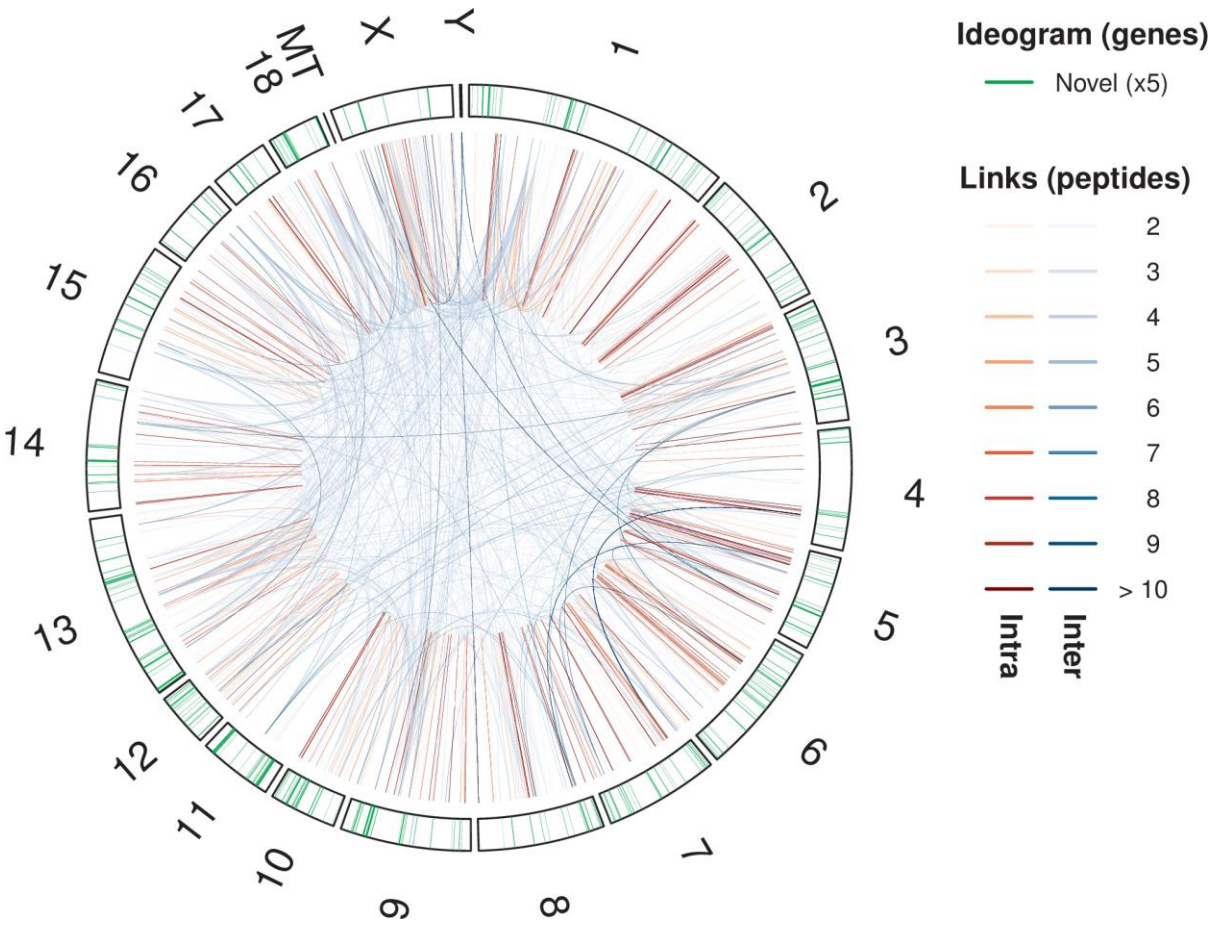

Figure 13. Gene level inference. Genome map (circle) of the autosomes (1-18), allosome (X, Y) and mitochondrial chromosome (MT). The ideogram contains 432 high confidence novel gene predictions (Augustus). Links indicate homology of genes in (red) and over (blue) chromosomes, color depth indicate the increasing amount of supporting evidence for gene loci.

Known and novel gene, transcript models and isoforms

In an attempt to illustrate the diversity of gene models from Augustus and Ensembl, we were assigning each gene model all distinct identified peptides and compared the identity (same-set) resulting in 8,819 gene models (Fig. 14A). Augustus outperforms the Ensembl pipeline including valid extrinsic information. To represent a non-redundant (assuming the longest model is valid) set of gene models, we omit sub-sets resulting in 6,834 gene models. The Augustus gene predictions include the peptide classifications on the genome level and therefore are more complement with 1,211 than 781 Ensembl non-redundant representative gene models. In 4,842 cases Augustus and Ensembl agree on the gene models.

The gene status for 6,348 Ensembl genes (6,299 distinguishable) is Known (4,353), Known by Projection (3), Putative (1), Novel (1991). Novel meaning sequence match outside Ensembl.

The identified Ensembl genes in each chromosome in comparison to the theoretical reference genes, results in an overall gene coverage of 24.34 %, where most are located on chromosome 2 (Fig. 14B). In accordance with the previous results on the peptide level we identify most genes in pancreas and the 28d stage (Fig. 14C).

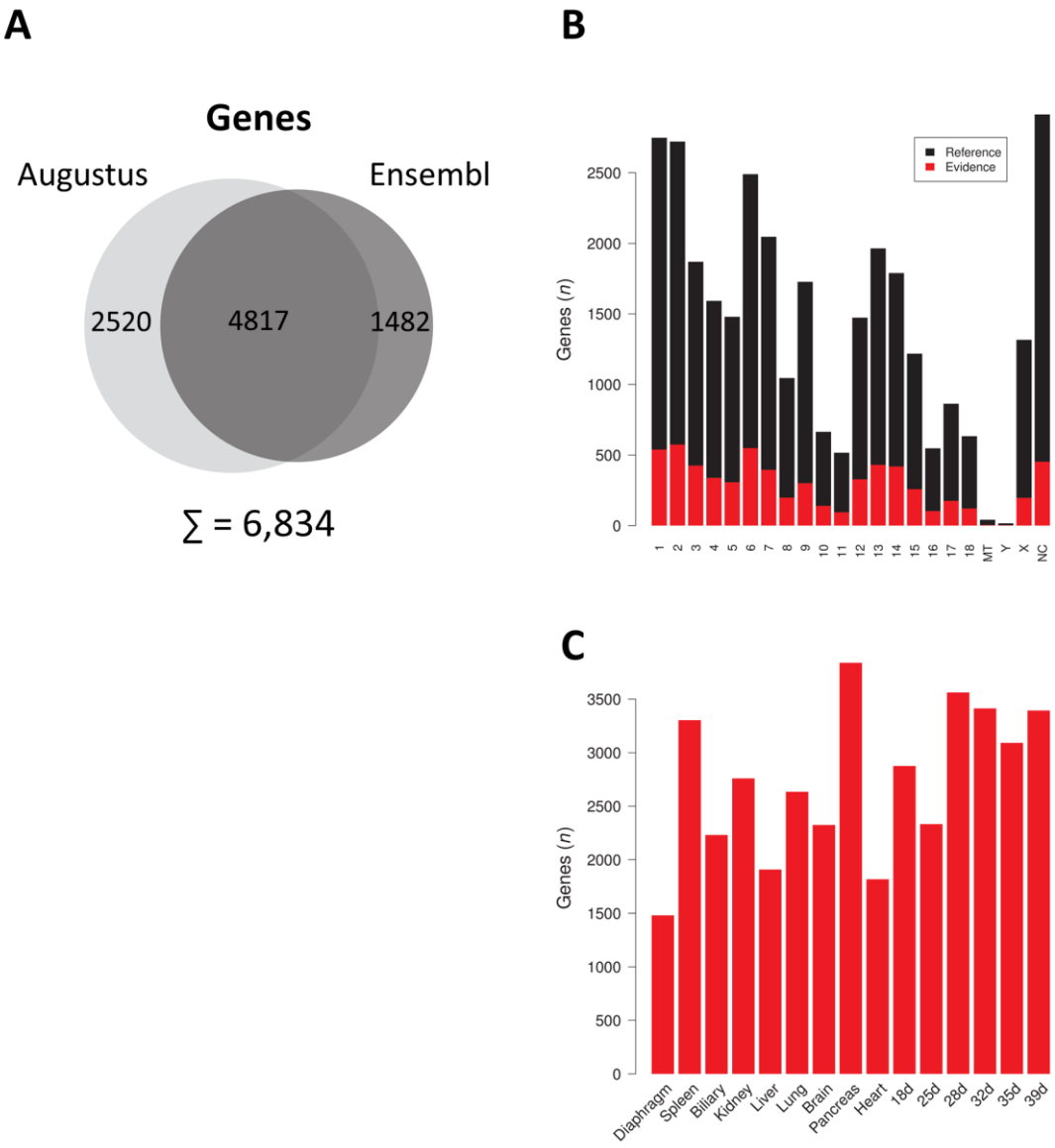

Figure 14. Gene models. (A) Same-set peptide cluster model comparison. Augustus and Ensembl gene models. (B) Reference genes (black) and with evidence (red) over chromosomes. (C) Evidence genes over samples.

In total we identify 12,120 transcript models (Fig. 15A), where Augustus provides most transcript models (7,478). The non-redundant set of Augustus and Ensembl transcript models results in 6,862 (slight increase to genes), including Genscan reduces the transcripts to 6,828 (transcripts merge genes). 1,086 (746 without subsets) novel transcript models (Fig. 15B). Grouping of unique peptides to novel models, comparing against each, no ambiguous classification (potential overlap with Ensembl models).

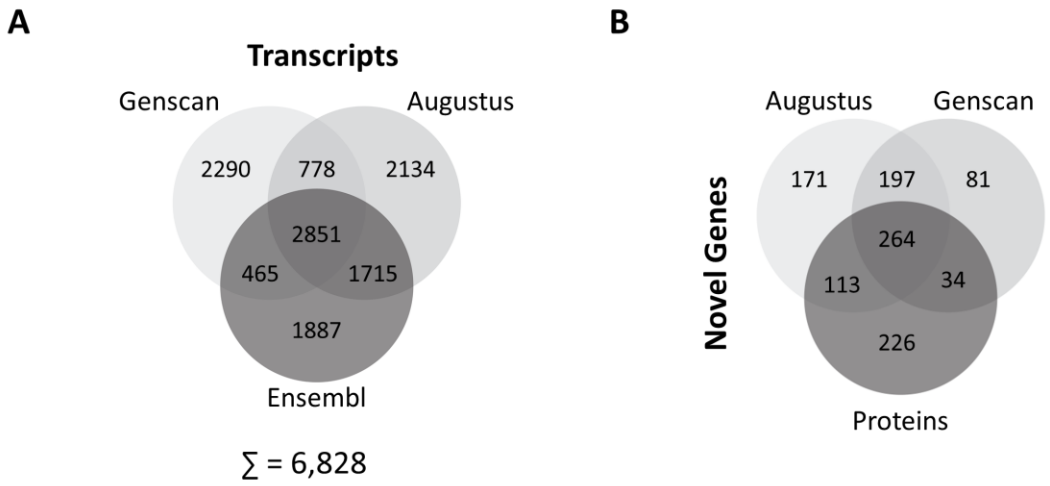

Figure 15. Transcript models. (A) Augustus, Ensembl and Genscan transcript models. (B) Overlap of novel genes between the gene predictors Augustus and Genscan to the Proteome.

9,632 protein groups, sharing 4,378 with 12,120 transcript models. Adding the transcripts, decreases the overall number of valid transcripts to 7,751 (Fig. 16A). 2,277 transcript models represent 5,180 protein groups, 3,845 in both, 1,629 protein groups represent 5,726 transcript models. 7,751 without subsets result in 7,108 after protein inference. Over all samples we identify a core proteome of 392 proteins (Fig. 16B). Pancreas (283) and brain (261) contain the most specific proteins (Fig. 16C). Each individual sample contributed to an incremental and unique number of protein identifications, ranking the samples based on the provided protein information reveals most in pancreas and least in liver (Fig. 16D).

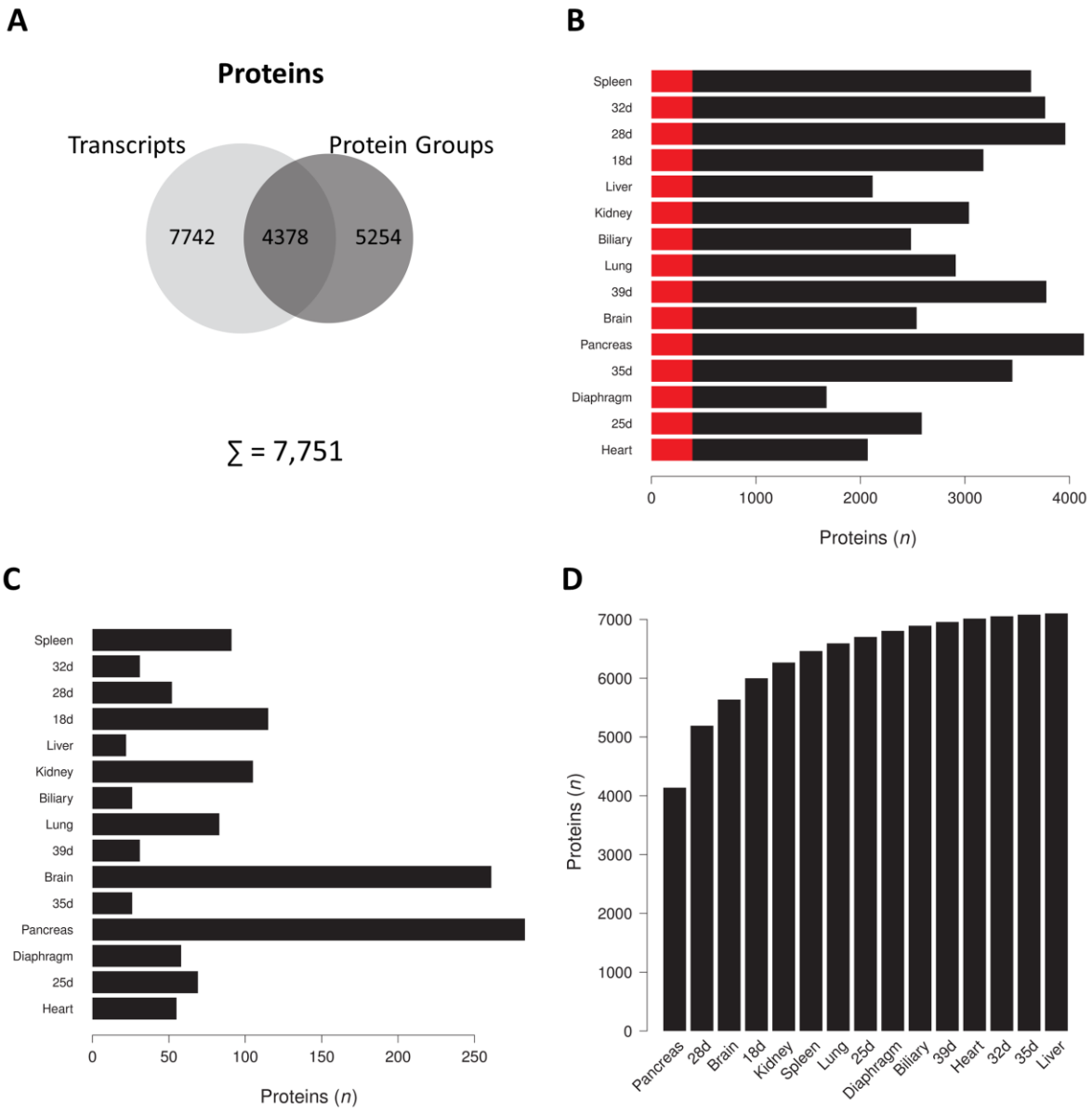

Figure 16. Protein models. (A) Overlap transcript and protein models. (B) Proteins over biological samples, red indicating the core proteome. (C) Proteins specific to biological samples. (D) Proteins over samples, in the order of a heuristic combination of samples.

To search against a comprehensive search space including models with the highest peptide evidence and therefore most probable the longest sequence, we apply a peptide centric clustering algorithm. The input databases are the Augustus transcript predictions (48,575) and all identified transcript models (cDNA, mRNA, PEP) resulting in a database of 108,816 entries. Performing a standard proteomics search increases the number of protein groups from 7,108 to 8,172 (no score filter, no gene inference, protein inference, MaxQuant grouping). In comparison our processing results in 8,067 protein groups (no score filter, no gene inference, protein inference, our grouping). We conclude that gene inference is conservative but can help in experiments addressing very specific issues (i.e. biomarker discovery) on a limited set of proteins.

Discussion

Even though the annotation of the porcine genome and proteome is in an early stage and differences are to be expected, were we able to provide comprehensive information to help to improve the annotation process of each consortium (UniprotKB, RefSeq, Ensembl). Our results suggest the usage of multiple databases for newly sequenced genomes are effective to maximize outcome of a discovery experiment. RNA-Seq data is easily attainable and capable of supplementing protein sequence databases. Furthermore the inclusion of other sources increases the coverage of novel genes, transcripts and boundaries. We were able to identify 19,008 peptides (86,811) not present in Ensembl and classified these into 6,266 intergenic and 7,464 intragenic events intersecting in 3,891 and provide evidence for 6,834 genes including 690 (432) novel, 176 refined, 34 pseudogenes, 18 non-coding and 10 IG-genes. In total we identify 7,108 proteins. Additional our data provides novel insight to proteins associated with specific functions in juvenile organs and embryonic stages.

Even though proteogenomics is advantageous in many aspects, is the search space definition and data processing subject of discussion in the community. The size and content of search spaces can lead to ambiguities in PSM assignment, due to incomplete ion series and therefore aggravating the search for single amino-acid variations [50]. Also peptide mapping is affected by the diversity in the source sequence space in comparison to the reference genome resulting in uncertainties in peptide coordinates [31], e.g. distinct peptides in proteins matching to different strands (2,365 peptides) and chromosomes (6,003 peptides). In addition is the validation of PSMs with local FDRs or rather posteriori error probabilities partly too conservative [34] leading to loss of valid peptide evidence.

We conclude that the early efforts in genome annotation are in most cases accurate, but can be vastly improved by proteogenomics. Furthermore an overestimation of protein coding genes (25,322) is prevalent and comparable to the situation of early stage genome annotation in human.

Acknowledgement

We would like to thank Mathias Wilhelm, Hannes Hahne for insightful discussions, Michael Kroetz-Fahning, Andrea Hubauer, Andreas Klaus, Steffen Loebnitz for assistance in the sample preparation and Fiona Pachl for measuring the samples.

Abbreviations

FDR	false discovery rate
IG	immunoglobulin
MS	mass spectrometry
ORF	open reading frame
PEPcEX	peptide centric exon graph
PSM	peptide spectrum match
SAP	single amino acid polymorphism

References

1. Verma, N., Rettenmeier, A. W., & Schmitz-Spanke, S. (2011, Feb). Recent advances in the use of Sus scrofa (pig) as a model system for proteomic studies. Proteomics, 11(4), 776-793.
2. Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012, Nov). Analyses of pig genomes provide insight into porcine demography and evolution. Nature, 491(7424), 393-398.
3. Humphray, S. J., Scott, C. E., Clark, R., Marron, B., Bender, C., Camm, N., et al. (2007). A high utility integrated map of the pig genome. Genome Biol, 8(7), R139.
4. Prather, R. S. (2013, Feb). Pig genomics for biomedicine. Nat Biotechnol, 31(2), 122-124.
5. Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M., et al. (2004, May). The Ensembl automatic gene annotation system. Genome Res, 14(5), 942-950.
6. Küster, B., Mortensen, P., Andersen, J. S., & Mann, M. (2001, May). Mass spectrometry allows direct identification of proteins in large genomes. Proteomics, 1(5), 641-650.
7. Ansong, C., Purvine, S. O., Adkins, J. N., Lipton, M. S., & Smith, R. D. (2008, Jan). Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic, 7(1), 50-62.
8. Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., et al. (2008, May). Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science, 320(5878), 938-941.
9. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., & Briggs, S. P. (2008, Dec). Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A, 105(52), 21034-21038.
10. Volkening, J. D., Bailey, D. J., Rose, C. M., Grimsrud, P. A., Howes-Podoll, M., Venkateshwaran, M., et al. (2012, Oct). A proteogenomic survey of the Medicago truncatula genome. Mol Cell Proteomics, 11(10), 933-944.
11. Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., et al. (2014, Jan). An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. Mol Cell Proteomics, 13(1), 157-167.
12. King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I., Eddes, J. S., Mallick, P., et al. (2006). Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. Genome Biol, 7(11), R106.
13. Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., et al. (2011, May). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and resurrected pseudogenes in the mouse genome. Genome Res, 21(5), 756-767.
14. Branca, R. M., Orre, L. M., Johansson, H. J., Granholm, V., Huss, M., Pérez-Bercoff, Å., et al. (2014, Jan). HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods, 11(1), 59-62.
15. Low, T. Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., et al. (2013, Dec). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep, 5(5), 1469-1478.
16. Bitton, D. A., Smith, D. L., Connolly, Y., Scutt, P. J., & Miller, C. J. (2010). An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. PLoS One, 5(1), e8949.

17. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., et al. (2005). Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol, 6(1), R9.
18. Castellana, N., & Bafna, V. (2010, Oct). Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics, 73(11), 2124-2135.
19. Choudhary, J. S., Blackstock, W. P., Creasy, D. M., & Cottrell, J. S. (2001, May). Interrogating the human genome using uninterpreted mass spectrometry data. Proteomics, 1(5), 651-667.
20. Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P., et al. (2007, Feb). Improving gene annotation using peptide mass spectrometry. Genome Res, 17(2), 231-239.
21. Nygard, A.-B., Cirera, S., Gilchrist, M. J., Gorodkin, J., Jørgensen, C. B., & Fredholm, M. (2010). A study of alternative splicing in the pig. BMC Res Notes, 3, 123.
22. Borchert, N., Dieterich, C., Krug, K., Schütz, W., Jung, S., Nordheim, A., et al. (2010, Jun). Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. Genome Res, 20(6), 837-846.
23. Wang, Z., Gerstein, M., & Snyder, M. (2009, Jan). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10(1), 57-63.
24. Ning, K., & Nesvizhskii, A. I. (2010). The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinformatics, 11 Suppl 11, S14.
25. Sheynkman, G. M., Shortreed, M. R., Frey, B. L., & Smith, L. M. (2013, Aug). Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol Cell Proteomics, 12(8), 2341-2353.
26. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., et al. (2014, Jan). Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res, 13(1), 21-28.
27. Wang, X., & Zhang, B. (2013, Dec). customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics, 29(24), 3235-3237.
28. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., et al. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol, 7, 548
29. Vogel, C., & Marcotte, E. M. (2012, Apr). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet, 13(4), 227-232.
30. Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics, 7, 62.
31. Specht, M., Stanke, M., Terashima, M., Naumann-Busch, B., Janssen, I., Höhner, R., et al. (2011, May). Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the Chlamydomonas reinhardtii genome. Proteomics, 11(9), 1814-1823.
32. Marx, H., Lemeer, S., Klaeger, S., Rattei, T., & Kuster, B. (2013, Jun). MScDB: a mass spectrometry-centric protein sequence database for proteomics. J Proteome Res, 12(6), 2386-2398.
33. Zhou, A., Zhang, F., & Chen, J. Y. (2010). PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. BMC Bioinformatics, 11 Suppl 6, S7.

34. Blakeley, P., Overton, I. M., & Hubbard, S. J. (2012, Nov). Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J Proteome Res, 11(11), 5221-5234.
35. Shevchenko, A., Wilm, M., Vorm, O., & Mann, M. (1996, Mar). Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal Chem, 68(5), 850-858.
36. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2010. http://www.repeatmasker.org.
37. Kolkman, J. A., & Stemmer, W. P. (2001, May). Directed evolution of proteins by exon shuffling. Nat Biotechnol, 19(5), 423-428.
38. Marx, H., Lemeer, S., Schliep, J. E., Matheron, L., Mohammed, S., Cox, J., et al. (2013, Jun). A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat Biotechnol, 31(6), 557-564.
39. Jones, A. R., Siepen, J. A., Hubbard, S. J., & Paton, N. W. (2009, Mar). Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics, 9(5), 1220-1229.
40. Käll, L., Storey, J. D., MacCoss, M. J., & Noble, W. S. (2008, Jan). Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res, 7(1), 40-44.
41. Gupta, N., Bandeira, N., Keich, U., & Pevzner, P. A. (2011, Jul). Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom, 22(7), 1111-1120.
42. Nesvizhskii, A. I., & Aebersold, R. (2005, Oct). Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics, 4(10), 1419-1440.
43. Grobei, M. A., Qeli, E., Brunner, E., Rehrauer, H., Zhang, R., Roschitzki, B., et al. (2009, Oct). Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Res, 19(10), 1786-1800.
44. Schirle, M., Heurtier, M.-A., & Kuster, B. (2003, Dec). Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics, 2(12), 1297-1305.
45. Wertz, N., Vazquez, J., Wells, K., Sun, J., & Butler, J. E. (2013, Oct). Antibody repertoire development in fetal and neonatal piglets. XII. Three IGLV genes comprise 70% of the pre-immune repertoire and there is little junctional diversity. Mol Immunol, 55(3-4), 319-328.
46. Butler, J. E., Wertz, N., & Sun, X. (2013, Oct). Antibody repertoire development in fetal and neonatal piglets. XIV. Highly restricted IGKV gene usage parallels the pattern seen with IGLV and IGHV. Mol Immunol, 55(3-4), 329-336.
47. Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., et al. (2004, Oct). Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res, 14(10B), 2048-2052.
48. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., et al. (2013, Jan). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol, 9(1), 59-64.
49. Svensson, O., Arvestad, L., & Lagergren, J. (2006, May). Genome-wide survey for biologically functional pseudogenes. PLoS Comput Biol, 2(5), e46.
50. Song, C., Wang, F., Cheng, K., Wei, X., Bian, Y., Wang, K., et al. (2014, Jan). Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. J Proteome Res, 13(1), 241-248.

Chapter 5

General conclusions

MS-based proteomics is the central high throughput-technology to profile the proteome, in conjunction with sophisticated computational strategies to process and analyze the resulting data. The objective of this thesis was to develop novel approaches for database searching, in particular improve aspects of the theoretical search space and means to validate the results. The thesis concentrates on the construction process and composition of sequence databases and the subsequent statistical validation of peptide identifications and phosphorylation site localization.

In database searching is the database choice of utmost importance, as the contents or rather proteins of the database restrict the success of a discovery experiment. To address this issue, a clustering algorithm was conceived, enabling the grouping of multiple protein sequence databases to construct a comprehensive search space as well as reflect the peptide centric nature of proteomics data. As part of a pipeline, referred to as mass spectrometry-centric database (MScDB), facilitates an increase in the peptide to protein ratio in contrast to common sequence clustering approaches. Analysis of database searching with MScDB against a cancer cell line and human placenta, results in peptide identifications and single amino acid polymorphisms undetectable by a sequence clustered database such as UniProtKB.

In the next generation sequencing era, a plethora of genomes and transcriptomes for a multitude of organisms are available. To make this resources attainable to database searching, novel approaches in the field of proteogenomics are required. To this end a tailored strategy was developed to combine the search results of multiple databases and control with an objective criteria the quality of the data in a genomic search space. The proteogenomic analysis of nine porcine juvenile organs and six embryonic stages, yielded 176 refined and 690 novel gene models.

The validation of the search results is an integral part of database searching, to discern true and false peptide identifications as well as the correctness of localization of post translational modifications. To address this predominant statistical issue, a synthetic reference peptide and phosphopeptide library was synthesized to derive a more objective criteria. The library enabled the validation of peptide identification and phosphorylation site localization algorithms. And also made the systematic analysis of the behavior of unmodified and modified peptides in a liquid chromatography system possible.

List of publications

1. Wilhelm M., Schlegl J., Hahne H., Gholami A. M., Lieberenz M., Savitski M. M., Ziegler E., Butzmann L., Gessulat S., **Marx H.**, Mathieson T., Lemeer S., Schnatbaum K., Reimer U., Wenschuh H., Mollenhauer M., Slotta-Huspenina J., Boese J.-K., Bantscheff M., Gerstmair A., Faerber F., Kuster B. Mass spectrometry based draft of the human proteome. *Nature* (2014) (in press)

2. **Marx H.**, Lemeer S., Schliep J. E., Matheron L., Mohammed S., Cox J., Mann M., Heck A. J. R., Kuster B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol.*31(6), 557-64 (2013)

3. **Marx H.**, Lemeer S., Klaeger S., Rattei T. and Kuster B. MScDB: a mass spectrometry-centric protein sequence database for proteomics. *J Proteome Res.* 12(6), 2386-98 (2013) (partly diploma thesis)

Acknowledgement

First and foremost, I would like to thank Bernhard Küster and Dmitrij Frishman for the supervision of my thesis.

On a personal note, I would like to thank my friends, family and better half for the constant and loyal support. Along the lines of my thesis, to boil things down to what really matters, a few wishes, puns, memes and attributes for each I value most.

Amin,	keep calm and carry on.
Andreas,	the best.
Ben,	a true friend. Thanks for lending me an ear in dire straits.
Erhard,	a cool dad.
Hannes,	the towel.
Horst,	forgive me for not giving you a proper farewell.
Mathias,	101010.
Ursula,	a bastion of calm, at times spiky but underneath pure gold.
Xüsha,	means the world to me.

Curriculum vitae

PERSONAL DETAILS

Name:	Harald Marx
Date of birth:	16.10.1981
Address:	Technische Universität München Chair of Proteomics and Bioanalytics Emil-Erlenmeyer-Forum 5 85354 Freising Germany
Phone:	+49(0)8161-714369
Email:	harald.marx@tum.de

EDUCATION

Doctor rerum naturalium, Bioinformatics — 16.12.2009 - present

Technische Universität München, Munich, Germany
Supervisors: Prof. Kuster and Prof. Frishman

Diploma, Bioinformatics — 01.10.2004 - 30.09.2009

Technische Universität München and Ludwig-Maximilians University, Munich, Germany
Supervisors: Prof. Rattei and Prof. Kuster

AWARDS

Fellowship, University of Bavaria e.V. (Elite Network of Bavaria) — 26.10.2010 - 30.06.2013

Fellowship, DFG funded International Research Training School RECESS (Regulation and Evolution of Cellular Systems) — 16.12.2009 - 01.01.2013

TEACHING

Courses:

Bioinformatics for Biosciences I — 01.10.2012 - 31.03.2013
Technische Universität München — 01.10.2011 - 31.03.2012

Bioinformatics for Biosciences II — 01.04.2012 - 30.09.2012
Technische Universität München — 01.04.2011 - 30.09.2011
01.04.2010 - 30.09.2010

Bachelor projects:

Refinement of human gene models by mass spectrometry based proteomics. — 04.03.2013 - 04.06.2013

Generation of phosphopeptide libraries based on public data repositories. — 25.03.2011 - 01.08.2011

Comparative shotgun proteomic analysis of porcine organs. — 10.05.2010 - 25.06.2010

TALKS

Micromethods in Protein Chemistry, Bochum, Germany — 24.06.2013 - 26.06.2013

POSTERS

61st ASMS conference on Mass Spectrometry and allied Topics, Minneapolis, USA — 09.06.2013 - 13.06.2013

60th ASMS conference on Mass Spectrometry and allied — 19.05.2012 - 24.05.2012

Topics, Vancouver, Canada	
German Conference on Bioinformatics, Freising, Germany	07.09.2011 - 09.09.2011
Moscow Conference on Computational Molecular Biology, Moscow, Russia	22.07.2011 - 24.07.2011
RECOMB Satellite conference on Computational Proteomics, San Diego, USA	11.03.2011 - 13.03.2011