Computational Proteomics: from numbers to biology

Henk van den Toorn

The research in this thesis was performed in the Biomolecular Mass Spectrometry and Proteomics Group at Utrecht University with financial support from the Netherlands Proteomics Centre (NPC) and Proteins@Work.

Print: ProefschriftMaken || www.proefschriftmaken.nl

Cover design: Annette van den Toorn

ISBN: 978-90-393-6821-3

In memory of my father Henk W. van den Toorn,
dedicated to my family and my partner.

Computational Proteomics: from numbers to biology

Computationele Proteomics: van getallen naar biologie
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge
het besluit van het college voor promoties in het openbaar te verdedigen
op maandag 16 oktober 2017 des middags te 2.30 uur door

Hendrik Willem Pieter van den Toorn

geboren op 8 december 1970 te Tricht, gemeente Geldermalsen

Promotor:	Prof. dr. A.J.R. Heck
Copromotor:	Dr. ir. B. van Breukelen

Table of contents

Detailed table of contents

1. Introduction

The field of biological research is increasingly trying to understand the complex processes that are too small to observe by eye, even when using a microscope. The mechanisms of life take place at the molecular level, therefore the only way to investigate their workings is to use molecular tools. Since the elucidation of the simple structure of DNA as the carrier of genetic information (Watson and Crick, 1953), and the identification of genes as physical functional units, it has become possible to get a handle on the biological networks within the cell. DNA and RNA sequencing have become indispensable in uncovering many mechanisms, and the techniques to sequence these molecules has become extremely powerful over the last decades, yielding enormous amounts of data. At the same time, it has always been clear that proteins ultimately perform most of the work in the cell and that studying them directly would give specific insight into how cells work, and what may go wrong. Studying all proteins is called proteomics (Wilkins et al., 1996), and since this is a high-throughput technique, computational analysis is indispensable. This introduction will describe the bioinformatics techniques necessary to analyze proteomics experiments, with references to the appropriate chapters which go more in-depth into the subject.

Transcriptomics techniques such as microarrays and RNA sequencing created the opportunity of reading out the RNA content of biological samples (Wang et al., 2009). It became possible to study which genes' expression changes upon external stimuli or mutations, or to find expression differences between tissue types. Unfortunately, RNA sequences do not necessarily represent expressed proteins because of degradation of spurious transcripts. This is because of differential splicing and post-translational modifications, for any gene regularly there exist multiple variations at the protein level, called proteoforms (Smith and Kelleher, 2013). Additionally, when looking at quantification, RNA levels do not fully explain protein levels, probably because of differences between RNA and protein turnover rates. Consequently, we need proteomics to investigate quantification and protein forms to gain biological insight. Historically, different proteomics techniques were developed, with 2d-gel electrophoresis (2DIGE) as an important technique. Nowadays, liquid chromatography coupled with mass spectrometry (LC-MS) has taken the main stage in the analysis. Understandably, many new computational techniques are necessary and

indeed are at the foundation of this field. In the following paragraphs a brief overview of LC-MS technology is given, touching upon the subjects where bioinformatics play an important role. These bioinformatics questions are then explained in more detail in further sections of this introduction, placing the following chapters into perspective.

An overview of bioinformatics in LC-MS technology

LC-MS proteomics

The mass spectrometer has been instrumental in advancing proteomics, enabling the high throughput analysis of many proteins within a relatively limited time. A mass spectrometer measures a mass spectrum of a mixture of masses. There are several techniques to use the information in these mass spectra for the identification of proteins. One of the simplest techniques is Peptide Mass Fingerprinting (PMF, Figure 1) (Henzel et al., 1993; James et al., 1993; Mann et al., 1993; Pappin et al., 1993; Yates et al., 1993), useful for the analysis of a single protein, typically isolated from a 2d-gel. The protein is digested with an endopeptidase to produce peptides after which the masses are used as a 'bar code' for the protein in the sample. This bar code is then compared to the masses calculated from the proteins already in a sequence database. For complex mixtures of peptides, the sample needs to be separated using one or more separation steps with high pressure liquid chromatography (HPLC). This separation dramatically reduces the complexity of the sample presented to the mass spectrometer so that individual peptides can be selected on their mass and fragmented inside the mass spectrometer. This allows for the reconstruction of the peptide amino acid sequence and the identification of the protein it originated from. There are several different techniques available to cause peptide dissociation. Some often-used techniques include Collision Induced Dissociation(Wells and McLuckey, 2005) (CID), Higher-energy Collision-induced Dissociation (Olsen et al., 2007) (HCD) and Electron Transfer Dissociation (Mikesh et al., 2006; Syka et al., 2004) (ETD). Globally, dissociation preferentially takes place at the peptide backbone, with a preferential difference of atomic bond depending on the fragmentation mechanism. CID and HCD rely on collision of the peptides with an inert

gas such as Helium or Argon, essentially heating the molecule up to a point where the vibrational energy causes a bond to break. Unfortunately, several modifications such as phosphorylation and glucosylation have a bond that is affected by these energies as well, and are lost from the peptide, creating ambiguity for the analysis of the localization of the modification. This problem is alleviated by ETD, where the transfer of an electron causes the backbone to break but will not affect the modification.

When ETD became available on the Orbitrap mass spectrometer, allowing for the selection of either the CID or ETD method for each peptide,

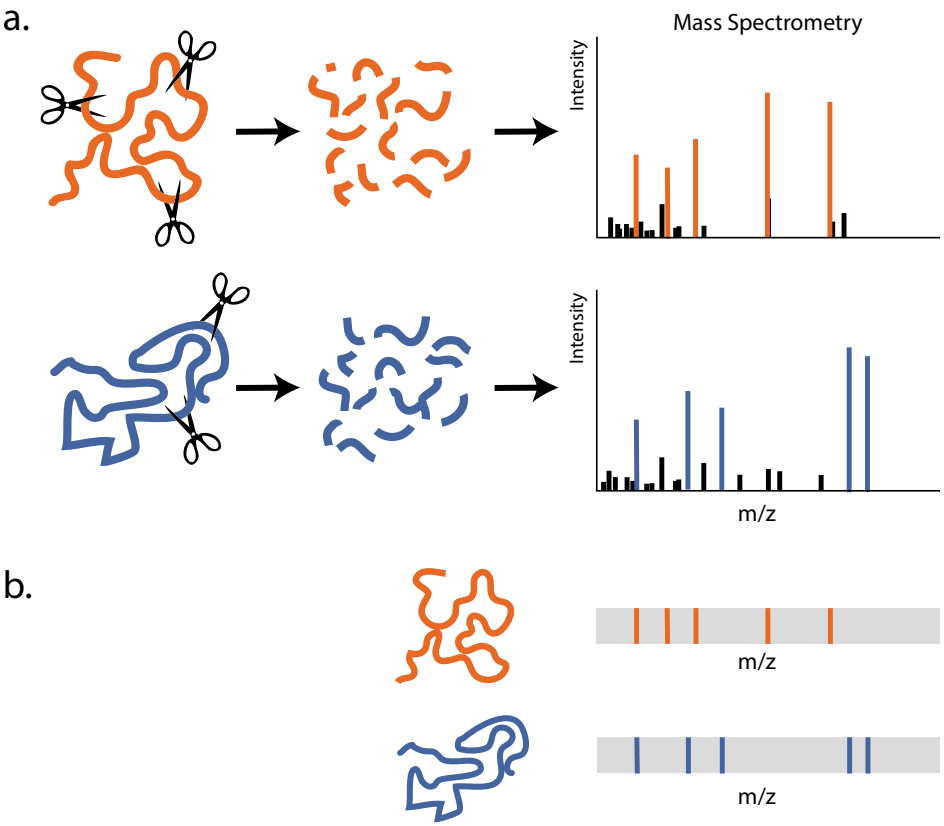


Figure 1. The principle of Peptide Mass Fingerprinting.

a. Proteins are cleaved by proteolytic enzymes (depicted by scissors) at defined amino acids. The masses of the resulting peptides are determined on the mass spectrometer. **b.** The peptide masses form a pattern which can be predicted using the known sequence databases and therefore form a 'fingerprint' for each protein, which can be compared to the mass spectra.

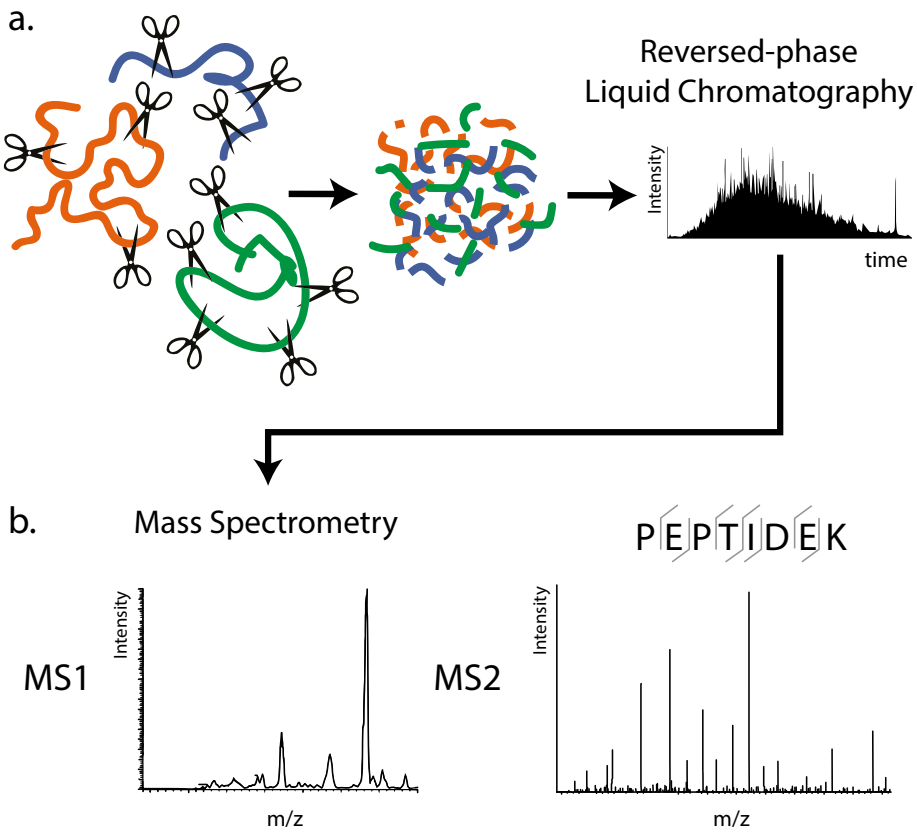


Figure 2. The principle of fragmentation based peptide identification.

a. Proteins are cleaved by proteolytic enzymes (depicted by scissors) at defined amino acids. The mixture of these proteins is subjected to chromatographic separation, such as reversed-phase liquid chromatography, to reduce complexity before mass spectrometry analysis. **b.** Single masses from the peptide survey scan (MS1) are selected and fragmented in the mass spectrometer, whereafter the fragment mass spectrum is analyzed for sequence identification.

it became feasible to compare the performance of the techniques on the same sample. Several studies were performed on the advantages of this technology, one of which is presented in chapter 4, describing the global overview of peptide properties of these differences. It is shown that for identification, ETD and CID are complementary as they target different types of peptides. Currently, it is standard practice to select the fragmentation method based on precursor mass and charge (Swaney et al., 2008).

Peptide identification algorithms for LC-MS proteomics

A single mass spectrometry run typically yields tens or hundreds of thousands of fragment spectra, for which we want to identify the peptides and ultimately, the proteins they belong to. Much data about what proteins may be present is available in online databases, which can therefore be used to create possible fragmentation spectra to compare to the observed spectra. The database search method uses a simplified model of the chemical and physical events that occur during the analysis of a protein molecule. First, the proteolytic enzyme is simulated, based on the preferred cleavage sites on the protein. From the resulting peptide sequences the masses can be calculated. These masses can then be used to create a lookup table from mass to peptide sequence, extended to take into account (post-translational) modifications. Ignoring the real-world peptide separation, the pipeline then continues to simulate the fragmentation method by calculating the appropriate ion series masses for each peptide entry (Steen and Mann, 2004).

Since the mass spectrometer works by selecting peptide masses for fragmentation, the selected peptide mass is used to find matching (modified or unmodified) peptide sequences in the lookup table, allowing for mass deviations. The observed fragments can be matched to the various theoretically derived fragmentation spectra to produce a Peptide to Sequence match (PSM). Since several peptides can match the mass of a single precursor (amongst other things, because of inaccuracies in the measurement), a score is needed to rank between the matches and to indicate confidence for the match. A major difference between different database search engines is the way they calculate the score. Sequest (Eng et al., 1994) was the first search algorithm available and is based on the correlation between theoretical and observed fragments XCorr. The Mascot (Perkins et al., 1999) search engine has a probabilistic scoring model known as the 'ions score' where the score is $-10*\log_{10}$ of a p-value that some match is found by chance. Unfortunately, since the scoring algorithm is proprietary it is not possible to describe its principle in detail. A scoring method that is available is used by MaxQuant (Cox and Mann, 2008), which is based on the binomial distribution:

$$s(q) = -10\,log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right].$$

Where q is the number of peaks retained in a 100 m/z window and chosen as the probability to obtain a match by chance, *k* is the number of matched ions and *n* is the number of theoretical ions. This score represents the chance of obtaining the number of matched ions by chance, based on the number of ions maximally retained in each 100 m/z window. Interestingly, this score does not take peaks into account that do not belong to the theoretical spectrum. The formula also clearly conveys the necessity to choose the correct theoretical ions to match with. The ion series are readily established from the fragmentation type used, but the number of neutral losses are mostly dependent on unknown factors, therefore the score algorithm chooses whether to include the neutral losses based on the highest score. While PSM scores are calculated for each individual match, the problem is how to set the threshold for inclusion in a final list, in an experiment. Another problem of the PSM scores is that they are hard to compare amongst each other, even within the same search engine but for instance between different fragmentation methods. In an attempt to unify these scores and to have standardized quality thresholds, the proportion of false positives amongst the reported hits, called the false discovery rate (FDR) is used. The FDR is unknown and therefore needs to be estimated, and different methods are available for that.

FDR estimation

There are different levels of identification for which FDR can be calculated: at the PSM level, at the (grouped) peptide level and at the protein level. Since it is not known which of the reported identifications is false, the FDR has to be estimated (see Table2). For PSMs, different strategies were devised, one of the most simple is the target-decoy approach proposed by Gygi and coworkers (Elias and Gygi, 2007). In this strategy, all spectra are searched against a database with existing 'target' sequences, as well as a database with shuffled (ideally non-existent) 'decoy' sequences. There are two variations on this strategy: either two separate searches are conducted, or the target and decoy databases are combined and used in a single search and as such the spectra 'compete' for matches. Moreover, there are several ways to construct the decoy sequences, although the differences between them are not large (Bianco et al., 2009). The decoy PSMs are expected to have a distribution of lower scores due to pure chance. The false positive PSMs

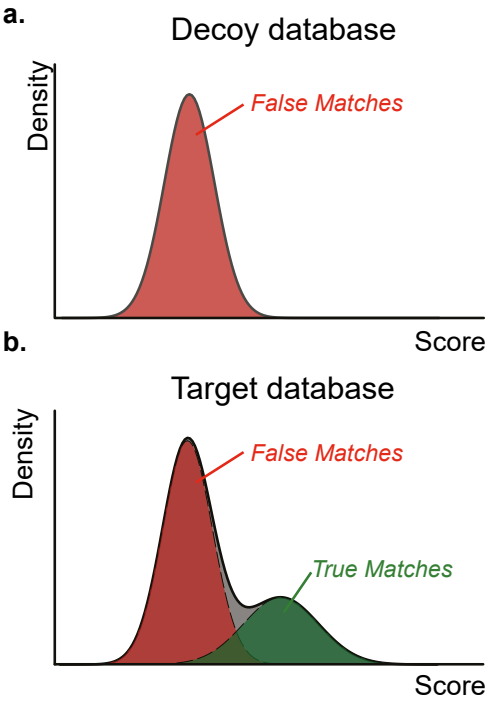


Figure 3. Hypothetical score distributions for scores of spectra to a decoy database (**a.**) and scores of spectra to a target database (**b.**). The 'target' PSMs consists of a mixture of false positive identifications with a low score distribution, similar to the 'decoy' distribution, and a true positive distribution with higher scores

from the target database are expected to have the same score distribution as random matches, i.e. the decoy PSM score distribution (Figure 3). The 'true' matches from the target database should have a distribution of higher scores. Therefore, for all PSMs with the target database, the score distribution will be a mixture of the low and high score distributions. The proportion of the decoys of the total number of identifications above some score threshold is the estimation of the false discovery rate. Similarly, a model-based approach is PeptideProphet algorithm (Keller et al., 2002), which tries to fit the two mixed score distributions. For Mascot, PeptideProphet simply uses the ions score. For Sequest there are several scores available, such as XCorr, deltaCN and mass delta, for which a weighted 'summary' score D is calculated. Building on the idea that multiple features can be used to discriminate 'good' versus 'bad' identifications, the Percolator algorithm uses support vector machines in a multidimensional feature space (Käll et al., 2008; Spivak et al., 2009). Percolator trains the SVM on the data itself, using cross-validation to estimate the FDR on the selected thresholds. Especially for the Mascot search engine, there was no straightforward way to filter the results, therefore we developed a suite of utilities that reads the Mascot ".dat" search result file directly, filters the PSMs based on a target-decoy or the Percolator algorithm and writes out a new .dat file that is compatible with Mascot

and other downstream tools. For manual inspection, the utility is able to freely plot any relation between numerical features of PSMs. Rockerbox is described in Chapter 2.

Quantification

If the identity of the proteins present is considered to be the qualitative state of the cell, a complete overview would need to include quantitative data. Cell activities are regulated by adjustments in the activity of the proteins. Protein activity on the one hand is regulated by physical changes at the molecular level such as conformation, translocation, modification or proteolytic cleavage and on the other hand by their levels. The mRNA levels are actively regulated by the cell upon stimuli, and can be measured with microarrays and quantitative RNA-seq experiments. Unfortunately, protein levels are not predicted well by mRNA levels (MacKay et al., 2004), even if there are corrections based on sequence that explain the difference so some extent (Vogel et al., 2010), so it is advantageous to be able to measure protein levels directly.
There are several protein quantification techniques, which can be divided in label-free and isotopically-labeled techniques. Label-free techniques read protein levels from one mass spectrometry analysis for each sample, whereas isotopically labeled peptides allow for simultaneous analysis of different samples in a single analysis at the cost of sensitivity. From these measurements, either an estimation of the absolute protein levels are made, or the measurements are divided to give a ratio value between two conditions.

Label free quantification

The simplest label-free technique is spectral counting. Since the speed of modern mass spectrometers is high, a peptide is likely to be fragmented several times, depending on its concentration(Griffin et al., 2010). Therefore, protein abundance can be estimated by counting the number of Peptide to Spectrum Matches (PSMs) or peptides for every protein. The main concerns raised against spectral counting are that it is prone to errors from instrument-specific parameters such as sequencing speed, ionization effi-

ciency and dynamic exclusion. Moreover, a substantial amount of spectral counts per protein is necessary to allow for robust statistical analysis (Old et al., 2005). Another measurement is the area under the elution profile of the peptides, extracted from the combined MS1-spectra, called extracted ion chromatogram (XIC). In general, these approaches are considered to provide better precision than spectral-counting (Schulze and Usadel, 2010). Moreover, this measure is shown to be approximately linear to the protein concentration (Purves et al., 1998). Several quantification techniques use this measurement, the simplest being Top 3 Protein Quantification (T3PQ) (Grossmann et al., 2010; Silva et al., 2006), i.e. the average of the XIC areas of the "top 3 peptides".

Since not all peptides are selected for fragmentation in an MS-run, this under-sampling can be addressed by combining identifications of another run ("match between run"). For that the peaks between different runs must be matched. Ideally this would be a trivial task, but variations in elution time between experiments are both non-linear and unpredictable to an extent, therefore many algorithms have been proposed(Smith et al., 2015). Some important variations are as Correlation Optimized Warping based on COmponent Detection Algorithm (COW-CODA) (Christin et al., 2010) or MaxLFQ (Cox et al., 2014). Alternatively, one can make use of Accurate Mass-time Tag (AMT) databases (Lipton et al., 2002), where the (normalized) elution time of high quality previously identified peptides are mapped onto the elution profile of a new sample. For experiments involving more than two conditions or replicated measurements, multiple pair-wise alignments may have to be performed to make all chromatograms comparable.

To avoid the inter-run matching, and to reduce the amount of mass-spec runs, it is also possible to use stable (non-radioactive) isotopes to label different samples, and analyze them together in a single mass spec run. In practice several methods are used: Stable Isotope Labeling of Cells (SILAC), is done by supplying amino acids containing heavy isotopes such as ^{13}C and ^{15}N in the of the growth medium to cultured cells, or the diet for test animals such as mice or rats. After enough cell divisions or generations, a "reference" strain with most proteins labeled is available. This strain can then be used to compare the protein levels to an experimental treated sample in a single mass spectrometry experiment. The application

Table 1. An overview of different protein quantification methods.

Name	abbr.	Remarks
Count based		
Protein Abundance Index (Rappsilber et al., 2002)	PAI	Peptide variants counted (charge, modifications) / theoretical number of observable peptides per protein
Exponentially modified PAI (Ishihama et al., 2005)	emPAI	10^{PAI} -1
Spectral Abundance Factor (Liu et al., 2004)	SAF	Spectral count / aa_{prot}
Normalized SAF (Zybailov et al., 2006)	NSAF	SAF / SAF_{total}
F_{abb} (Aye et al., 2010)	F_{abb}	Spectral count / protein mass
Absolute Protein Expression(Braisted et al., 2008; Lu et al., 2006)	APEX	spectral counts normalized for peptide detectability
Precursor intensity based, label free		
Top 3 peptides (Grossmann et al., 2010; Silva et al., 2006)	T3PQ	Sum of the intensities of the 3 most highly expressed peptides
Intensity based absolute quantification (Schwanhausser et al., 2009)	iBAQ	
Label free quantification		Calculate area of XIC, eg. SuperHirn (Mueller et al., 2007), ProtQuant (Bridges et al., 2007), MaxLFQ (Cox et al., 2014).
Precursor intensity based, labeled		
Metabolic labeling		
^{15}N labeling(Conrads et al., 2001; Oda et al., 1999)		Modification mass delta is dependent on peptide sequence
Stable isotope labeling by amino acids in cell culture (Ong et al., 2002)	SILAC	Currently also available as selected completely labeled organisms
Chemical (in vitro) labeling		
isotope coded affinity tags(Gygi et al., 1999)	ICAT™	Cysteine binding thiol
^{18}O labeling(Yao et al., 2001)		Heavy / Light labeled peptide
Dimethyl labeling		Heavy / Medium / Light labeled peptide
Isobaric labeled peptides distinguished at the fragment ion level		
Tandem Mass Tags(McAlister et al., 2012; Thompson et al., 2003)	TMT	Up to 10-plex
Isobaric Tags for Relative and Absolute Quantification(Ross et al., 2004)	iTRAQ®	
Targeted approaches, MS2 based elution profiles		
Selected Reaction Monitoring / Multiple Reaction Monitoring(Kondrat et al., 1978; Lange et al., 2008)	SRM, MRM	

of SILAC is limited to the existing reference samples or organisms, and by the time to create a new reference. To circumvent these problems, alternatively samples can be labeled after protein extraction and cleavage. Labeling techniques include dimethyl labeling, where Lysines and the N terminus of the peptides are covalently linked to isotopic variants of dimethyl(Boersema et al., 2009; Hsu et al., 2003; Huang et al., 2006), and ^{18}O labeling that incorporate the heavy isotope during proteolytic cleavage (Yao et al., 2001).

Programs like Proteome Discoverer (Thermo Fisher Scientific, Bremen, Germany), Peaks, MSQuant (Gouw and Krijgsveld, 2012; Mortensen et al., 2010), MaxQuant (Cox and Mann, 2008) and ASAPRatio (Li et al., 2003) enable the quantification of peptides and proteins . Another example, based on centroided data is the pview algorithm, using space-partitioning algorithms for quick access, combined with methods from graph theory to group together peptide to substantially decrease processing

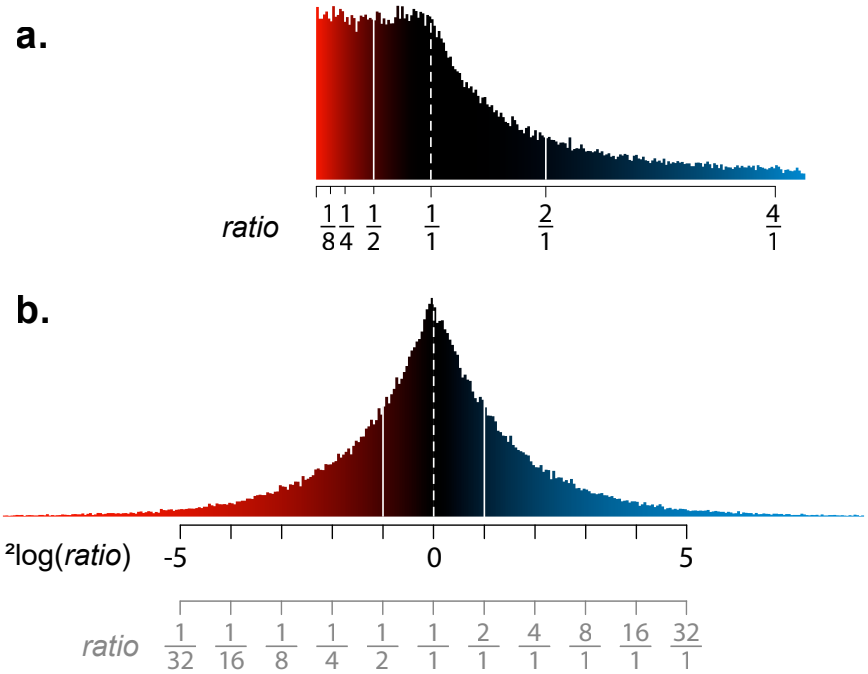


Figure 4 Effect of log2-transforming ratio data. **a.** Example data, ratios of random numbers drawn from a uniform distribution, **b.** Log2 transform of the same ratios, with a secondary axis to demonstrate the symmetry of their values around the 1:1 ratio. In both figures a solid white line indicates a 2x relative difference, the dashed line is the 1:1 ratio,

time (Eastman, 2009; Khan et al., 2009).
The different channels can then be extracted from the chromatogram when looking at a known mass difference, within an elution time window for so called *peak pairs* (or *triplets*). Most software will try to extract peak pairs first, and later assign peptide sequencing events to the peaks. For ^{15}N-labeling this is problematic, since the mass difference between two peaks depends on the amino acid composition of the peptides. MSQuant (Gouw et al., 2010; Mortensen et al., 2010) address this problem by identifying peptides before extracting precursor XICs. A problem that is not currently addressed in most packages is the possibility that the isotopic clusters of two labels may overlap. This effect can be corrected for by *in silico* deconvolution (Cappadona et al., 2011).

Normalization of ratio data

An (exaggerated) example of ratio data is shown in Figure 4, where measurements emulated are by dividing random numbers from a uniform distribution. Clearly, ratio data is inherently asymmetrical, with half the data concentrated between 0 and 1, and the other half occupying the numbers between 1 and infinity. Therefore, the first step in the normalization is the transformation into log2 values. Values with a log2-ratio of 0 represent equal expression, values of -1 and 1 represent 2x down- and 2x upregulated values respectively (Figure 4b).
In most experiments only a relatively low subset of proteins is expected to be regulated, so the majority of the log-ratios are expected to be centered on zero. If an experimental bias occurs, e.g. by pipetting error, the mean or median of the log-ratios may be different from zero. In that case, simply subtracting the mean or median log-ratio from all data will compensate for the bias. Moreover, if the log-ratios, M, are plotted against the log of the intensity values, A (an "MA plot", Fig. 5a), often non-linear trends are observed. These trends may be removed by performing a moving average or locally weighted regression (loess) algorithm along the A axis, and subtracting the fitted line. A graphical software package implementing these methods is DanteR (Polpitiya et al., 2008).
A common observation is, that some proteins will not be measured in one of the channels. A common solution is to replace the zero intensity values with very low numbers, sometimes related to the noise level. Un-

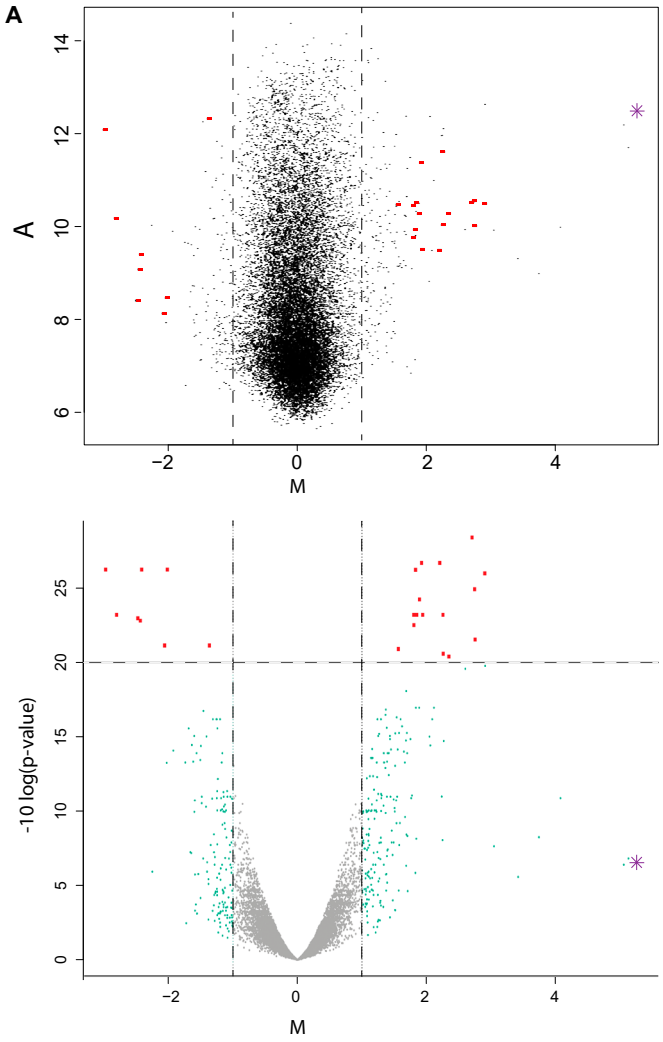


Figure 5: two representations of ratio data. **a.** an A (log10 of the intensity) versus M (log 2 of the ratio) plot. Red dots indicate significantly regulated plots, determined by a t-test. **b.** A volcano plot, with the same M values on the horizontal axis as in a., and on the vertical axis the -10 log transformed p-value. The purple star shows an apparent extremely regulated data point, that is not significantly upregulated.

fortunately, the resulting ratios are usually large, disrupting normalization and significance testing. Therefore, it may be preferable to remove these measurements before further processing and treat them as special cases.

Significance analysis

An important aspect after measuring and normalizing peptide and protein ratios is the determination of regulated proteins. Traditionally, a simple fold-cutoff would be performed, reporting all proteins with a fold change

higher than a pre-set threshold, of for instance two-fold. Unfortunately, this method gives no information about the quality of the results, i.e. the number of false positives in the reported protein list. Applying statistical analyses enables control over the quality of the results. There are different analysis tools available, which can be applied in different situations.
If enough replicate experiments are done, the variability of each protein can be estimated from the replicate values and used in a t-test. Increasing the number of replicates in an experiment increases the sensitivity. The type of replicates that need to be performed depend on the experimental conditions, while the replicate numbers may be estimated by performing a power analysis (Karp and Lilley, 2007; Karp et al., 2005) This test shows how confidently (or how persistently) the ratio is different from "unchanging", i.e. zero, however low the ratio may be. Therefore, a second fold change threshold is often applied to account for biological function. This is represented in a volcano plot that shows the relation between ratio and p-value, together with the chosen thresholds (Fig. 5b).
Time and financial constraints often limit the number of replicates that can be performed, for a single protein usually only one measurement is done. A technique used to estimate the variability is to assume most proteins are not influenced by the experimental conditions, so using the population variance is used to estimate the null distribution, and outliers are assumed to be differentially regulated. In Perseus, the statistical analysis tool accompanying MaxQuant, there are two ways of assigning differential expression based on the overall variation, named Significance A and Significance B (Cox and Mann, 2008). Significance A will calculate the variance over the whole population and perform one-sample t-tests, to establish if the log-ratios are significantly different from zero. Because standard deviation tends to vary with measured intensities (as in Fig. 5a), Significance B performs the analyses in intensity bins. It is important to realize that these techniques often yield false positives, since their discriminating power is very low.

Multiple testing correction

When considering a protein for significance based on a statistical test, a so-called false positive rate (usually written as α) is chosen: the proportion of times the test is expected to indicate a significant difference, when there

is no real difference. Traditionally, α is set to 5%, indicating that if a test is performed 100 times, it is expected that 5 of the tests indicate significance purely by chance. Likewise, if a proteomics experiment quantifies 5000 protein ratios, we'd expect 250 significant proteins, even if the experimental condition has no effect.
The Bonferroni correction tries to remove these false positive findings by simply dividing α by the number of tests, so for 5000 proteins that would mean α becomes 0.00001. Since this threshold is so conservative, the number of significant tests decreases dramatically. Instead, controlling the false discovery rate (FDR), i.e. allowing a fraction of the significant hits to be false, has gained popularity because of its greater sensitivity. The Benjamini-Hochberg step-up procedure is applicable for most high throughput experiments (Benjamini and Hochberg, 1995). Similar to *p*-values resulting from statistical tests, denoting the α value that would just pass a measurement as significant, the *q* value denotes the FDR in a multiple testing corrected context (Storey and Tibshirani, 2003).
Although many tools give nice results in an automated fashion, sometimes it is necessary to go into more detail to investigate outliers or correct artifacts by hand. For that have developed StatQuant (Chapter 3), an interactive tool to load in quantification data and perform t-tests with Benjamini-Hochberg correction. By displaying the quantified PSMs for every protein and the possibility to disable any of these PSMs in the calculations, it is possible to have better confidence in the results.

Application of proteomics and bioinformatics on biological problems

The proteomics of the most primitive animal

A key event in the evolution of animals is the invention of true multi-cellularity, where cells differentiate to perform different functions and are dependent on each other. Adaptation have caused speciation to occur for some branches, while other branches have stayed in a relative primitive state. Based on their DNA sequence, morphology and physiology, some species have been hypothesized to represent the most direct lineage to the

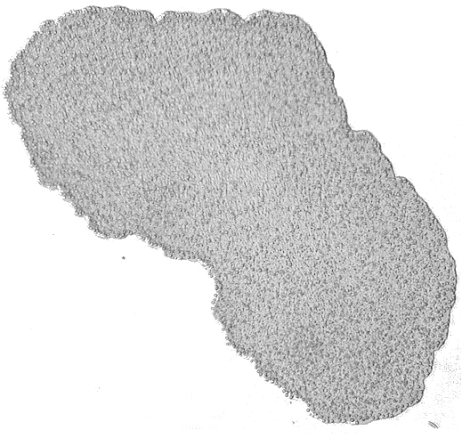


Figure 6 A microscope image of *Trichoplax adhaerens*. The animal consists of a few cell layers, here seen from above. There are no axes of symmetry. Photograph by Oliver Voigt, CC BY-SA 3.0 via Commons, https://commons.wikimedia.org/wiki/File:Trichoplax_mic.jpg#/media/File:Trichoplax_mic.jpg

common ancestor of animals, such as sponges (*Porifera* class *Demospongiae*) (Srivastava et al., 2008, 2010), but more likely the placozoa *Trichoplax adhaerens* (Schierwater et al., 2009). The placozoan genome is by far the smallest of these and has been regarded as the best living surrogate for the hypothetical Cnidaria-Bilateria ancestor genome or even metazoan genome in general (Schierwater and Kuhn, 1998; Schierwater et al., 2009). The placozoan *Trichoplax adhaerens* is morphologically the simplest of all animals, lacking a body axis, basal lamina and extracellular matrix (ECM), and containing only 5 somatic cell types(Guidi et al., 2011). It can be found in tropical and subtropical sea waters and appears as a flat disc of 2-3mm diameter consisting of two epithelial layers with a loose layer of fiber cells in between (Schierwater, 2005). *Trichoplax* reproduces *in vitro* by fission and budding, and although *in vitro* the egg stadium does not develop into an embryonic stage beyond 64-128 cells, there are clear indications for a bisexual reproduction cycle, which left its signature in the DNA (Eitel et al., 2011; Signorovitch et al., 2005). The *Trichoplax* genome contains 11,500 genes and interestingly include important genes characteristic of more complex bilaterian animals such as developmental signaling pathways, neuroendocrine processes, and extracellular matrix proteins (Srivastava et al., 2008). The available genome, however, does not reveal which proteins are expressed, to what level and whether proteins are functionally regulated by posttranslational modifications (PTMs). Using high-resolution mass spectrometry based proteomics we monitor for the first time which *Trichoplax* genes are actually translated and expressed. Moreover, as the functionality of proteins is

to a large extent determined by posttranslational modifications (PTMs), which can only be studied at the protein level, we look into more detail at some important PTMs such as phosphorylation and acetylation. In summary, here we show that studying the proteome of *Trichoplax*, one of the most ancient extant multicellular animals, may provide significant insight into the mechanisms underlying the emergence of metazoan multicellularity.

Annotating genomes from proteins: Proteogenomics

Apart from protein identification and quantification, proteomics is also a meaningful technique in genome annotation. Newly sequenced genomes are subjected to statistical gene models to establish potential gene locations. This is hard both in prokaryotic and eukaryotic organisms for different reasons: Prokaryotic DNA is densely populated to the extent that coding sequences may overlap, whereas eukaryotes have RNA splicing, so the prediction of intron-exon boundaries is a complicating factor. Moreover, alternative splicing is sometimes used as an adaptive mechanism, therefore a statistical model is not going to establish the correct outcomes, unless it would be able to know and integrate many elements of the cellular machinery. A common aid in the annotation of ORFs are Expressed Sequence Tags (ESTs), short pieces of end-sequenced RNA that are traditionally used to help with determining expressed transcripts(Adams et al., 1991). More recently, RNA-seq data and proteomics data is used to improve the gene annotation, in a process called evidence-driven gene annotation(Yandell and Ence, 2012). Early work in this field includes the mapping of *Mycoplasma pneumoniae* peptides, where 81% of the predicted ORFs were detected, and also several new ORFs or ORF extensions were found(Jaffe et al., 2004). Most search engines support 6-frame translation searches, originally introduced in the Sequest engine (Yates et al., 1995). These peptides are then mapped to the genome, so gene annotation tools such MAKER (Cantarel et al., 2008), PASA (Haas et al., 2003, 2011), EVM (Haas et al., 2008) or Gnomon (Thibaud-Nissen et al., 2013) can take them into account either during the gene discovery stage, or by validating the predicted genes. Interestingly, proteogenomics also has a place in the area of meta-analysis where larger numbers of organisms (mostly prokaryotes) are analyzed at once to obtain "community-wide" information of gene and

protein expression (Delmotte et al., 2009; Rodríguez-Valera, 2004). We have applied the proteogenomics approach for the first time on a single organ of a higher organism, i.e. rat (*Rattus norvegicus*) liver samples. We compared two well-characterized strains (BN-*Lx*: Brown Nose and SHR: Spontaneous Hypertensive Rat), at an age where the hypertension is not observed. First, as expected, we observed that protein expression and RNA levels do not correlate very well, but we do observe that the relation between RNA and protein levels is consistent between different strains, hinting at a gene-specific net translation rate. This was also found in a larger data set by the Kuster group, enabling the prediction of protein levels from mRNA levels for different human samples (Wilhelm et al., 2014). Combining expression of both RNA and protein level, we observed a strong downregulation of *Cyp17a1,* a cytochrome implicated with human hypertension, which could be related to a SHR-specific mutation in the promoter of the gene. The rat study clearly shows that the integration of different omics technologies, in this case DNA-sequencing for snp detection, RNA-seq for splice variants and expression differences, and proteomics for the confirmation of the variants and quantification. Overall, with the different technologies providing complementary information on biology, the future of molecular biology most likely is in the tactical combination of different techniques.

My contributions

Several authors contributed to each chapter. I list the work I did for each publication here. I wrote and tested the RockerBox software described in Chapter 2 and wrote the publication. I optimized and partly wrote the StatQuant program described in chapter 3. I performed all data analyses for the comparison of CID and ETD in chapter 4 and wrote the publication. I did most of the work for the phosphorylation database and web front end described in chapter 5. I analyzed sequences in the Trichoplax data in chapter 6 for the presence of kinase motifs, and helped in creating the kinase trees. Lastly, I performed all analysis on the proteomics data, and integrated the proteomics and transcriptomics data in chapter 7.

References

Adams, M., Kelley, J., Gocayne, J., Dubnick, M., Polymeropoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B., Moreno, R., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science (80-.). 252, 1651–1656.

Aye, T.T., Scholten, A., Taouatas, N., Varro, A., Van Veen, T.A.B., Vos, M.A., and Heck, A.J.R. (2010). Proteome-wide protein concentrations in the human heart. Mol. Biosyst. 6, 1917.

Benjamini, Y., and Hochberg, Y.C.N.-M. (96d:62143) (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B 57, 289–300.

Bianco, L., Mead, J.A., and Bessant, C. (2009). Comparison of Novel Decoy Database Designs for Optimizing Protein Identification Searches Using ABRF sPRG2006 Standard MS/MS Data Sets. J. Proteome Res. 8, 1782–1791.

Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S., and Heck, A. J.R.C.N.-0018 (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat. Protoc. 4, 484–494.

Braisted, J.C., Kuntumalla, S., Vogel, C., Marcotte, E.M., Rodrigues, A.R., Wang, R., Huang, S.-T., Ferlanti, E.S., Saeed, A.I., Fleischmann, R.D., et al. (2008). The APEX Quantitative Proteomics Tool: Generating protein quantitation estimates from LC-MS/MS proteomics results. BMC Bioinformatics 9, 529.

Bridges, S.M., Magee, G.B., Wang, N., Williams, W.P., Burgess, S.C., and Nanduri, B. (2007). ProtQuant: a tool for the label-free quantification of MudPIT proteomics data. BMC Bioinformatics 8 Suppl 7, S24.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.

Cappadona, S., Munoz, J., Spee, W.P., Low, T.Y., Mohammed, S., van Breukelen, B., and Heck, A.J. (2011). Deconvolution of overlapping isotopic clusters improves quantification of stable isotope-labeled peptides. J. Proteomics 74, 2204–2209.

Christin, C., Hoefsloot, H.C., Smilde, A.K., Suits, F., Bischoff, R., and Horvatovich, P.L. (2010). Time alignment algorithms based on selected mass traces for complex LC-MS data. J Proteome Res 9, 1483–1495.

Conrads, T.P., Alving, K., Veenstra, T.D., Belov, M.E., Anderson, G.A., Anderson, D.J., Lipton, M.S., Paša-Tolić, L., Udseth, H.R., Chrisler, W.B., et al. (2001). Quantitative Analysis of Bacterial and Mammalian Proteomes Using a Combination of Cysteine Affinity Tags and 15 N-Metabolic Labeling. Anal. Chem. 73, 2132–2139.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics 13, 2513–2526.

Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., von Mering, C., and Vorholt, J.A. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc. Natl. Acad. Sci. U. S. A. 106, 16428–16433.

Eastman, Q. (2009). Space-partitioning speeds up data processing - Journal of Proteome Research (ACS Publications).

Eitel, M., Guidi, L., Hadrys, H., Balsamo, M., and Schierwater, B. (2011). New Insights into Placozoan Sexual Reproduction and Development. PLoS One 6.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989.

Gouw, J.W., and Krijgsveld, J. (2012). MSQuant: a platform for stable isotope-based quantitative proteomics. Methods Mol Biol 893, 511–522.

Gouw, J.W., Krijgsveld, J., and Heck, A.J.R. (2010). Quantitative Proteomics by Metabolic Labeling of Model Organisms. Mol. Cell. Proteomics 9, 11–24.

Griffin, N.M., Yu, J.Y., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J.A., and Schnitzer, J.E. (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat. Biotechnol. 28, 83-U116.

Grossmann, J., Roschitzki, B., Panse, C., Fortes, C., Barkow-Oesterreicher, S., Rutishauser, D., and Schlapbach, R. (2010). Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. J Proteomics 73, 1740–1746.

Guidi, L., Eitel, M., Cesarini, E., Schierwater, B., and Balsamo, M. (2011). Ultrastructural analyses support different morphological lineages in the phylum placozoa Grell, 1971. J. Morphol. 272, 371–378.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999.

Haas, B.J., Delcher, A.L., Mount S.M., Wortman, J.R., Smith, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654–5666.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7.

Haas, B.J., Zeng, Q., Pearson, M.D., Cuomo, C. a, and Wortman, J.R. (2011). Approaches to Fungal Genome Annotation. Mycology 2, 118–141.

Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc. Natl. Acad. Sci. 90, 5011–5015.

Hsu, J.-L., Huang, S.-Y., Chow, N.-H., and Chen, S.-H.C.N.-0117 (2003). Stable-isotope dimethyl labeling for quantitative proteomics. Anal. Chem. 75, 6843.

Huang, S.Y., Tsai, M.L., Wu, C.J., Hsu, J.L., Ho, S.H., and Chen, S.H. (2006). Quantitation of protein phosphorylation in pregnant rat uteri using stable isotope dimethyl labeling coupled with IMAC. Proteomics 6, 1722–1734.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. Mol. Cell. Proteomics 4, 1265–1272.

Jaffe, J.D., Berg, H.C., and Church, G.M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4, 59.

James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993). Protein identification by mass profile fingerprinting. Biochem. Biophys. Res. Commun. 195, 58–64.

Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J. Proteome Res. 7, 29–34.

Karp, N. a, and Lilley, K.S. (2007). Design and Analysis Issues in Quantitative Proteomics Studies. Proteomics 7, 42–50.

Karp, N.A., Spencer, M., Lindsay, H., O'Dell, K., and KS, L. (2005). Impact of Replicate Types on Proteomic Expression Analysis.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R.C.N.-1136 (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383.

Khan, Z., Bloom, J.S., Garcia, B.A., Singh, M., and Kruglyak, L. (2009). Protein quantification across hundreds of experimental conditions.

Kondrat, R.W., McClusky, G.A., and Cooks, R.G. (1978). Multiple reaction monitoring in mass spectrometry/mass spectrometry for direct analysis of complex mixtures. Anal. Chem. 50, 2017–2021.

Lange, V., Picotti, P., Domon, B., and Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. Mol. Syst. Biol. 4, 222.

Li, X.J., Zhang, H., Ranish, J.A., and Aebersold, R. (2003). Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem 75, 6648–6657.

Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., et al. (2002). Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. Proc. Natl. Acad. Sci. U. S. A. 99, 11049.

Liu, H., Sadygov, R.G., and Yates, J.R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2006). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25, 117–124.

MacKay, V.L., Li, X., Flory, M.R., Turcott, E., Law, G.L., Serikawa, K.A., Xu, X.L., Lee, H., Goodlett, D.R., Aebersold, R., et al. (2004). Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. Mol. Cell. Proteomics 3, 478–489.

Mann, M., Højrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

McAlister, G.C., Huttlin, E.L., Haas, W., Ting, L., Jedrychowski, M.P., Rogers, J.C., Kuhn, K., Pike, I., Grothe, R.A., Blethrow, J.D., et al. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. Anal. Chem. 84, 7469–7478.

Mikesh, L.M., Ueberheide, B., Chi, A., Coon, J.J., Syka, J.E.P., Shabanowitz, J., and Hunt, D.F. (2006). The utility of ETD mass spectrometry in proteomic analysis. Biochim. Biophys. Acta 1764, 1811–1822.

Mortensen, P., Gouw, J.W., Olsen, J. V, Ong, S.-E., Rigbolt, K.T.G., Bunkenborg, J., Cox, J., Foster, L.J., Heck, A.J.R., Blagoev, B., et al. (2010). MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. J. Proteome Res. 9, 393–403.

Mueller, L.N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., and Müller, M.C.N.-0172 (2007). SuperHirn a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics 7, 3470.

Oda, Y., Huang, K., Cross, F.R., Cowburn, D., and Chait, B.T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. Proc. Natl. Acad. Sci. U. S. A. 96, 6591–6596.

Old, W.M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K.G., Mendoza, A., Sevinsky, J.R., Resing, K.A., and Ahn, N.G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol. Cell. Proteomics 4, 1487–1502.

Olsen, J. V, Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods 4, 709–712.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics 1, 376–386.

Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. Curr. Biol. 3, 327–332.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567.

Polpitiya, A.D., Qian, W.-J., Jaitly, N., Petyuk, V.A., Adkins, J.N., Camp, D.G., Anderson, G.A., and Smith, R.D. (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics 24, 1556–1558.

Purves, R.W., Gabryelski, W., and Li, L. (1998). Investigation of the quantitative capabilities of an electrospray ionization ion trap/linear time-of-flight mass spectrometer. Rapid Commun. Mass Spectrom. 12, 695–700.

Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-Scale Proteomic Analysis of the Human Spliceosome. Genome Res. 12, 1231–1245.

Rodríguez-Valera, F. (2004). Environmental genomics, the big picture? FEMS Microbiol. Lett. 231, 153–158.

Ross, P.L., Huang, Y.N.L.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., et al. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol. Cell. Proteomics 3, 1154–1169.

Schierwater, B. (2005). My favorite animal,Trichoplax adhaerens. BioEssays 27, 1294–1302.

Schierwater, B., and Kuhn, K. (1998). Homology of Hox genes and the zootype concept in early metazoan evolution. Mol. Phylogenet. Evol. 9, 375–381.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Dellaporta, S.L., Kolokotronis, S.-O., and DeSalle, R. (2009). Concatenated Analysis Sheds Light on Early Metazoan Evolution and Fuels a Modern "Urmetazoon" Hypothesis. PLoS Biol. 7, e1000020.

Schulze, W.X., and Usadel, B. (2010). Quantitation in Mass-Spectrometry-Based Proteomics. Annu. Rev. Plant Biol. Vol 61 61, 491–516.

Schwanhausser, B., Gossen, M., Dittmar, G., and Selbach, M. (2009). Global analysis of cellular protein translation by pulsed SILAC. Proteomics 9, 205–209.

Signorovitch, A.Y., Dellaporta, S.L., and Buss, L.W. (2005). Molecular signatures for sex in the Placozoa. Proc. Natl. Acad. Sci. U. S. A. 102, 15518–15522.

Silva, J.C., Gorenstein, M. V, Li, G.-Z., Vissers, J.P.C., and Geromanos, S.J. (2006). Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol. Cell. Proteomics 5, 144–156.

Smith, L.M., and Kelleher, N.L. (2013). Proteoform: a single term describing protein complexity. Nat. Methods 10, 186–187.

Smith, R., Ventura, D., and Prince, J.T. (2015). LC-MS alignment in theory and practice: a comprehensive algorithmic review. Brief. Bioinform. 16, 104–117.

Spivak, M., Weston, J., Bottou, L., Käll, L., and Noble, W.S. (2009). Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. J. Proteome Res. 8, 3737–3745.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., et al. (2008). The Trichoplax genome and the nature of placozoans. Nature 454, 955.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., et al. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466, 720–726.

Steen, H., and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 5, 699–711.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genome-wide studies. In Proc Natl Acad Sci U S A, (Department of Biostatistics, University of Washington, Seattle, WA 98195, USA. jstorey@u.washington.edu), pp. 9440–9445.

Swaney, D.L., McAlister, G.C., and Coon, J.J. (2008). Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat. Methods 5, 959–964.

Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 101, 9528–9533.

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M., and Kitts, P. (2013). Eukaryotic Genome Annotation Pipeline (National Center for Biotechnology Information (US)).

Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Hamon, C., Schäfer, J., Kuhn, K., et al. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. 75, 1895–1904.

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., Penalva, L.O., et al. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. 6, 400.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Watson, J.D., and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738.

Wells, J.M., and McLuckey, S.A. (2005). Collision-induced dissociation (CID) of peptides and proteins. Methods Enzymol. 402, 148–185.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587.

Wilkins, M.R., Pasquali, C., Appel, R.D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J.X., Gooley, A.A., Hughes, G., Humphery-Smith, I., et al. (1996). From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. Bio/Technology 14, 61–65.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342.

Yao, X., Freas, A., Ramirez, J., Demirev, P.A., and Fenselau, C. (2001). Proteolytic 18 O Labeling for Comparative Proteomics: Model Studies with Two Serotypes of Adenovirus. Anal. Chem. 73, 2836–2842.

Yates, J.R., Speicher, S., Griffin, P.R., and Hunkapiller, T. (1993). Peptide Mass Maps: A Highly Informative Approach to Protein Identification. Anal. Biochem. 214, 397–408.

Yates, J.R., Eng, J.K., and McCormack, A.L. (1995). Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. 67, 3202–3210.

Zybailov, B., Mosley, A.L., Sardiu, M.E., Coleman, M.K., Florens, L., and Washburn, M.P. (2006). Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J. Proteome Res. 5, 2339–2347.

2. RockerBox: analysis and filtering of massive proteomics search results

Reprinted (adapted) with permission from van den Toorn, H.W.P., Muñoz, J., Mohammed, S., Raijmakers, R., and Heck, A.J.R. (2011). RockerBox: analysis and filtering of massive proteomics search results. J. Proteome Res. 10, 1420–1424. Copyright (2011) American Chemical Society.

A major problem in the analysis of mass spectrometry-based proteomics data is the vast growth of data volume, caused by improvements in sequencing speed of mass spectrometers. This growth affects analysis times and storage requirements so severely that many analysis tools are no longer able to cope with the increased file sizes. We present a tool, RockerBox, to address size problems for search results obtained from the widely used Mascot search engine. RockerBox allows for a fast evaluation of large result files by means of a number of commonly accepted metrics which can often be viewed through charts. Moreover, result files can be filtered without altering their informative content, based on a number of FDR calculation methods. File can be reduced dramatically, often to a tenth of their original size, thus relaxing the need for storage and computation power, and boosting analysis of current and future proteomics experiments.

Introduction

Mass spectrometry- (MS-) based proteomics has emerged as a powerful technique for high-throughput characterization of proteins in biological samples. Technical advances in peptide separation, precision, sensitivity and duty cycle (Cox and Mann, 2009; John et al., 2009; Kim et al., 2010; Olsen et al., 2009) have come at the price of increased data volume in terms of spectra and identifications. Consequently, the computational analysis of large data dependent mass spectrometric experiments has become increasingly difficult. Although database search engines such as Mascot (Perkins et al., 1999), OMSSA (Geer et al., 2004), Sequest (Eng et al., 1994) and others are able to handle the increased number of spectra, many downstream analysis tools are currently restricted by storage capacity and by the amount of available system memory.

A large proportion of MS/MS spectra have a low signal-to-noise ratio or correspond to fragmentation of contaminants or peptides not present in the protein database utilized by the search engine. Most of the filtering methods proposed in the past to circumvent these restrictions, are try to remove MS/MS spectra based on data quality assessment by applying prior knowledge on peptide fragmentation (Junqueira et al., 2008; Mujezinovic et al., 2010; Salmi et al., 2009). An alternative strategy is to filter

peptide-spectrum matches (PSMs) based on search engine results.
To discriminate between low quality and high quality PSMs, numerical thresholds have to be established for metrics such as score and mass delta between the observed precursor mass and the mass of the corresponding PSM. Several methods exist to automatically determine these thresholds (Brosch et al., 2009; Elias and Gygi, 2007; Elias et al., 2004, 2005; Higgs et al., 2007; Huttlin et al., 2007; Joo et al., 2010; Käll et al., 2007, 2008a; Keller et al., 2002; Moore et al., 2002). Score thresholds can be determined with the False Discovery Rate (FDR, Supplementary Table ST1), a commonly accepted measure for the quality of a MS proteomics results (Elias and Gygi, 2007; Moore et al., 2002). A particular FDR directly affects the sensitivity and specificity of an experiment: a higher FDR threshold will allow for more PSMs to be accepted at the cost of more false positives (see supplementary Table ST1). To estimate the FDR, spectra are matched against a database containing known sequences ("target" database) and an equivalent number of randomized or reversed sequences ("decoy" database) (Elias and Gygi, 2007; Moore et al., 2002). The FDR at a certain score threshold can be estimated by calculating the ratio of "decoy" PSMs (i.e. false positives), divided by the total number of PSMs (Elias et al., 2005). There are two common ways to perform a "decoy" search: non-competitive and competitive. In the non-competitive method, a standard option in Mascot, all spectra are searched separately against a "target" database and a "decoy" database. In the competitive method, spectra are searched against a so-called concatenated database, in which the "target" and "decoy" sequences are combined. New strategies have been proposed that make use of machine learning techniques to classify PSMs based on several features (Käll et al., 2007, 2008b). Percolator is an example of such strategy, based on the support-vector machine algorithm (Käll et al., 2007).
Here, we present RockerBox, an application designed to process and analyze .dat files . The toolkit provides charts to quickly assess data and filtering methods to drastically reduce file size. RockerBox allows manual entry of cutoff values for Mascot ions score and mass window. Moreover, it allows the researcher to select peptide modifications as a filtering criterion, to focus on a particular subset of the search result. FDR values can also be used to determine cutoff values. RockerBox will calculate FDR based on either the concatenated or separate database decoy search methods, or

using the Percolator algorithm (Käll et al., 2007).
All the intricate internal references (Grosse-Coosmann et al., 2005) are maintained in the filtered .dat files, so they are fully compatible with downstream processing tools, including Mascot's own .html formatting script. In this article, we present the results of filtering a large data set to illustrate how RockerBox removes 'superfluous' matches, while retaining good-quality search results and spectra.

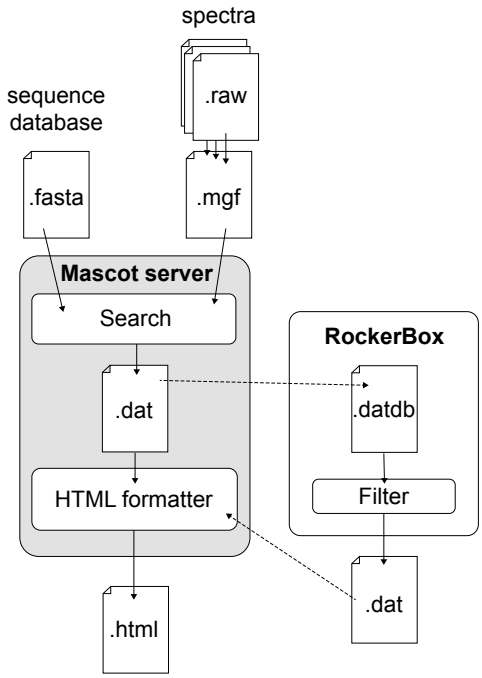


Figure 1. General workflow outlining the mass-spectrometric analysis of a typical experiment with the Mascot search engine. Raw files representing separate pre-fractionation samples are converted to peak lists and combined into a single .mgf file. Subsequently, all spectra in the mgf file are matched to proteolytic peptides that are predicted from the sequence database chosen by the user. The output of this process, PSMs, are stored in .dat files on the Mascot server and formatted into an html file for viewing and further analysis. RockerBox is used as an external tool for the Mascot search engine to view and filter .dat files. The new .dat files can then be used to create HTML files for further analyses.

Methods

RockerBox can be integrated in the Mascot search workflow as shown in Figure 1. It is a desktop utility written in the Java programming language (version 1.6). Development was carried out with the NetBeans Integrated Development Environment (IDE) (version 6.9, Oracle, www.netbeans.org). All operations are highly optimized for memory usage, using stream-based file access throughout. A disk-based caching strategy is used for charts and statistics to keep memory requirements to a minimum.

Data analysis

To quickly assess the quality of the search results from Mascot, RockerBox enables the crea-

tion of ROC curves, FDR charts (FDR value as a function of score), score as a function of mass delta and mass delta as a function of scan number (Supplementary Figure 1b). Charts can be saved in the .png bitmap file format and the vector file formats .svg and .pdf. Furthermore, charts can be saved as .fcht files that can be opened inside RockerBox for later viewing.

For quick reference, when .dat files are opened in RockerBox, a summary of the information including search parameters, database size and number of queries is displayed for these files. To allow further analysis in spreadsheets or statistical packages, RockerBox features extensive export capabilities to tab-delimited files with information of all the first-ranking PSMs from a .dat file. Moreover, in-depth feature tables are exported that show delta scores and PSM matching statistics. (all features are summarized in Supplementary table ST2).

Furthermore, RockerBox can extract .mgf files for matched spectra. This allows the researcher to perform a second search with different search parameters. To further facilitate that, all search parameters used to create a .dat file can be exported as a .par file, which can be read and edited in the Mascot Daemon tool (Matrix Science) to submit the new search.

Filtering methods

RockerBox features three filtering methods: manual filter, FDR-based filter and Percolator-based filter (Supplementary Figure 1c and 2).

Manual filtering parameters include a cutoff on Mascot ions score, a range for the mass error of the precursor ion and a selection of post-translational modifications (PTMs) allowed for each peptide (Supplementary Figure 1c-I).

In FDR-based filtering, the FDR for every score cutoff can be estimated either using the "automatic decoy" created by Mascot or by counting the number of "decoy" hits from a concatenated sequence database. Both FDR calculations and filtering can optionally be limited to a mass error window (Supplementary Figure 1c-II).

The Percolator-based filtering method provides the means to use multiple features for filtering Mascot results. These features (summarized in supplementary Table ST2) are extracted and calculated from the .dat file for each PSM (Supplementary Figure 1c-III) and used as input for the Perco-

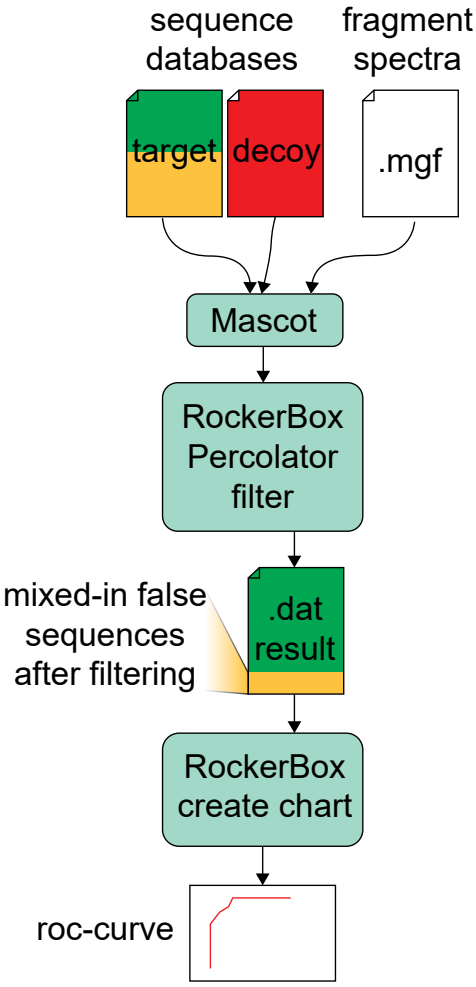


Figure 2. A graphical depiction of the method used to estimate the performance of Percolator filtering in RockerBox. The Mascot search included the 'automatic decoy' method, where a search against a scrambled decoy database (red) alongside the standard 'target' database is performed automatically by the Mascot search engine. Here the 'target' also contained 'false' or non-existant sequences, to keep track of the filtering performance. Features from both database searches were extracted and used to train the Percolator algorithm, which assigned q-values to the 'target' PSMs. A new .dat file was created by filtering the input 'target' PSMs at a q-value threshold of 0.01. Within the 'target' PSMs, database hits were found from the mixed-in false sequences. This allows for FDR estimation on the resulting data.

lator executable. Each PSM feature-set is annotated as either "decoy" or "target" based on one of the "target-decoy" search methods. When using a concatenated decoy strategy, features are calculated as if the PSMs were obtained using a non-competitive decoy strategy, to prevent undersampling of decoy PSMs. Percolator assigns a q-value to every PSM, which is then used to filter the PSMs at a selected FDR level. During the development of RockerBox we have used Percolator version 1.14. Similar to the MascotPercolator application(Brosch et al., 2009), we included the option to substitute the Mascot ions scores in the filtered .dat file with a score based on Posterior Error Probabilities (PEP), or a score based on

q-values as estimated by the Percolator application.
In some cases, e.g. a 2D LC pre-fractionation strategy, the researcher concatenates peak lists from several raw files. If the input peak list (.mgf file) used for searching contains information about the original raw file then both FDR-based and Percolator-based filter modes can be run on the data from separate raw files. Several software packages that prepare peak list files from raw data files such as MassHunter (Agilent Technologies), Quant (MaxQuant), or ProteomeDiscoverer (Thermo Fischer Scientific) supply this information and are automatically recognized by RockerBox.

The .datdb file format

To reduce processing times on input/output operations and to minimize disk usage, we introduce the .datdb file format, a form of the .dat file based on a the sqlite one-file database (Figure 1). It is possible to extract .datdb

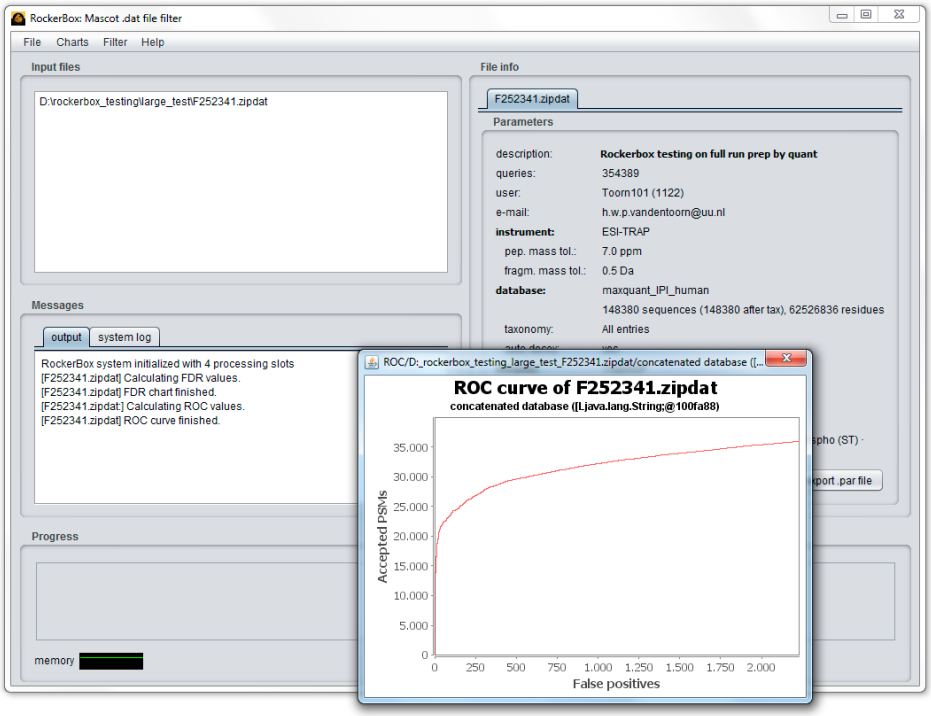


Figure 3. A screenshot of the RockerBox application, showing file information and a ROC curve.

files to obtain the original .dat files that can be used on the Mascot server or other downstream tools. Therefore, the .datdb file format is well equipped for data archiving and exchange.

Performance testing of the RockerBox filtering methods

The dataset used for testing the performance of the available have already been presented by Gauci *et al.* (Gauci et al., 2009). The dataset consists of a HEK293 cell lysate, which was digested with trypsin and separated by SCX and measured on a nano LC coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fischer Scientific) yielding 354,389 fragmentation spectra. Peak list (.mgf) files were prepared with the Quant application from MaxQuant (version 1.0.13.8). Database searching was performed with Mascot (version 2.2.04) with a peptide tolerance of 7 ppm and a fragment tolerance of 0.5 Da. The database used was IPI Human (v3.52) concatenated with a "decoy" database by the SequenceReverser tool from MaxQuant, with enzyme trypsin allowing no missed cleavages. Variable modifications were acetylation of the protein N-termini, oxidation of methionines and phosphorylation of serine, threonine and tyrosine residues. Carbamidomethylation of cysteines was set as fixed modification. In addition, we enabled the Mascot automatic "decoy" generation method, which allowed us to train the Percolator algorithm on a fully randomized database as decoy, while the PSMs presented to the algorithm as "target" contained reversed sequences from the concatenated "decoy" database (Figure 2).

Availability

Rockerbox executables and source code can be obtained from https:/www.hecklab.com/software/rockerbox.

Results and discussion

To address file size-related issues in data processing, we implemented a strategy to filter .dat files by removing low-quality data. As shown in Figure 1, the RockerBox application conveniently hooks into a typical proteomic workflow. A screenshot of the application (Figure 3) shows two of the available analysis techniques: a quick overview of the .dat file that is

opened and a ROC curve generated from the PSM data in the file. Smaller Mascot .html files can then be generated by the Mascot .html formatter for use in further downstream analyses.

We demonstrate the performance of RockerBox by filtering the Mascot search results of a data set of a MudPIT experiment on a HEK293 cell extract.

Figure 4 shows an overview of the unfiltered .dat file in the form of a chart with the score as a function of the difference between the measured precursor mass and the mass of the matched peptide (mass delta) of both "target" and "decoy" PSMs. The results are based on the data from the concatenated search database, disregarding the automatic "decoy" search generated by Mascot. "decoy" PSMs display a uniform score distribution across the mass delta range (lower scatterplot). The same distribution

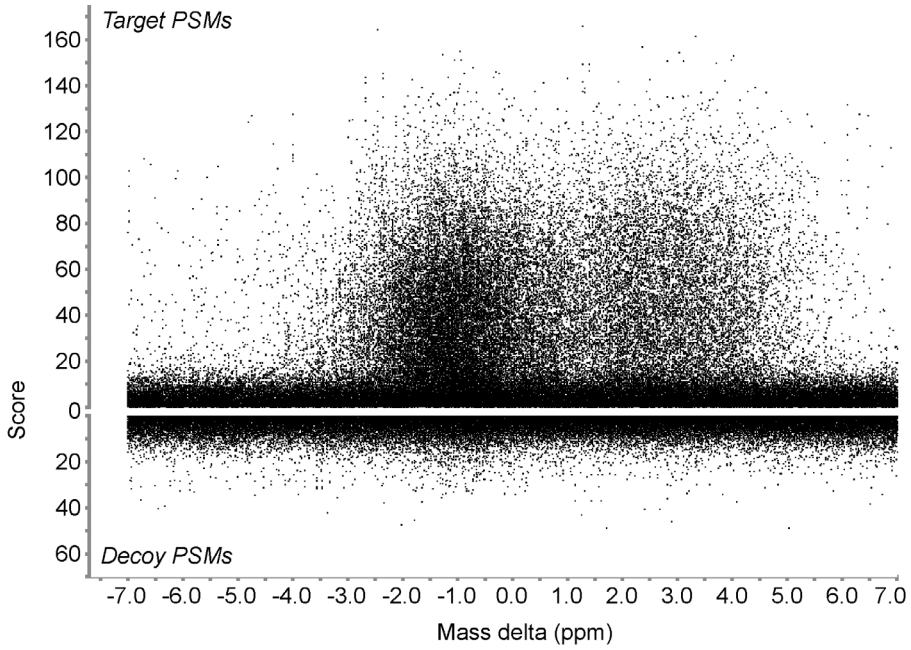


Figure 4. Analytical chart of the data used for analysis of RockerBox performance. The Mascot ions score is plotted against the relative mass difference between the matched peptide and the measured precursor mass (mass delta, ppm). The upper chart represent PSMs that match a real sequence (target PSMs), the lower part represent PSMs that match a reversed sequence (Decoy PSMs). For comparison, the Score axis for the "decoy" PSMs is mirrored with respect to the "target" PSMs. While "decoy" PSMs display a uniform distribution, "target" PSMs show a clear additional cloud of high scoring PSMs around 0ppm.

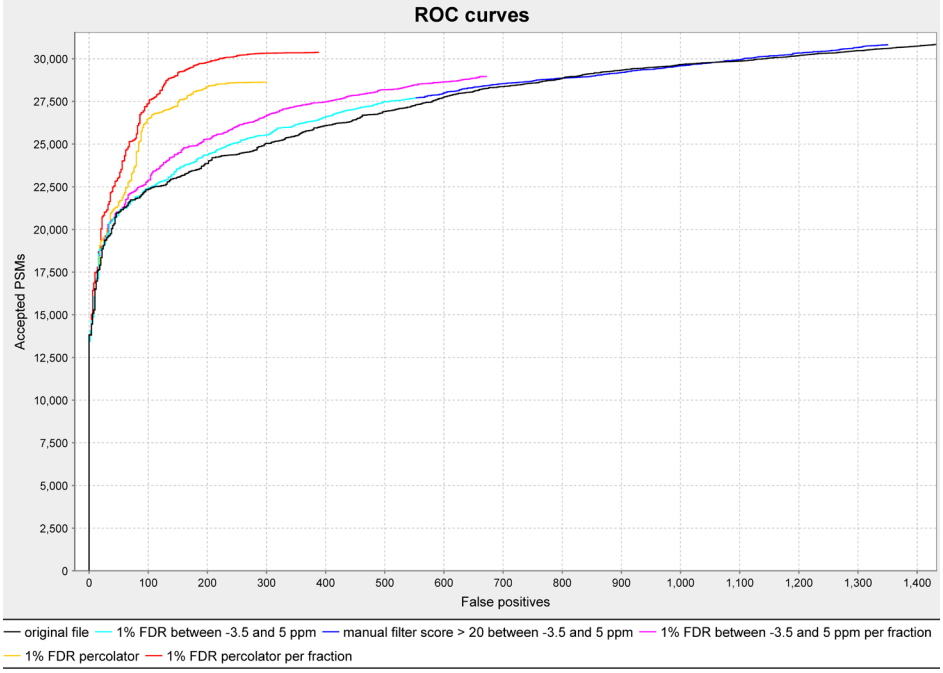


Figure 5. ROC curves showing the accuracy of the different described filtering methods. The black line shows the accuracy of the Mascot ions score alone in separating false and true positives in the unfiltered file. The blue curve shows the ROC curve after applying a manual filter on Mascot ions score and mass window (FDR 1.4%). The cyan line shows the effect of an automatic cut-off score based on a 1% FDR value within the mass window of -3.5 to 5.0 ppm. The purple line shows the ROC curve for the automatic FDR filter within the mass window, where the FDR is calculated for all separate SCX fractions. The yellow line shows the result of the calculations performed by the Percolator algorithm over the whole file. The red line shows the results of filtering based on the Percolator algorithm with a "target" FDR of 1% for each fraction separately.

can be found among the "target" PSMs (top scatterplot), although there is a marked concentration of higher scoring PSMs clustered around the center of the mass delta values. These findings are in accordance with the data of earlier work(Everley et al., 2006)[26] and indicate that mass delta values may be used as a discriminator of PSM quality alongside Mascot ions score values. The "target" hits show a dip in the center of the distribution, indicating some of the fractions that were included in this search displayed a shift in the MS calibration during their mass spectrometry analysis.
To compare the accuracy of the different filtering methods, we con-

structed ROC curves for datasets resulting from all filtering methods. As shown in Table 1, both size and number of spectra are reduced considerably after filtering for all methods. A ROC curve for original file could be drawn (Figure 5, black line) using the "target" and "decoy" PSMs at different score cutoffs, showing the accuracy of Mascot ions scores for discrimination between true positives and false positives. Based on the plot shown in Figure 4, we estimated cutoff criteria as a mass delta window of -3.5 to 5.0 ppm and a Mascot ions score cutoff of 20. We entered these values into the 'manual filter' dialog of RockerBox, leaving mostly the high-scoring "target" PSMs while the uniformly distributed low-scoring PSMs are no longer present (supplementary Figure 3b). The ROC curve shows a slight increase in accuracy (Figure 5 blue curve), which is caused by the promotion of several lower-ranking PSMs to the first rank since they fall within the score and mass window criteria, whereas the original first-ranking PSM did not. Importantly, filtering caused the result file size to be an order of magnitude smaller than the original file (69 Mb vs. 897 Mb, Table 1). By estimating the FDR of the manually filtered file, using the target-decoy method of Elias *et al.* (Elias et al., 2005), we obtain a value of 1.4% which is above a desired FDR of 1%.

A further method, which estimates the FDR rate for all score cutoff values allowing a "target" FDR to be set as a parameter, leaving the mass window as a manual parameter ("FDR based filter"). A "target" FDR of 1% resulted in a ions score cutoff of 21.77 (filtering result shown in supplementary Figure 3c). The cyan ROC curve of the filtered file (Figure 5) follows the manually filtered curve (blue), but ends at lower Accepted and Target PSMs because of the more stringent score cutoff to obtain a FDR of 1%.

In certain experimental setups, such as a 2D LC based analysis, the search result is obtained from the combination of several raw files. Raw files containing relatively low quality spectra (e.g. SCX fractions containing many highly charged peptides (Toorn et al., 2008)) contribute a high number of "decoy" PSMs, possibly increasing the score cutoff and reducing the amount of accepted PSMs. Calculating FDR values from combined results disregards such information. For this reason, RockerBox offers the option to calculate the cutoff score separately for each raw file. The purple ROC curve in Figure 5 shows the accuracy using this method for an FDR of 1% (Table 1, supplementary Figure 3d).

The two methods outlined so far only rely on the control of score values as a cutoff, with the addition of a manually chosen mass window to increase the discrimination accuracy. There are, however, more features of peptide-spectrum matches that may contribute to a better partitioning between false positive and true positive PSMs. Examples of features that are not automatically taken into account for FDR calculations are the mass delta between observed and theoretical peptide and the difference of score between the first and second ranking PSM for a given spectrum. The Percolator application (Käll et al., 2007), although originally written to re-rank Sequest search engine results, may also be used to discriminate on a generic list of features. Percolator uses a Support-Vector Machine (SVM) based algorithm to classify the data based on the features from "decoy" and "target" PSMs. RockerBox can utilize the output of Percolator to filter and optionally re-score Mascot search results. Using Percolator-based filtering results in a higher accuracy compared to the other methods presented here (Figure 5, purple line) and a high number of accepted PSMs (Table 1, supplementary Figure 3e) at an FDR level of 1%. This filtering method yielded a slightly lower number of remaining spectra than the method calculating 1% FDR for each raw input file, indicating that FDR calculation methods greatly benefit from restriction to input raw files. Therefore, we included the possibility to apply the Percolator algorithm on separate raw files as well. This method yielded the highest number of accepted PSMs of all filtering methods tested here (Table 1, supplementary Figure 3f) alongside with the highest accuracy (Figure 5, red line).

Conclusions

RockerBox greatly benefits the analysis of large data sets by offering global charts, file conversion, export and filtering methods. Moreover, result file sizes can be reduced considerably using any of the available filtering methods, the choice of which should be made by the researcher based on the experiment and personal preference. Therefore, a significant reduction in storage requirements and processing time enables analysis of complex samples even with modest computer hardware requirements.

Acknowledgements

We express our gratitude to Sharon Gauci and coworkers for making their data available for the performance testing, and all coworkers at the Heck lab for testing RockerBox. We thank Salvatore Cappadona and Danny Navarro for their comments and discussion on the manuscript. This research has been supported with a grant from the Netherlands Bioinformatics Institute and the Netherlands Proteomics Centre.

References

Brosch, M., Yu, L., Hubbard, T., and Choudhary, J. (2009). Accurate and sensitive peptide identification with Mascot Percolator. J. Proteome Res. *8*, 3176–3181.

Cox, J., and Mann, M.(2009). Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. J. Am. Soc. Mass Spectrom. *20*, 1477.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207.

Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., and Gygi, S.P.C.N.-0155 (2004). Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotech *22*, 214.

Elias, J.E., Haas, W., Faherty, B.K., and Gygi, S.P.C.N.-0310 (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Meth *2*, 667.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976–989.

Everley, P.A., Bakalarski, C.E., Elias, J.E., Waghorne, C.G., Beausoleil, S.A., Gerber, S.A., Faherty, B.K., Zetter, B.R., and Gygi, S.P. (2006). Enhanced Analysis of Metastatic Prostate Cancer Using Stable Isotopes and High Mass Accuracy Instrumentation. J. Proteome Res. *5*, 1224.

Gauci, S., Helbig, A.O., Slijper, M., Krijgsveld, J., Heck, A.J.R., and Mohammed, S. (2009). Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. Anal. Chem. *81*, 4493–4501.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004). Open mass spectrometry search algorithm. J. Proteome Res. *3*, 958.

Grosse-Coosmann, F., Boehm, A.M., and Sickmann, A.C(2005). Efficient analysis and extraction of MS/MS result data from Mascot™ result files. BMC Bioinformatics *6*, 290.

Higgs, R.E., Knierman, M.D., Bonner Freeman, A., Gelbert, L.M., Patil, S.T., and Hale, J.E. (2007). Estimating the Statistical Significance of Peptide Identifications from Shotgun Proteomics Experiments. J. Proteome Res. *6*, 1758.

Huttlin, E.L., Hegeman, A.D., Harms, A.C., and Sussman, M.R.C.N.-0037 (2007). Prediction of Error Associated with False-Positive Rate Determination for Peptide Identification in Large-Scale Proteomics Experiments Using a Combined Reverse and Forward Peptide Sequence Database Strategy. J. Proteome Res. *6*, 392.

John, R.Y., Cristian, I.R., Aleksey, N., Yates, J.R., Ruse, C.I., and Nakorchevsky, A. (2009). Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. Annu. Rev. Biomed. Eng. *11*, 49.

Joo, J.W.J., Na, S., Baek, J.-H., Lee, C., and Paek, E. (2010). Target-Decoy with Mass Binning: A Simple and Effective Validation Method for Shotgun Proteomics Using High Resolution Mass Spectrometry. J. Proteome Res. *9*, 1150.

Junqueira, M., Spirin, V., Santana Balbuena, T., Waridel, P., Surendranath, V., Kryukov, G., Adzhubei, I., Thomas, H., Sunyaev, S., and Shevchenko, A. (2008). Separating the wheat from the chaff: unbiased filtering of background tandem mass spectra improves protein identification. J. Proteome Res. *7*, 3382.

Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925.

Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008a). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J. Proteome Res. *7*, 29–34.

Käll, L., Storey, J.D., and Noble, W.S. (2008b). Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics *24*, i42.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R.(2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. *74*, 5383.

Kim, M.-S., Kandasamy, K., Chaerkady, R., and Pandey, A. (2010). Assessment of resolution parameters for CID-based shotgun proteomic experiments on the LTQ-Orbitrap mass spectrometer. J. Am. Soc. Mass Spectrom. *21*, 1606.

Moore, R.E., Young, M.K., and Lee, T.D. (2002). Qscore: an algorithm for evaluating SEQUEST database search results. J. Am. Soc. Mass Spectrom. *13*, 378.

Mujezinovic, N., Schneider, G., Wildpaner, M., Mechtler, K., and Eisenhaber, F. (2010). Reducing the haystack to find the needle: improved protein identification after fast elimination of non-interpretable peptide MS/MS spectra and noise reduction. BMC Genomics *11 Suppl 1*.

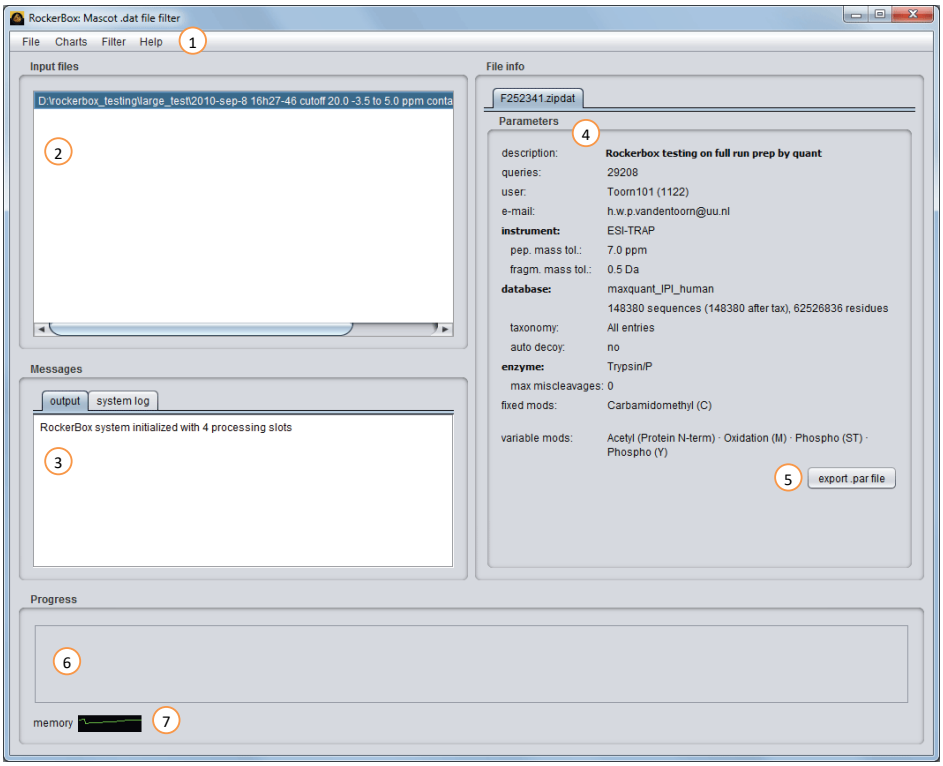
Olsen, J. V, Schwartz, J.C., Griep-Raming, J., Nielsen, M.L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., et al. (2009). A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol. Cell. Proteomics MCP *8*, 2759.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551–3567.

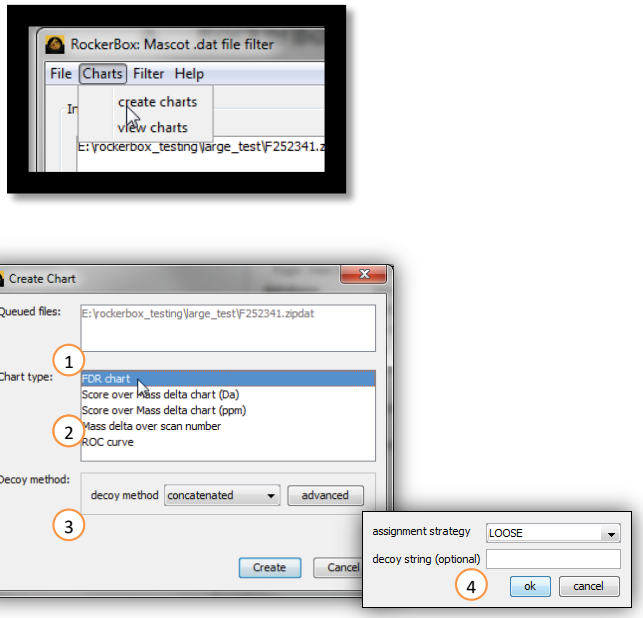
Salmi, J., Nyman, T.A., Nevalainen, O.S., and Aittokallio, T. (2009). Filtering strategies for improving protein identification in high-throughput MS/MS studies. Proteomics *9*, 848.

Toorn, H.W.P. van den, Mohammed, S., Gouw, J.W., Breukelen, B. van, and Heck, A.J.R.R.C.N.-0000 (2008). Targeted SCX Based Peptide Fractionation for Optimal Sequencing by Collision Induced, and Electron Transfer Dissociation. J. Proteomics Bioinform. *1*, 379.

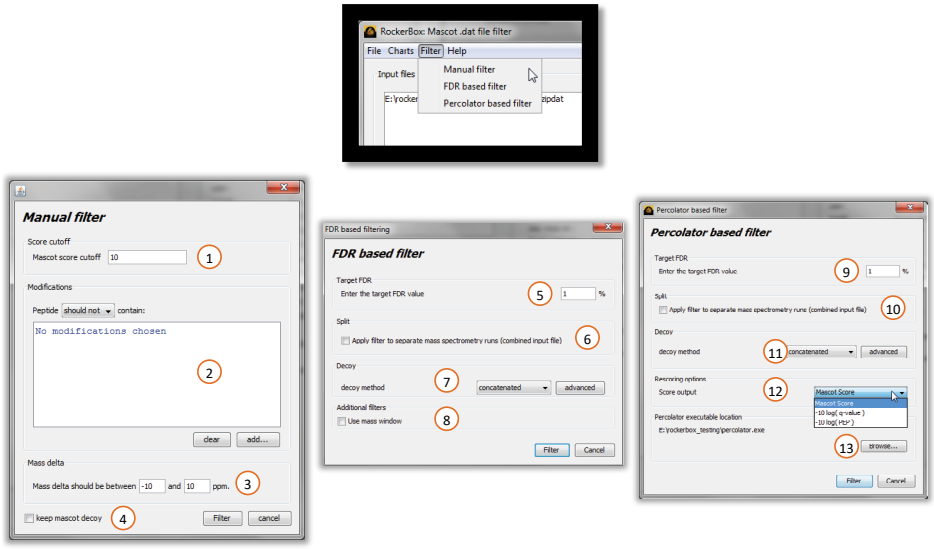
Supplementary figures



Supplementary figure S1a. RockerBox main window. (1) The menu area which gives access to all functionality, see below. (2) The file list, showing the files that are loaded to be processed. Multiple files can be listed here, enabling batch processing of files. (3) Message area, showing progress on the current operation. (4) Information area, showing search parameters for the currently selected file. (5) Button to export a parameter (.par) file for Mascot Daemon (6) Area for progress bar(s). If multiple processes run simultaneously, more than one progress bar may be visible. (7) Memory meter, indicating memory usage of the application. Clicking on it will force the release of unused memory.



Supplementary figure S1b. Charts dialog. (1) A list of files is shown in the 'queued files' list for which charts will be created. (2) In the chart type list, a chart type is selected. (3) All charts can be made from either 'concatenated' database searches or from a 'automatic decoy' search as performed by Mascot. For the 'concatenated' option, the 'advanced' button becomes active, which gives the option to enter a string for recognition of decoy proteins (e.g. "REV_") if this string is not present in the current implementation of RockerBox. (4) Furthermore, the mode of operation for whether to label a peptide as 'decoy' when any of the matching proteins come from the decoy database (STRICT) or label the peptide as 'target' if any of the matching proteins come from the target database (LOOSE).

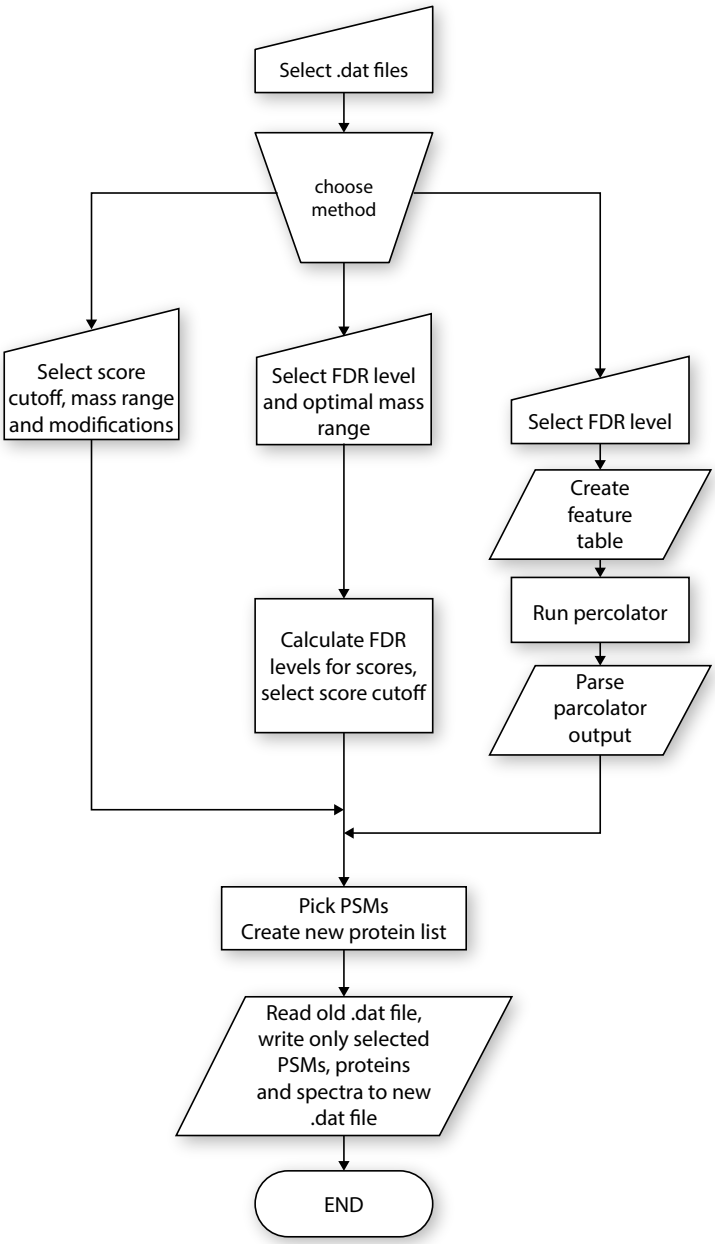


S1c. The three different filtering strategies present in RockerBox, which can be found in the "filter" menu in the main window.

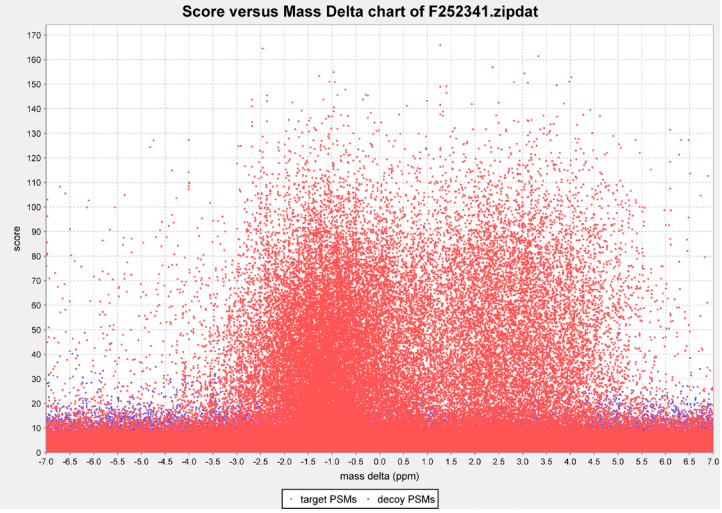
I. Manual filter dialog. There are four criteria to filter PSMs from a .dat file: (1) The mascot score cutoff is the minimum mascot ions score that a PSM should have to pass filtering. (2) The modifications chooser allow for the choice of one or more modifications and whether to include or exclude the chosen modifications. (3) Precursor masses are filtered according the mass window entered here. Optionally, the Mascot automatic decoy search results can be kept in the output file. Decoy PSMs are filtered from the output .dat file as well.

II. FDR based filter dialog. (5) A target FDR value is entered here (in percentage). (6) Using the Split option, mass spectrometry runs inside the .dat file can be filtered separately, or all together. The result of the filtering is always a single file. (7) The algorithm needs input about the type of decoy used for the experiment. The function of the 'advanced' button is described in b): *"charts dialog"*. (8) An additional filter can be added to the FDR based filter to restrict FDR calculations and filtering to a mass window. Clicking this option box will reveal a mass delta input field similar to the one in the manual filter dialog (3).

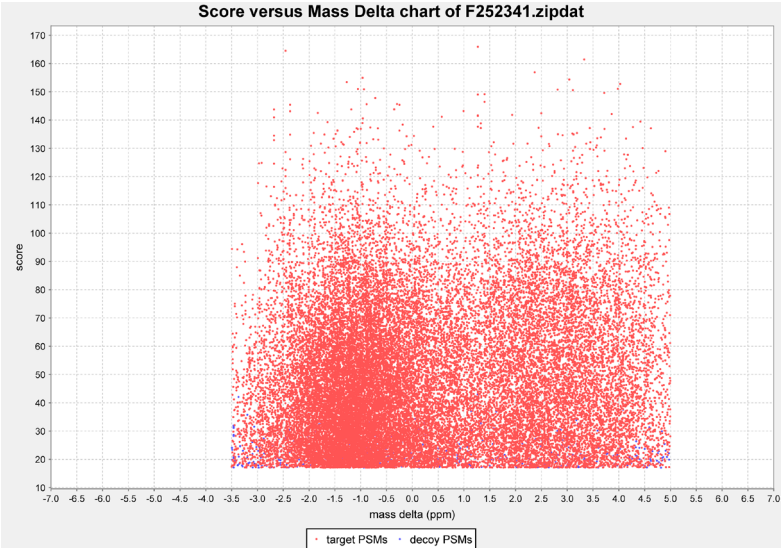
III. Percolator based filter dialog. (9), (10) and (11) correspond to (5), (6) and (7) of the FDR based filter dialog. (12) In the score output pull-down menu, the output score can be chosen. Choosing Mascot Score retains the Mascot ions score found from the input file. Choosing -10 log(q-value) or -10 log(PEP) replaces the scores in the output file with output values from the Percolator algorithm. (13) The location of the percolator executable, which can be downloaded from http://per-colator.com/ should be set here before this filter will work.



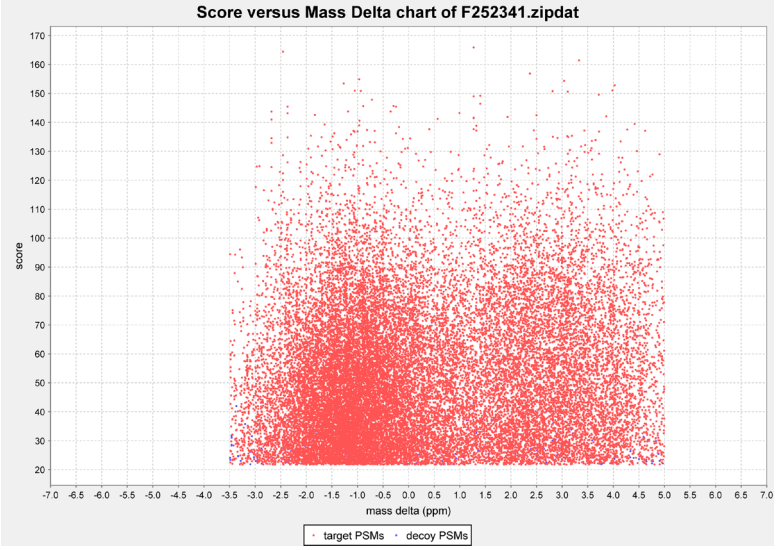
Supplementary figure S2. A (simplified) flow chart showing the filtering operations implemented in RockerBox.



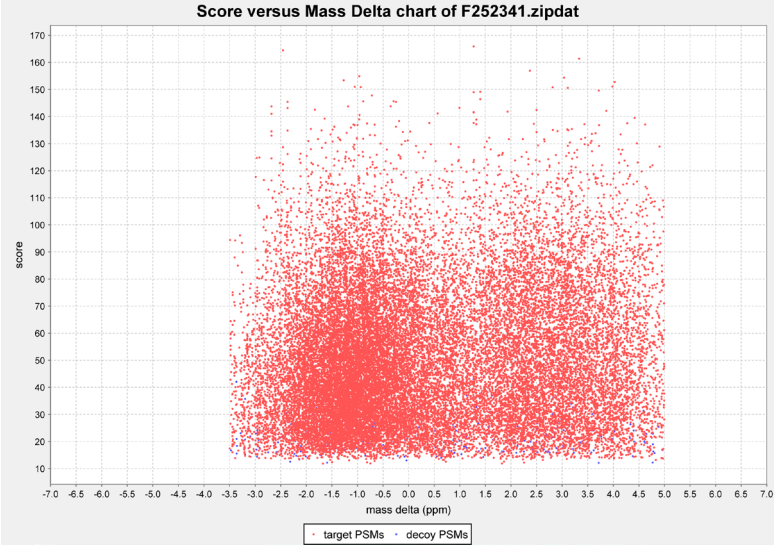
Supplementary figure S4a. An overview of Mascot score distribution related to mass delta values for the unfiltered file used in this article, also depicted in figure 2. Here, the decoy PSMs are shown on the same score scale as the target PSMs using blue dots.



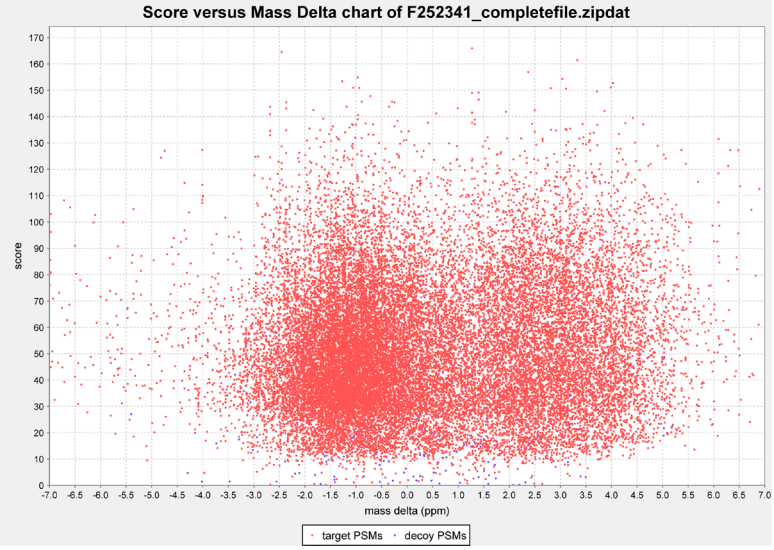
S4b. The result of a manual filtering operation with score cutoff is 20 and mass window of -3.5 to 5.0 ppm on the file described in Supplementary figure 1a. The uniformly distributed PSMs at the lower score range is deleted, as well as data points outside of the selected mass window. The estimated FDR after this operation is 4.3%.



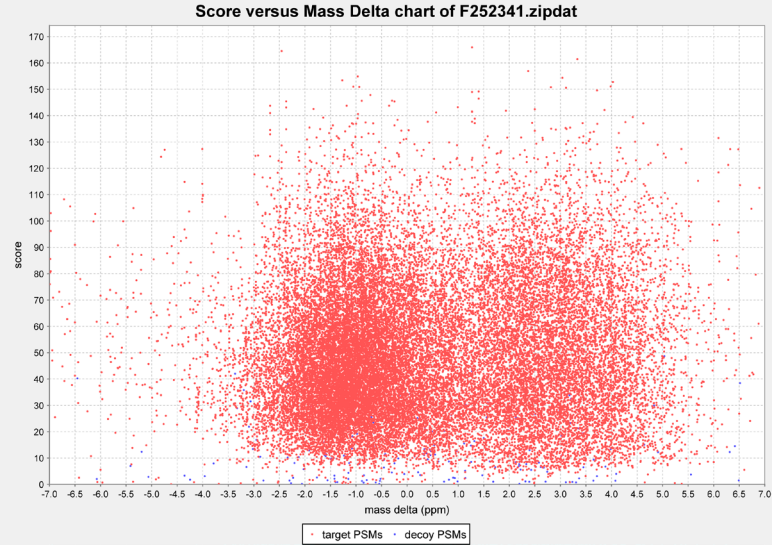
S4c. The result of a filtering operation where the score cutoff was determined automatically to obtain an FDR of 1%, within the mass window of -3.5 to 5.0 ppm.



S4d. The result of a filtering operation where the score cutoff was determined automatically to obtain an FDR of 1% *for every SCX fraction within the file*, in the mass window of -3.5 to 5.0 ppm. Compared to c. and d. the lower limit for the Mascot ions score is less defined because of different calculated score cutoff values for every fraction.



S4e. The result of filtering using the percolator algorithm, to achieve a FDR of 1%. The percolator algorithm allowed for many PSMs outside the mass window as determined by eye in the manual or FDR based filters.



S4f. The result of filtering using the percolator algorithm, to achieve a FDR of 1% calculated for every SCX fraction in the input file. Here, more PSMs are accepted with a lower mascot score and higher mass delta values.

	Accepted	Rejected
True	TP	FN
False	FP	TN

Supplementary table ST1a. Definition of terms. TP: true positives, FP: false positives, TN: true negatives, FN: false negatives. In terms of peptide identifications, true positives are accepted and true PSMs. Likewise, true negatives are rejected and false PSMs. On the other hand are false negatives that are true PSMs that are rejected and false positives are false PSMs that are accepted. In practice the numbers of FP, FN, FP and TN cannot be established directly and have to be inferred from other quantities.

Formulas for metrics of multiple statistical testing	
Precision	$\frac{TP}{TP + FP}$
FDR	$\frac{FP}{TP + FP} = 1 - precision$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$

ST1b. Formal list of terms as determined by their formulas.

Feature	Description
id	Identifier. RockerBox uses the form *db*_querynumber_rank, in which *db* may be 'target' for Mascot automatic decoy real database, 'decoy' for Mascot automatic decoy scrambled database or 'combined' for a concatenated decoy strategy
label	-1 if decoy, 1 if target PSM
charge	Precursor charge
score	Mascot score
deltaScore	Difference between current rank score and 'next' rank score
mr	Measured precursor mass
deltaM	Delta mass between precursor mass and matched peptide mass
deltaMPpm	deltaM relative to matched peptide mass
absDeltaM	Absolute value of deltaM
absDeltaMPpm	Absolute value of deltaMPpm
isoDeltaM	Delta mass allowing for 1, 2, 3 or 4 Dalton difference
isoDeltaMPpm	isoDeltaM relative to matched peptide mass
missedCleavages	Number of missed cleavages
fragMassError*	RMS error of the MS2 spectrum to the theoretical spectrum
totalIntensity*	Total intensity of the MS2 spectrum
intMatchedTot*	Total intensity of matched MS2 peaks
relIntMatchedTot*	intMatchedTot divided by totalIntensity
fracIonsMatched*	Fraction of all MS2 peaks matched
peptide	Peptide sequence
proteins	The list of proteins from the search database that contain the peptide sequence

Supplementary table ST2 An overview of all features exported from the RockerBox application in order to train the Percolator algorithm.

*Features marked with an asterisk are based on an internal peptide to spectrum matching algorithm, based on the ion series that Mascot has used for identification.

3. StatQuant: A post quantification analysis toolbox for quantitative mass spectrometry in proteomics

van Breukelen, B., van den Toorn, H.W.P., Drugan, M., and Heck, A.J.R. (2009). StatQuant: a post-quantification analysis toolbox for improving quantitative mass spectrometry. Bioinformatics 25, 1472–1473.

Motivation: Mass spectrometric protein quantitation has emerged as a high-throughput tool to yield large amounts of data on peptide and protein abundances. Currently, differential abundance data can be calculated from peptide intensity ratios by several automated quantitation software packages available. There is, however, still a great need for additional processing to validate and refine the quantitation results. Here, we present a software tool, termed StatQuant, that offers a set of statistical tools to process, filter, compare and represent data from several quantitative proteomics software packages such as MSQuant. StatQuant offers the researcher post-processing methods to achieve improved confidence on the obtained protein ratios.

Availability: StatQuant can be downloaded from: https://www.hecklab.com/software/statquant/ (binary and source code). Contact: b.vanbreukelen@uu.nl

Introduction

Mass spectrometric protein quantification has become a powerful tool to determine differences in protein abundances between proteomes. Such data allows the identification of those proteins that play a role in, for example, cancer development, responses to stimuli and cell signaling cascades. Typically, samples containing proteins or digested proteins are labeled with stable isotopes and subsequently mixed with an equal amount of unlabeled or differentially labeled (control) sample(s). In case of intact proteins the samples are first digested before mass spectrometric analysis. The mixture is then separated by nanoliter flow liquid chromatography and tandem mass spectrometric analysis (LC-MS/MS). Typically, isotopomers of peptides will co-elute from the LC column and thus simultaneously appear in the mass spectra. Based on their differences in mass they can be differentiated and quantified on their relative intensities. Peptides can subsequently be identified using mass spectrometric peptide sequencing and assigned to their corresponding proteins. Following mass spectrometric analysis, a variety of dedicated software tools can be used to calculate protein abundances. Usually, such software reads the raw mass spectrometer data which it combines with peptide identifications to calculate the abun-

dance for each peptide. Proteins are commonly identified with confidence by two or more peptides. Therefore, a protein ratio will be the average of its peptide ratios. As a consequence, variation in peptide ratios results in a protein ratio with a particular SD. Typically, the software outputs a table of protein identifications; its associated peptides; and the relative abundance ratios. In successive rounds of analysis this data is used for quality control and in-depth analysis. Often this part is performed manually using a spreadsheet. In general, protein ratio values as derived from these tools are verified in follow-up experiments before the conclusions about the significantly up- or downregulated proteins are drawn.

In order to address the lack of proper post-analysis tools and to provide the user with an easy to use graphical user interface, we developed a software tool in the Java programming language (using JAVA version 1.6 or higher) named StatQuant. StatQuant provides methods for data normalization, cross-experiment comparison, outlier detection, data visualization and significance testing (P-values). P-values are calculated using the protein abundance ratios and their SD of its associated peptide ratios using one sample t-tests. This provides the user with an additional tool to assess the reliability or confidence on the obtained protein ratios.

Features

StatQuant reads data from protein quantitation software. In generic mode, StatQuant requires simple TAB delimited data with only few required columns, including protein ID, peptide sequence, signal (or intensity) 1 and signal 2. Furthermore a dedicated importer for MSQuant (Andersen et al., 2003; Schulze and Mann, 2004) results has been implemented.
Upon importing quantitation data several options are provided, for example, to combine multiple data files that are part of the same experiment (Figure 1). The user can choose to present the data in a peptide or protein-centric view. Imported data are represented in two tables in the main view (Figure 2). The top table presents an overview of all features, which are, depending on the selected view; either proteins or peptides; the amount of peptides associated to that feature; ratios both in normal and ^{2}Log scale; SD of the feature ratio; the corresponding P-value; and a Q-val-

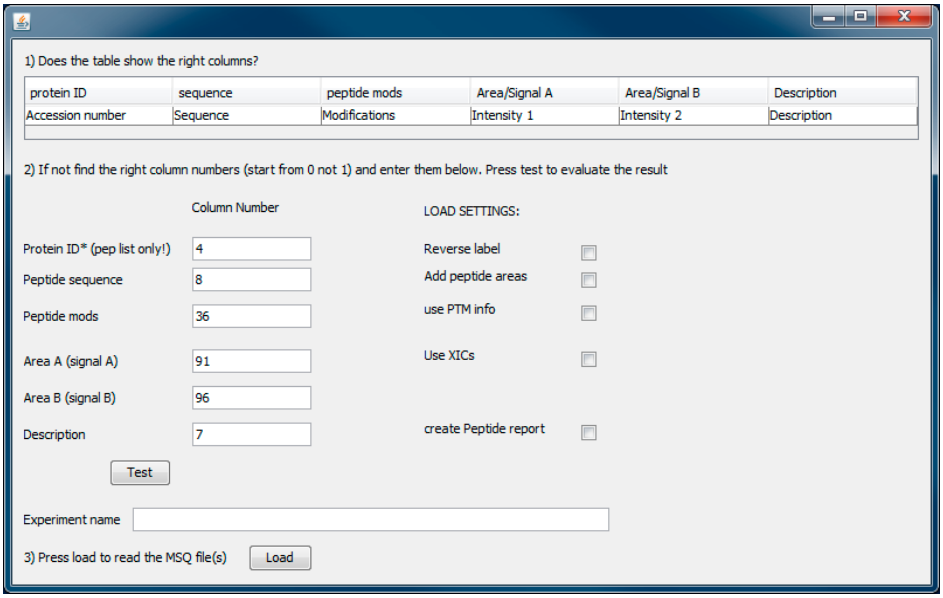


Figure 1: Loading options screen. The user is presented with several option for loading (importing) the data into StatQuant. In the top table the selected columns are shown. Here the user can select different columns if desired. Additionally the user can choose to add peptide areas, import XIC's or Signal intensities, create a peptide report, swap labels (in case of a reverse label experiments) and use the MSQuant specific PTM information.

ue. The bottom table presents detailed information on the feature including: the peptides and their modifications (PTMs); search engine scores; signal or intensity values and their ratios. The bottom panel also allows the user to select peptides and disqualify them for further analysis.

StatQuant offers the possibility for normalization based on the median intensity, a selected group of features or on a user-defined value. The effect on the distribution of the ratios is visualized through graphs, that on the X-axis present the ^{2}Log ratio (M) and on the Y-axis show the average (A) ^{2}Log intensity of both signals (MA plots, Figure 1 right, Figure 4).

Stable isotope labeling with amino acids in cell culture (SILAC) is a common method to generate samples for quantitative proteomics (Ong et al., 2002, 2003a). SILAC experiments may be hampered by the in-cell conversion of Arg to Pro impeding accurate quantitation (Hwang et al., 2006). To accommodate this problem StatQuant can modify the experimental-

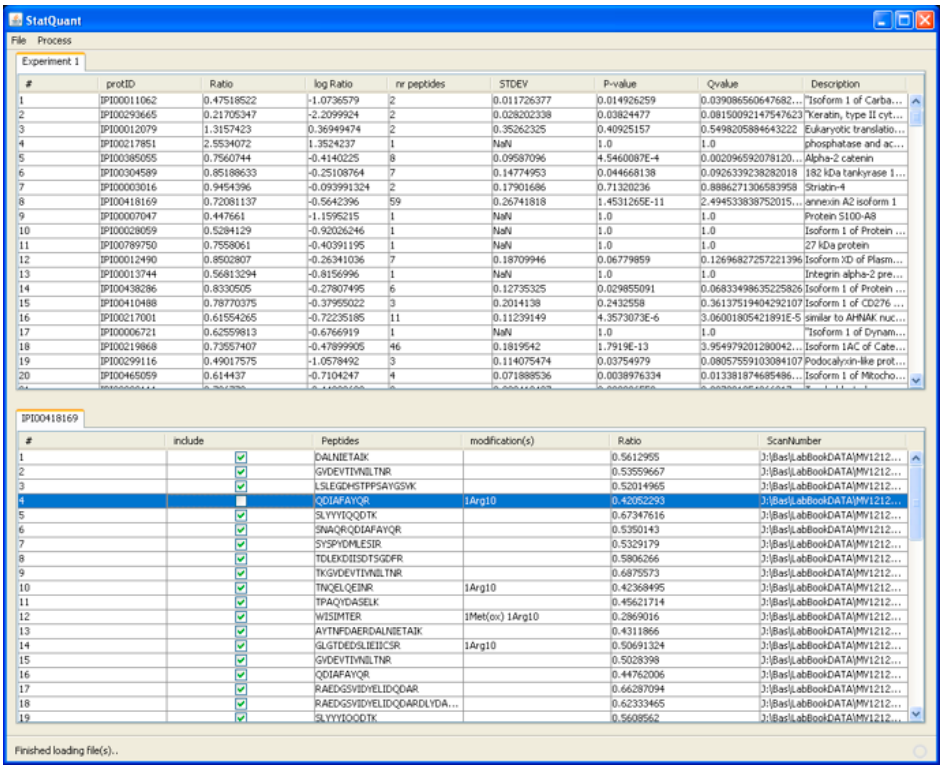


Figure 2: Data is presented in a table. The top table shows the proteins, their ratios and statistics. In the bottom table a per protein peptide overview is shown with the peptide ratio and additional peptide information. Moreover, peptides can be excluded from quantification here.

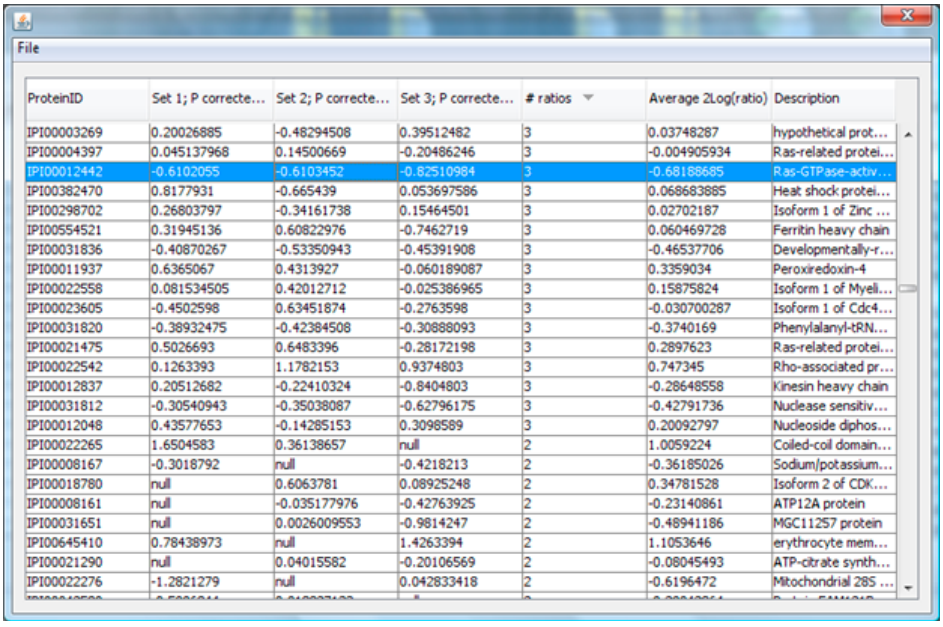
ly derived, Pro containing, peptide ratios based on a precalculated conversion rate which can be experimentally determined (Ong et al., 2003b; Van Hoof et al., 2007). A binomial distribution is then used to calculate the conversion rate for all other peptides containing one or more Pro residues, which improves quantitation significantly (Gruhler et al., 2005; Figure 5).

On top of the data processing tools StatQuant contains a series of filter steps to refine the data and to increase the confidence of the obtained ratios. Outlier detection is performed by applying the Z-score algorithm, which automatically unselects all peptides that have a ratio outside the 2 SDs interval of the feature ratio. Outliers are not always that obvious and do not always originate from experimental error. For example, proteins can be expressed in multiple isoforms or variants. As a result it is pos-

sible that multiple proteins share identical peptide sequences. To address that problem, StatQuant has a method to check for 'peptide uniqueness', which searches all peptides against a sequence database to determine if this peptide is unique to a single entry. StatQuant offers multiple export functions. It can export all visible tables in tabular format, but can also produce a comparison of multiple experiments, where only the overlapping proteins are presented (Figure 3).

Discussion

Most protein quantitation software packages provide the user with ratio data only, which often needs manual post-processing to ascertain the data and obtain values for its reliability and/or significance. Here, we present a versatile tool that can be generally used for post-processing of quantitative proteomics data, termed StatQuant. The graphical user interface offers options for data normalization, filtering, comparison and representation.


File

ProteinID	Set 1; P correcte...	Set 2; P correcte...	Set 3; P correcte...	# ratios	Average 2Log(ratio)	Description
IPI00003269	0.20026885	-0.48294508	0.39512482	3	0.03748287	hypothetical prot...
IPI00004397	0.045137968	0.14500669	-0.20486246	3	-0.004905934	Ras-related protei...
IPI00012442	-0.6102055	-0.6103452	-0.82510984	3	-0.68188685	Ras-GTPase-activ...
IPI00382470	0.8177931	-0.665439	0.053697586	3	0.068683885	Heat shock protei...
IPI00298702	0.26803797	-0.34161738	0.15464501	3	0.02702187	Isoform 1 of Zinc ...
IPI00554521	0.31945136	0.60822976	-0.7462719	3	0.060469728	Ferritin heavy chain
IPI00031836	-0.40870267	-0.53350943	-0.45391908	3	-0.46537706	Developmentally-r...
IPI00011937	0.6365067	0.4313927	-0.060189087	3	0.3359034	Peroxiredoxin-4
IPI00022558	0.081534505	0.42012712	-0.025386965	3	0.15875824	Isoform 1 of Myeli...
IPI00023605	-0.4502598	0.63451874	-0.2763598	3	-0.030700287	Isoform 1 of Cdc4...
IPI00031820	-0.38932475	-0.42384508	-0.30888093	3	-0.3740169	Phenylalanyl-tRN...
IPI00021475	0.5026693	0.6483396	-0.28172198	3	0.2897623	Ras-related protei...
IPI00022542	0.1263393	1.1782153	0.9374803	3	0.747345	Rho-associated pr...
IPI00012837	0.20512682	-0.22410324	-0.8404803	3	-0.28648558	Kinesin heavy chain
IPI00031812	-0.30540943	-0.35038087	-0.62796175	3	-0.42791736	Nuclease sensitiv...
IPI00012048	0.43577653	-0.14285153	0.3098589	3	0.20092797	Nucleoside diphos...
IPI00022265	1.6504583	0.36138657	null	2	1.0059224	Coiled-coil domain...
IPI00008167	-0.3018792	null	-0.4218213	2	-0.36185026	Sodium/potassium...
IPI00018780	null	0.6063781	0.08925248	2	0.34781528	Isoform 2 of CDK...
IPI00008161	null	-0.035177976	-0.42763925	2	-0.23140861	ATP12A protein
IPI00031651	null	0.0026009553	-0.9814247	2	-0.48941186	MGC11257 protein
IPI00645410	0.78438973	null	1.4263394	2	1.1053646	erythrocyte mem...
IPI00021290	null	0.04015582	-0.20106569	2	-0.08045493	ATP-citrate synth...
IPI00022276	-1.2821279	null	0.042833418	2	-0.6196472	Mitochondrial 28S ...

Figure 3: Experimental comparison report. In this table StatQuant shows all protein and their ratio's as obtained from the different experiments. This view can be exported as a TAB delimited text file, and can be loaded into ms-excel or other analysis tools such as SpotFire to create heatmaps.

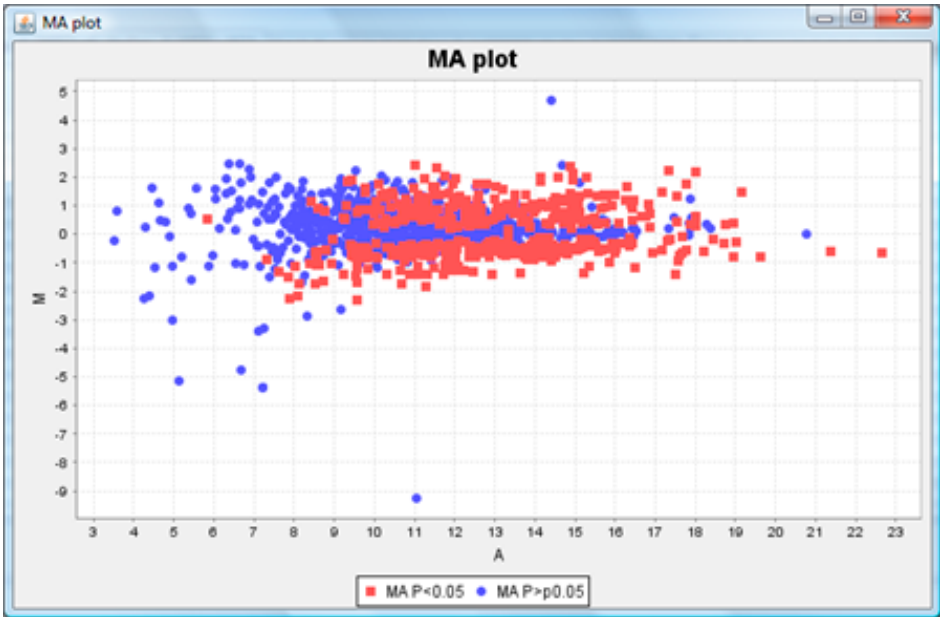


Figure 4: Visualization of protein/peptide ratios using a MA plot. On the x-axis the sum of the peptide intensities (per protein) A= $\left(\frac{\log_2 \sum signal\ A + \log_2 \sum Signal\ B}{2}\right)$ is plotted against the $\log_2$-Ratio on the y-axis (M). Protein ratios with a significance (p-value) above 0.05 are shown as blue squares whereas the protein ratios with a p-value below 0.05 are shown in red.

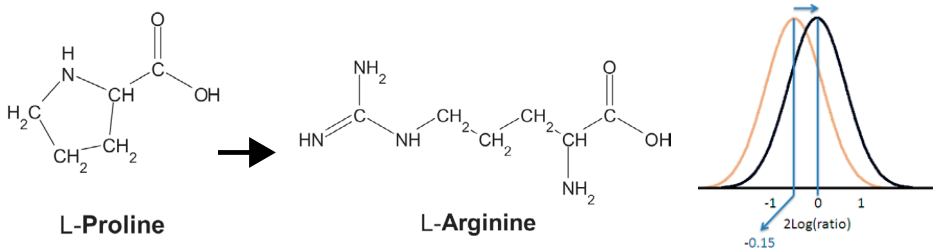


Figure 5: Proline conversion shifts ratios in SILAC experiments, which can be compensated for in StatQuant

StatQuant was written such that it can accommodate data from different protein quantitation software packages. Processed data from multiple experiments can be compared, processed and exported.

StatQuant has already been successfully implemented to process and filter large datasets, aiding the refinement of data such that only reliable protein

ratios were reported (Boersema et al., 2008; Raijmakers et al., 2008).

Acknowledgements

We would like to thank Shabaz Mohammed and Reinout Raijmakers for beta-testing StatQuant, their comments and suggestions.
Funding: Netherlands Proteomics Centre; Netherlands Bioinformatics Centre.

Conflict of Interest: none declared.

References

Andersen JS, et al. Proteomic characterization of the human centrosome by protein correlation profiling, Nature, 2003, vol. 426 (pg. 570-574)

Boersema PJ, et al. Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates, Proteomics, 2008, vol. 8 (pg. 4624-4632)

Gruhler A, et al. Stable isotope labeling of Arabidopsis thaliana cells and quantitative proteomics by mass spectrometry, Mol. Cell. Proteomics, 2005, vol. 4 (pg. 1697-1709)

Hwang SI, et al. Systematic characterization of nuclear proteome during apoptosis: a quantitative proteomic study by differential extraction and stable isotope labeling, Mol. Cell. Proteomics, 2006, vol. 5 (pg. 1131-1145)

Ong SE, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics , Mol. Cell. Proteomics, 2002, vol. 1 (pg. 376-386)

Ong SE, et al. Mass spectrometric-based approaches in quantitative proteomics, Methods, 2003, vol. 29 (pg. 124-130)

Ong SE, et al. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC), J. Proteome Res., 2003, vol. 2 (pg. 173-181)

Raijmakers R, et al. Automated online sequential isotope labeling for protein quantitation applied to proteasome tissue-specific diversity, Mol. Cell. Proteomics, 2008, vol. 7 (pg. 1755-1762)

Schulze WX, Mann M. A novel proteomic screen for peptide-protein interactions, J. Biol. Chem., 2004, vol. 279 (pg. 10756-10764)

Van Hoof D, et al. An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics, Nat. Methods, 2007, vol. 4 (pg. 677-678)

4. Targeted SCX based peptide fractionation for optimal sequencing by collision induced, and electron transfer dissociation

Toorn, H.W.P. van den, Mohammed, S., Gouw, J.W., Breukelen, B. van, and Heck, A.J.R. (2008). Targeted SCX Based Peptide Fractionation for Optimal Sequencing by Collision Induced, and Electron Transfer Dissociation. J. Proteomics Bioinform. 1, 379.

Electron transfer dissociation (ETD) of peptide ions has been introduced as a tool for mass spectrometry based peptide sequencing, complementary to the commonly used collision induced dissociation (CID). It has been proposed that ETD may have better performance than CID for more highly charged and/or larger peptides. Here, we compare the performance of ETD and CID on data generated in a large-scale proteomics experiment. First, tryptic proteolytic peptides of Drosophila melanogaster oocytes were off-line separated based on their in-solution net charge state using strong cation exchange chromatography (SCX), followed by an on-line reverse-phase (RP) liquid chromatography separation coupled to an ion trap mass spectrometer with ETD capabilities. The mass spectrometer selected MS peaks were subjected to both ETD and CID thus allowing a fair comparison.. Around 2300 peptides were exclusivelyidentified by CID and similiarly more than 3000 by ETD with approximately 1400 by both ETD and CID. In total nearly 7,000 peptides were identified with a very conservative Mascot peptide cut-off score of 60 clearly verifying that ETD and CID are complementary techniques. In the early SCX fractions, which contain peptides with a 'low' net charge, more than 90% of the peptides could be successfully identified by CID whereas in the later SCX fractions more than 90% of the identified peptides could be successfully identified by using ETD only. The chosen strategy, with a combination of SCX and RP-LC-MS/MS, allows the user to make targeted decisions on whether to optimally use CID and/or ETD. Analysis of the sequence and amino acid contents of all identified peptides clearly revealed that the impressive performance of ETD for peptides possessing charge states above three do not require CID based sequencing which, at best, would be solely confirmatory.

Introduction

Several strategies are available for performing large-scale analyses of complex protein mixtures(Aebersold and Mann, 2003; Brunner et al., 2007; Chen et al., 2006; Kolkman, 2005; Krijgsveld et al., 2006; Witze et al., 2007). The 'shotgun' peptide-centric approach is popular for such analyses, involving the generation of in-solution tryptic digests of whole lysates. The complexity of the sample introduced into the mass spectrometer is re-

duced by using multidimensional separation techniques where, typically, the first dimension consists of strong cation exchange (SCX) chromatography (Wu et al., 2003), hydrophilic interaction chromatography (HILIC) (Boersema et al., 2007) or peptide iso-electric focusing (IEF)(Cargile et al.; Krijgsveld et al., 2006). In particular, the combination of SCX as a first dimension for separation of the peptides with nanoflow reversed phase (RP) chromatography has been shown to be extremely powerful (MacCoss et al., 2002). Unfortunately, when using such an approach there will be an undersampling of the total peptide population. This is partly caused by the fact that the separation power of multidimensional chromatography is still insufficient and consequently too many peptides will co-elute and compete with each other for ionization and mass spectrometric sequencing. This drawback can be partly overcome by a repeating the analysis for each sample several times since peptide sampling by the mass spectrometer is partly random(de Groot et al., 2007; Lipton et al., 2002; Liu et al., 2004; Shen et al., 2005). Another reason why not all peptides are successfully sequenced lies in the fact that most current electrospray based mass spectrometers have an optimal m/z range for analysis which lies between 300-1500 Th. In-solution digestion using trypsin may not allow a complete analysis due to certain proteolytic peptides falling outside this optimal window (MacCoss et al., 2002; Mohammed et al., 2008). Larger and highly charged tryptic peptides are often sequenced poorly by CID based tandem MS, partly due to insufficient mass resolution to assign the correct charge state for the precursor and product ions as well as poor fragmentation(Paizs and Suhai, 2005). All in all, new methods that will enable improved proteome coverage by using techniques complemantary to CID, would be welcome.

Recently electron transfer dissociation (ETD) has been introduced as a new peptide sequencing method(Good et al., 2007; Syka et al., 2004), and through its mode of operation exhibits properties that are complementary to collision induced dissociation (CID). In ETD, an electron is transferred from a radical anion, usually fluoranthene, to the protonated peptide, inducing fragmentation and formation of c and z type ions. The exact mechanism of how ETD promotes fragmentation is however still under debate(-Leymarie et al., 2003; Syka et al., 2004; Zubarev et al., 1998). It has been shown that ETD can be effective at fragmenting peptides with the higher

charge state peptides that CID would often struggle to identify. However, earlier studies have shown that doubly charged peptides do not efficiently fragment in ETD experiments due to the fact that dissociation efficiency by ETD is related to the number of charges present on the precursor ion. To circumvent the difficulties of analyzing doubly protonated peptides by ETD techniques a limited amount of collisional activation is applied to precursor cations after electron transfer for more efficient fragmentation, so-called ETcaD (Pitteri et al., 2005; Swaney et al., 2007).

To obtain a more in-depth sense for the performance of ETD as compared to CID, we performed a medium-scale proteome analysis of early Drosophila melanogaster embryos in which all peptides were subjected to CID and ETD. We generated a dataset of approximately 7000 peptides that were sequenced by CID and ETD when applying a conservative Mascot cut-off score of 60. Annotated spectra were extracted from Mascot result files (.dat) that fulfill the score requirements using an in-house developed software tool, while systematic statistical analysis on the dataset were performed with simple Perl scripts. Strikingly, the result was the discovery for the overlap between the the ETD and CID data-sets being less than 25%. Looking into the specifics on each peptide we conclude from the data that CID favors smaller and less basic peptides, whereas ETD favors

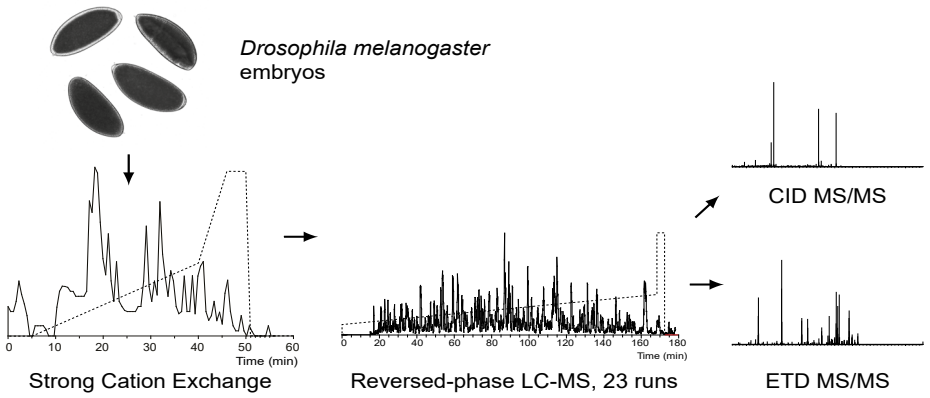


Figure 1: Scheme of the experimental setup. Tryptic peptide digests of Drosophila melanogaster embryo lysates were first separated by off-line strong cation exchange (SCX), where each fraction was analysed by reversed-phase liquid chromatography on-line coupled to a mass spectrometer. Individual peptide ions were selected by the mass spectrometer and fragmented using sequentially both collision induced dissociation (CID) and electron transfer dissociation (ETD).

longer and more highly charged and therefore more basic peptides. As SCX largely separate peptides on charge(Beausoleil et al., 2004), which is also clearly revealed by the current data-set, an optimal strategy can be proposed whereby the early fractions are predominantly analyzed by CID-MS, whereas the late fractions would solely require ETD.

Experimental

Fly stock and embryo collection and sample preparation

Wild-type OregonR flies were maintained by standard methods at 25 °C. Wild-type embryos were collected on agarose-agar plates, washed in water and dechorionated by incubation in 2.5% sodium hypochlorite for 90 s followed by another wash and kept at -20 °C. About 5 mg of embryos were lysed in 8 M urea and 50 mM ammonium bicarbonate. Cellular debris was pelleted by centrifugation at 20,000 g for 20 minutes. Prior to digestion, proteins were reduced with 1 mM DTT and alkylated with 2 mM iodoacetamide. The mixture was diluted 4-fold to 2 M urea using 250 µL of 50 mM ammonium bicarbonate and 50 µL of trypsin solution, 0.1 mg/mL, and incubated overnight at 37 °C.

Strong Cation Exchange.

Strong cation exchange was performed using a Zorbax BioSCX-Series II column (0.8 mm i.d. × 50 mm length, 3.5 µm), a FAMOS autosampler (LC-packing, Amsterdam, The Netherlands), a Shimadzu LC-9A binary pump and a SPD-6A UV-detector (Shimadzu, Tokyo, Japan). Prior to SCX chromatography, protein digests were desalted using a small plug of C18 material (3 M Empore C18 extraction disk) packed into a GELoader tip (Eppendorf) similar to what has been previously described(Rappsilber et al., 2003), onto which ~10 µL of Aqua C18 (5 µm, 200 Å) material was placed. The eluate was dried completely and subsequently reconstituted in 20% acetonitrile and 0.05% formic acid. After injection, a linear gradient of 1% min−1 solvent B (500 mM KCl in 20% acetonitrile and 0.05% formic acid, pH 3.0) was performed. A total of 45 SCX fractions (1 min each, i.e., 50 µL elution volume) were manually collected and dried in a vacuum

centrifuge, of which 23, that contained most peptides, were subjected to our mass spectrometric analysis by RP-LC MS/MS.

Nanoflow-HPLC-MS.

Dried residues were reconstituted in 50 µL of 0.1 M acetic acid and were analyzed by nanoflow liquid chromatography using an Agilent 1100 HPLC system (Agilent Technologies) coupled on-line to a LTQ-XL mass spectrometer (Thermo-Fisher Scientific). The liquid chromatography part of the system was operated in a setup essentially as described previously (Licklider et al., 2002; Meiring et al., 2002). Aqua C18 (Phenomenex), 5 µm resin was used for the trap column, and ReproSil-Pur C18-AQ, 3 µm, (Dr. Maisch GmbH) resin was used for the analytical column. Peptides were trapped at 5 µL/min in 100% solvent A (0.1 M acetic acid in water) on a 2 cm trap column (100 µm i.d., packed in-house) and eluted to a 20 cm analytical column (50 µm i.d., packed in-house) at ~100 nL/min in a 150-min gradient from 10 to 40% solvent B (0.1 M acetic acid in 8/2 (v/v) acetonitrile/water). The eluent was sprayed via standard coated emitter tips (New Objective), butt-connected to the analytical column. The mass spectrometer was operated in the data dependent mode to automatically switch between MS and MS/MS ETD and MS/MS CID. Survey MS spectra were acquired from m/z 350 to m/z 1500 in the LTQ after accumulation to a target value of 30,000 in the linear ion trap. The two most intense ions were fragmented in the linear ion trap at a target value of 10,000. To prevent repetitive analysis of the same ion, dynamic exclusion technology (Thermo Fischer Scientific) was used.

Data extraction and analysis

The MS2-data was extracted from the raw data file with a beta-release of Bioworks 3.4 (Thermo-Fisher Scientific) into separate spectrum (.dta) files using the Sequest preprocessor, without additional filtering. The standard method of charge state assignment is to use the Charger program (Thermo-Fisher Scientific) (Sadygov et al., 2008) on every ETD tandem mass spectrum. We let the Charger program analyze our experimental data to assess its performance. All other analyses were performed without the dependency on the Charger program via concatenating the spectra in Mas-

cot Generic Files (.mgf), where the CHARGE field for each peak list corresponding to an individual spectrum would contain values from 2+ to 7+. Tandem MS ion searches were performed with Mascot 2.2 (Matrix Science inc.) on a concatenated database of Drosophila melanogaster sequences in the Swissprot and the TREMBL databases, with a peptide tolerance of 3.0 Da and a MS/MS tolerance of 0.9 Da. Subsequently, the Mascot result files (.dats) were retrieved from the server and each underwent an extraction procedure. For all SCX fractions, the highest scoring peptide hit for each spectrum was retrieved. Sequence, ion score, charge state and precursor mass were stored in a text file. Similarly, for protein identifications the best protein hit for each peptide identification was retrieved. Protein identification required a minimum of two peptides to be identified. All algebraic operations regarding peptides and protein identifications were performed with Perl (Activestate Perl 5.8.8) scripts and visualized with Microsoft Excel 2007. All the Perl scripts, the original input text files and the Excel workbook (in Excel 2007 format) are made available in the supplementary material (http://bioinformatics.chem.uu.nl/vdtoorn_etdcid). All raw data with identifications has been submitted to the PRIDE repository at the EBI (http://www.ebi.ac.uk/pride, in a project with the name "Targeted SCX based peptide fractionation for optimal sequencing by collision induced, and electron transfer dissociation", accession numbers 8697-8742 inclusive)

Results and Discussion

Drosophila melanogaster embryos were lysed and the peptide mixture generated by trypsin proteolysis was subjected to SCX fractionation (Figure 1). Twenty three 1 minute fractions were subjected to analysis by RP-LC-MS/MS. The linear ion trap mass spectrometer was operated in the data dependent mode and switched automatically between MS, ETD and CID.

Precursor ion charge state determination

The linear ion trap mass spectrometer has limited mass resolution when performing a standard scan and therefore correct determination of the

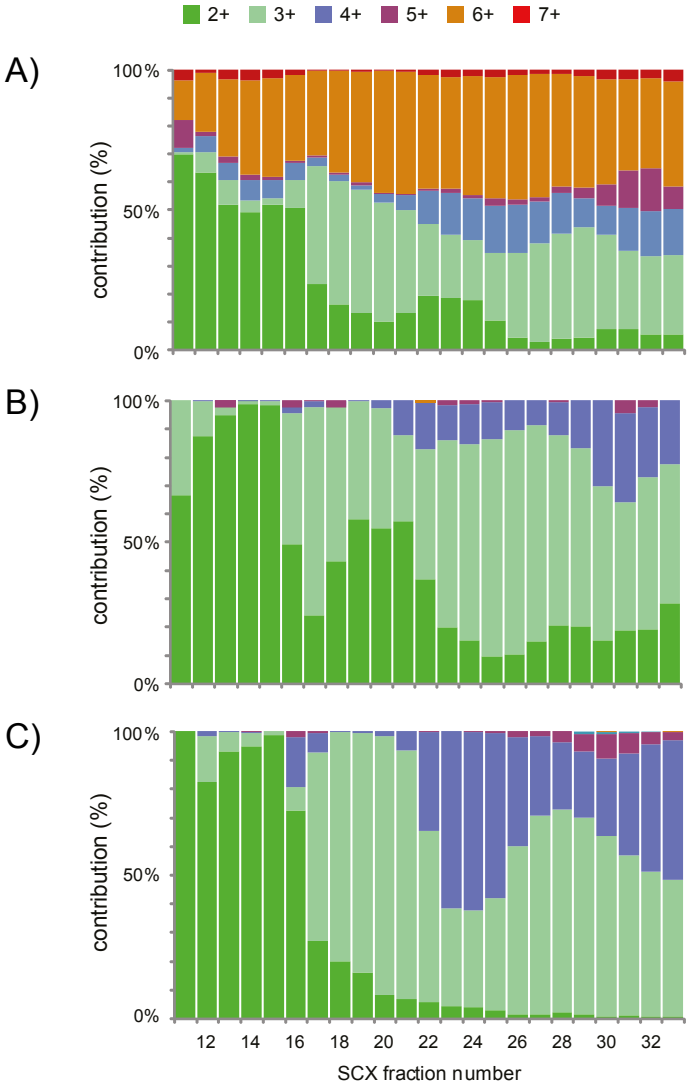


Figure 2: Correct determination of precursor ion charge states are vital for optimal peptide sequencing. In **(A)** the charge states were determined by the Bioworks 'charger' preprocessor program.. For each SCX fraction the relative contribution of ions of each charge state is shown. The charge assigned in **(B)** and **(C)** were determined by the charge of the highest scoring peptide (minimum peptide score of 60) in a Mascot database search. Mascot was instructed to search with the charge state being between 2+ and 7+ for each MS-spectrum. The charge state contributions for both CID data **(B)** and ETD data **(C)** are shown. The color legend at the top provides correlation with theassigned charges. The absolute number of identified peptides in each SCX fraction is listed in supplementary Figure 4.

precursor ion charge state requires additional time-consuming scans. A specific method of charge state determination is available for ETD MS spectra, which exploits the knowledge that the transferred electron(s) might not necessarily induce dissociation, but lead to intact peptides ions with a reduced charge state. The Charger program, which is part of the Bioworks package (Thermo-Fisher Scientific) (Sadygov et al., 2008) tries to determine charge states using these charge-reduced species that are present within the ETD spectra. An overview of Charger program output for the mass spectrometric data acquired is shown in Figure 2A. We started our comparison at fraction 11 which was the first to contain a reasonable number of peptides, and stopped at fraction 33. As expected, there is a trend of increasing peptide charge states with increasing fraction number for the SCX run. Notably, in every SCX fraction analysis many spectra are, quite unrealistically, assigned as 6+ charge states, especially compared to the number of peptides with 4+ and 5+ charges, suggesting there is a weakness in identifying 3+ peptides confidently possibly caused by poor peak detection. In order to check the confidence in charge state assignment we instructed the Mascot search engine to consider all charge states between 2+ and 7+ for each submitted ETD spectrum. The peptide sequence that was assigned with the highest Mascot score was assumed to indicate the correct charge state. For the identifications, we set an ion score cut-off of 60 to solidify our assumption, which allowed a False Discovery Rate (FDR) of below 0.3 % for all peptide rich fractions as determined by the use of a decoy database search (Supplementary figure 1). Figure 2B and C summarize charge state trends detected by Mascot based discrimination. Figure 2B indicates the charge states for all peptides sequenced by CID, whereas figure 2C contains the analogous data obtained from the ETD analysis. From the data presented in figure 2C it is apparent that, indeed, the higher number of 6+ charge states assigned by the Bioworks Charger program were erroneous. When the 6+ and 7+ charged peptides are removed from the Charger results (supplementary figure 2), the results show improved, though not perfect, agreement with the data presented in Figure 2C.

Our analysis reveals that for all peptides successfully identified charges of up to 4+ are detected frequently but higher charge states are much less common. The higher number of identified peptides in lower fractions for CID as compared to the higher number of identified peptides in higher

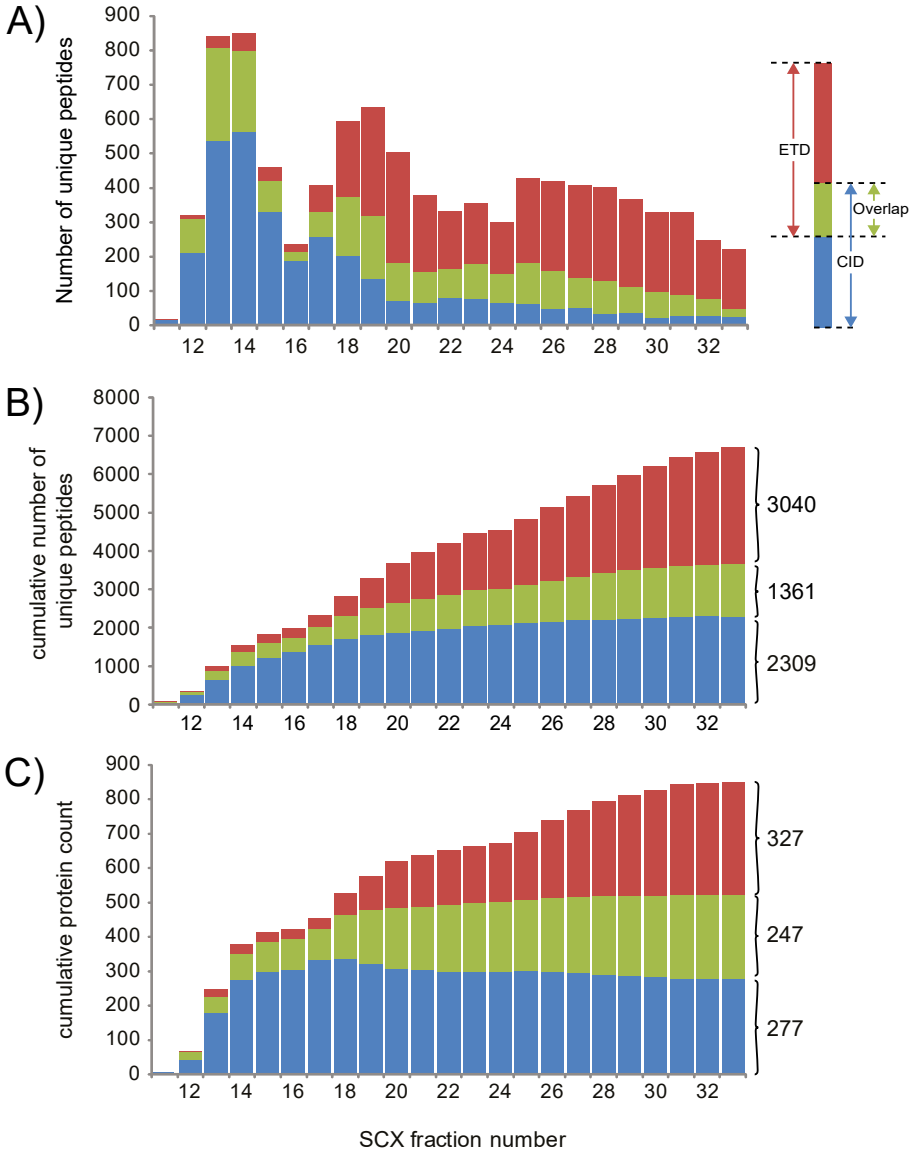


Figure 3: ETD and CID are complementary. **(A)** Number of unique peptide identifications per SCX fraction. The number of peptides exclusively identified with CID are in blue, the number of peptides exclusively identified with ETD are in red, and the number of peptides identified with both activation methods are in green. **(B)** The cumulative number of unique peptides identified based on all SCX fractions. On the right the accumulated peptide numbers are shown. The absolute number of identified peptides in each individual SCX fraction is listed in supplementary Figure 4. **(C)** The cumulative number of unique proteins over all SCX factions.

scx fractions (Figure 2B and 2C) indicate that CID has a preference for 2+ peptide precursor ions, whereas ETD is relatively more successful in sequencing higher charge state precursor peptide ions.

Comparison of performance between ETD and CID

In order to compare the performance between the ETD and CID methods we calculated the number of unique peptides identified in each SCX fraction. We separated the identified peptides (Mascot peptide score > 60) into three categories (figure 3); peptides exclusively identified by CID (blue), peptides exclusively identified by ETD (red), and peptides mutually identified by both CID and ETD (green). We also analyzed data using a cut-off of 40, which is presented in Supplementary Figure 1B. From figure 3A it can be observed there are approximately three broad maxima in the total peptide numbers observed over the full SCX run: around fractions 13/14, around fractions 18/19, and around fractions 25 to 29. These maxima correspond most likely to the elution profiles of differentially charged peptides in the SCX separation i.e. the 2+, 3+ and >4+ peptides, conforming to the results shown in Figures 2B and 2C. Around the first maximum (i.e. 2+ net charge) many more unique peptides are found with CID compared to ETD. Moreover, ETD adds little to the overall number of peptide identifications for these SCX fractions (note: supplemental activation was applied for the 2+ peptides analyzed by ETD). SCX fractions, containing 3+ peptides, 17 onwards, show a significant increase in peptides identified with ETD and from fraction 25 ETD outperforms CID where by using only ETD data, 90% of the total number of peptides identifications can be attained. It should be noted that ETD peptide fragment ions originating from 3+ peptides will have a maximum charge of 2+ while CID fragments from a similar precursor will have a maximum charge of 3+ thus making CID spectra potentially more difficult to interpret using an LTQ which has a limited resolving power. Figure 3B shows the cumulative identification of peptides over all SCX fractions revealing that the total number of peptides identified is dominated by CID in the early fractions and by ETD in the later SCX fractions.

The overall result is 6710 spectral annotations (peptide identifications), with a MASCOT score cut-off of 60. CID uniquely idenfied 2309, while ETD obtained 3040 unique identifications and 1361 were identified by

both CID and ETD (see also supplementary Figure 4). Interestingly, in a recent comparative study of CID and ETD Molina et al.(Molina et al., 2008) reported that the use of ETD hardly added to the number of peptide identifications already made through CID, i.e. CID outperformed ETD as far as number of peptide identification were concerned. This apparent discrepancy between these results and ours, wherein ETD identifies more peptides than CID, can be largely explained by the differences in experimental design. In the present work, by using SCX as peptide pre-fractionation technique, a more targeted analysis was performed towards sequencing of highly charged, larger peptides, which we show are more successfully sequenced by ETD. However, observed differences in the outcome of these two experiments may also originate from the different search engines used for peptide identifications (i.e. Spectrum Mill versus Mascot) and different mass spectrometers. Our findings are actually more consistent with the results reported earlier by Good et al. (2007).

When switching to the context of unique protein identification which we based upon the identification of a minimum of two unique peptides, the numbers for ETD, CID and the overlapping group are 327, 277 and 247 respectively (Figure 3C). It is clear that the trend observed for peptides identifications is reflected in protein identifications.

We delved further into the data and determined the basic residue (Arginine, Lysine and Histidine) occurrence in the sequenced peptides with respect to SCX fraction with the expectation to find an increasing number of basic residues in the later SCX fractions. In Figure 4A and 4B the results of this analysis is provided for the peptides identified by CID and ETD, respectively. Although the data for the CID and ETD sets show some resemblance, there are also some significant differences observed. For instance, as expected, ETD fails to identify peptides with no basic residues, contrary to CID which is able to identify a few in SCX fraction 11. Moreover, in agreement with our charge state analysis (see Figure 2), the contribution of peptides with an increasing number of basic residues gradually increases over the SCX run. We observed a marked increase in the appearance of a basic histidine in the identified peptide sequences from SCX fraction 16/17 onwards (see figure 4C), which we believe is responsible, alongside Arginine and Lysine, for the distinctive shift in the peptide net charge and the observed peptide charge within the mass spectrometer found for these

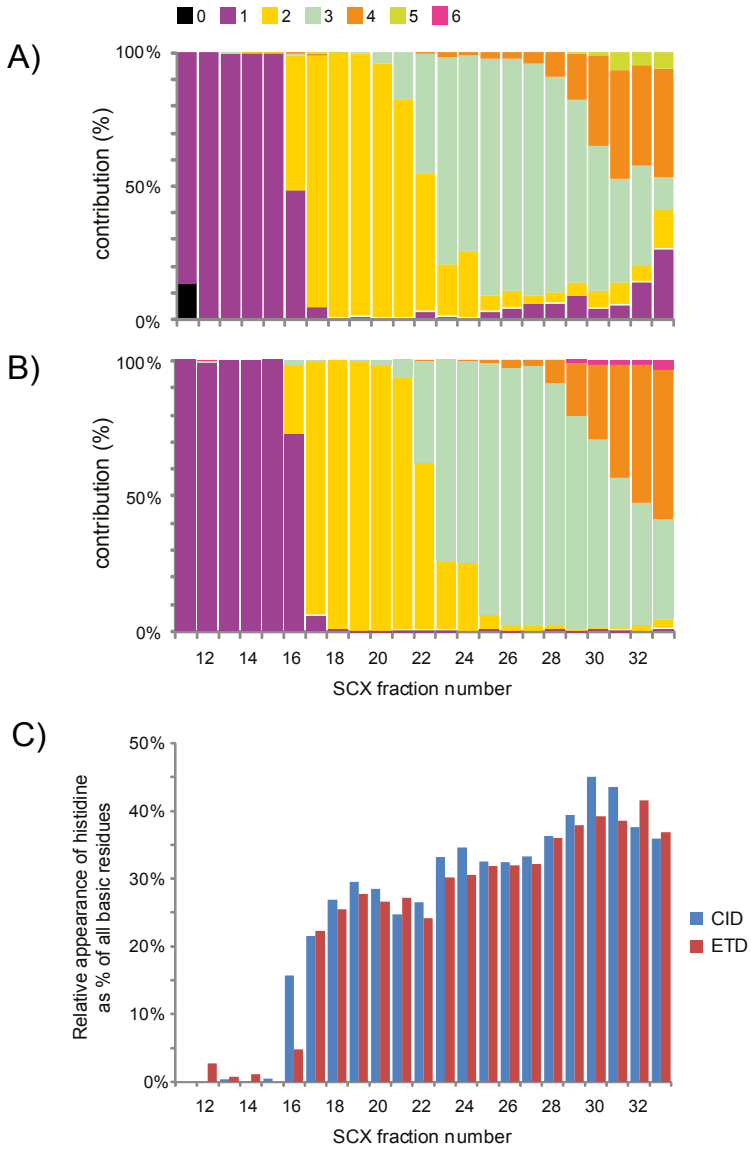


Figure 4: Abundance of basic residues in identified peptides. **(A)** Relative appearance of basic residues in peptides identified in each SCX fraction by CID. The colors code denotes the number of basic residues per peptide. **(B)** Relative appearance of basic residues in peptides identified in each SCX fraction by ETD. **(C)** Relative appearance of histidine basic residues as % of all basic residues in peptides identified in each SCX fraction by CID (blue) and ETD (red). The absolute number of identified peptides in each individual SCX fraction is listed in the supplementary Figure 4.

SCX fractions. The CID data (figure 4A) indicates initially a decrease in the number of peptides with 1 and 2 basic residues and then a steady increase in the number of peptides with 1 and 2 basic residues which is in contrast with the apparent clear separation and observed increase of peptide net charge for the SCX chromatography. These identifications might be false positives or artifacts of the SCX method. It should be noted that in these fractions the total number of identified peptides is rather low, making conclusive remarks about these fractions statistically less valid. Incidentally, results obtained with a lower Mascot score cutoff indicate such peptides are still present in the later fractions however, it can be seen that the clear patterns as found in figure 2A and 2B are gradually lost (Supplementary data figure 3) with decreasing score threshold, indicating that the number of false positives likely increases.

Conclusion

Although suggested previously, it is very apparent, from our data, that CID and ETD identify largely complementary peptide data sets. Whereas smaller, less basic tryptic peptides are ideally sequenced by CID, larger and more basic peptides with higher charges are more readily sequenced by ETD. As shown here, these general characteristics make SCX an ideal pre-fractionation method for the peptides, as this peptide separation is largely based on charge. Our data indicate that the most efficient and simple use of sequencing time can be achieved when a switch is made between CID and ETD from the SCX fractions wherein peptides with 3 or more charges become dominant. The complimentarity of the two techniques in conjunction with SCX is also reflected when observed in the context of protein identifications. When instruments with an inherent higher MS resolution with both CID and ETD capabilities become more readily available(Meiring et al., 2002) decision making on the charge state should be made "on-the-fly", making it easier to decide on the activation method to be used. However, when the peptide separation power in SCX is further improved, so that fractions of different charge states could be baseline resolved, such an on-the-fly assessment of the charge is not really required.

Acknowledgements

We would like to acknowledge Andreas Huhmer and Rovshan Sadygov of Thermo-Fisher Scientific for the use of the Bioworks beta version. We also like to thank Dr. Maarten Altelaar for his valuable comments. This work is supported by the Netherlands Proteomics Centre and the Netherlands Bioinformatics Institute.

References

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature 422, 198–207.

Beausoleil, S.A., Jedrychowski, M., Schwartz, D., Elias, J.E., Villén, J., Li, J., Cohn, M.A., Cantley, L.C., and Gygi, S.P.C.N. (2004). Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc. Natl. Acad. Sci. U. S. A. 101, 12130.

Boersema, P.J., Divecha, N., Heck, A.J.R., and Mohammed, S.C.N.-0057 (2007). Evaluation and Optimization of ZIC-HILIC-RP as an Alternative MudPIT Strategy. J. Proteome Res. 6, 937.

Brunner, E., Ahrens, C.H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E.W., Panse, C., de Lichtenberg, U., Rinner, O., et al. (2007). A high-quality catalog of the Drosophila melanogaster proteome. Nat. Biotechnol. 25, 576–583.

Cargile, B.J., Bundy, J.L., Freeman, T.W., and Stephenson, J.L. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J. Proteome Res. 3, 112–119.

Chen, E.I., Hewel, J., Felding-Habermann, B., and Yates, J.R.C.N.-0103 (2006). Large Scale Protein Profiling by Combination of Protein Fractionation and Multidimensional Protein Identification Technology (MudPIT). Mol. Cell. Proteomics 5, 53.

Good, D.M., Wirtala, M., McAlister, G.C., and Coon, J.J. (2007). Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry. Mol. Cell. Proteomics 6, 1942–1951.

de Groot, M.J.L., Daran-Lapujade, P., van Breukelen, B., Knijnenburg, T.A., de Hulster, E.A.F., Reinders, M.J.T., Pronk, J.T., Heck, A.J.R., and Slijper, M. (2007). Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes. Microbiology 153, 3864–3878.

Kolkman, A. (2005). Double Standards in Quantitative Proteomics: Direct Comparative Assessment of Difference in Gel Electrophoresis and Metabolic Stable Isotope Labeling. Mol. Cell. Proteomics 4, 255–266.

Krijgsveld, J., Gauci, S., Dormeyer, W., and Heck, A.J.R. (2006). In-Gel Isoelectric Focusing of Peptides as a Tool for Improved Protein Identification. J. Proteome Res. 5, 1721–1730.

Leymarie, N., Costello, C.E., and O'Connor, P.B. (2003). Electron Capture Dissociation Initiates a Free Radical Reaction Cascade. J. Am. Chem. Soc. 125, 8949–8958.

Licklider, L.J., Thoreen, C.C., Peng, J., and Gygi, S.P. (2002). Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. Anal. Chem. 74, 3076–3083.

Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., et al. (2002). Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. Proc. Natl. Acad. Sci. U. S. A. 99, 11049.

Liu, H., Sadygov, R.G., John R. Yates, I., Liu, H.B., Sadygov, R.G., and Yates, J.R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201.

MacCoss, M.J., McDonald, W.H., Saraf, A., Sadygov, R., Clark, J.M., Tasto, J.J., Gould, K.L., Wolters, D., Washburn, M., Weiss, A., et al. (2002). Shotgun identification of protein modifications from protein complexes and lens tissue. Proc. Natl. Acad. Sci. U. S. A. 99, 7900–7905.

Meiring, H.D., van der Heeft, E., ten Hove, G.J., and de Jong, A.P.J.M. (2002). Nanoscale LC-MS(n): technical design and applications to peptide and protein analysis. J. Sep. Sci. 25, 557–568.

Mohammed, S., Lorenzen, K., Kerkhoven, R., Van Breukelen, B., Vannini, A., Cramer, P., and Heck, A.J.R. (2008). Multiplexed proteomics mapping of yeast RNA polymerase II and III allows near-complete sequence coverage and reveals several novel phosphorylation sites. Anal. Chem. 80, 3584–3592.

Molina, H., Matthiesen, R., Kandasamy, K., and Pandey, A. (2008). Comprehensive comparison of collision induced dissociation and electron transfer dissociation. Anal. Chem. 80, 4825–4835.

Paizs, B., and Suhai, S. (2005). Fragmentation pathways of protonated peptides. Mass Spectrom. Rev. 24, 508–548.

Pitteri, S.J., Chrisman, P.A., Hogan, J.M., and McLuckey, S.A. (2005). Electron transfer ion/ion reactions in a three-dimensional quadrupole ion trap: Reactions of doubly and triply protonated peptides with SO 2???-. Anal. Chem. 77, 1831–1839.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop And Go Extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. 75, 663–670.

Sadygov, R.G., Hao, Z., and Huhmer, A.F.R. (2008). Charger: Combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. Anal. Chem. 80, 376–386.

Shen, Y., Zhang, R., Moore, R.J., Kim, J., Metz, T.O., Hixson, K.K., Zhao, R., Livesay, E.A., Udseth, H.R., and Smith, R.D. (2005). Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics. Anal. Chem. 77, 3090–3100.

Swaney, D.L., McAlister, G.C., Wirtala, M., Schwartz, J.C., Syka, J.E.P., and Coon, J.J. (2007). Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors. Anal. Chem. 79, 477–485.

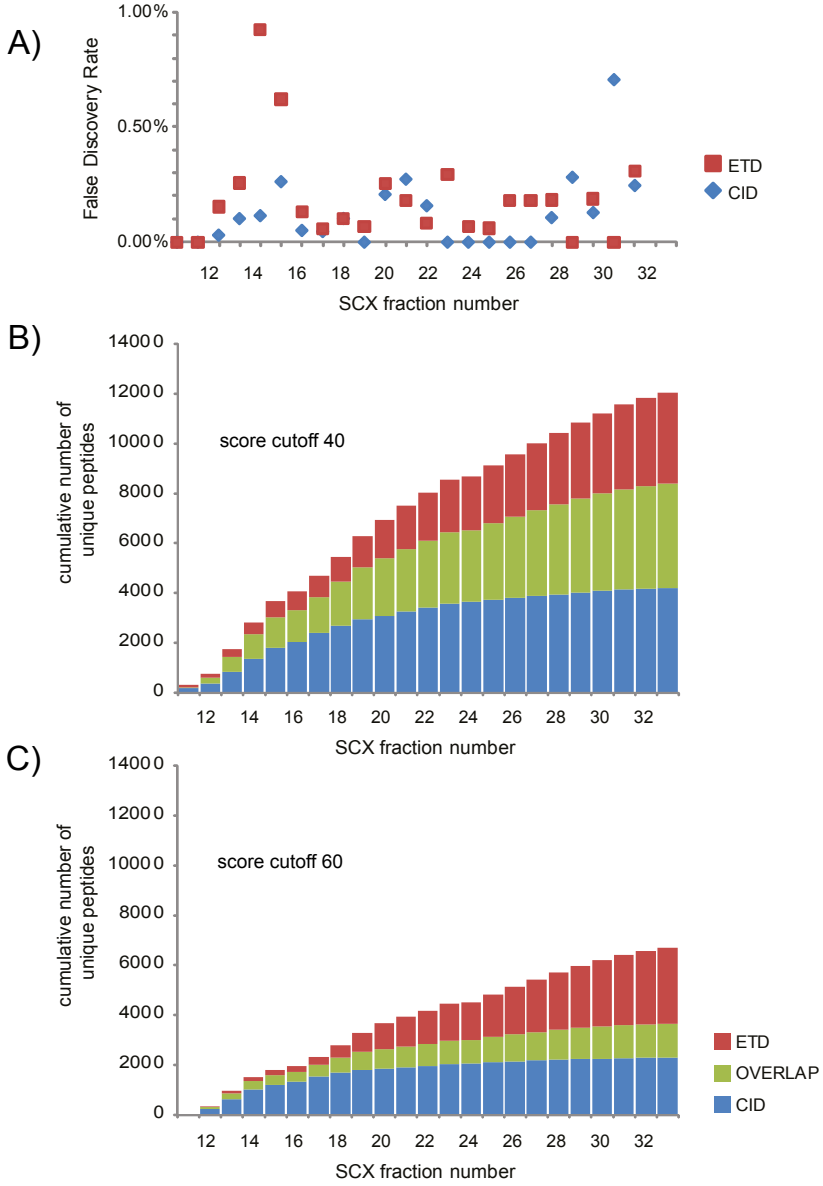
Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 101, 9528–9533.

Witze, E.S., Old, W.M., Resing, K.A., and Ahn, N.G. (2007). Mapping protein post-translational modifications with mass spectrometry. Nat. Methods 4, 798–806.

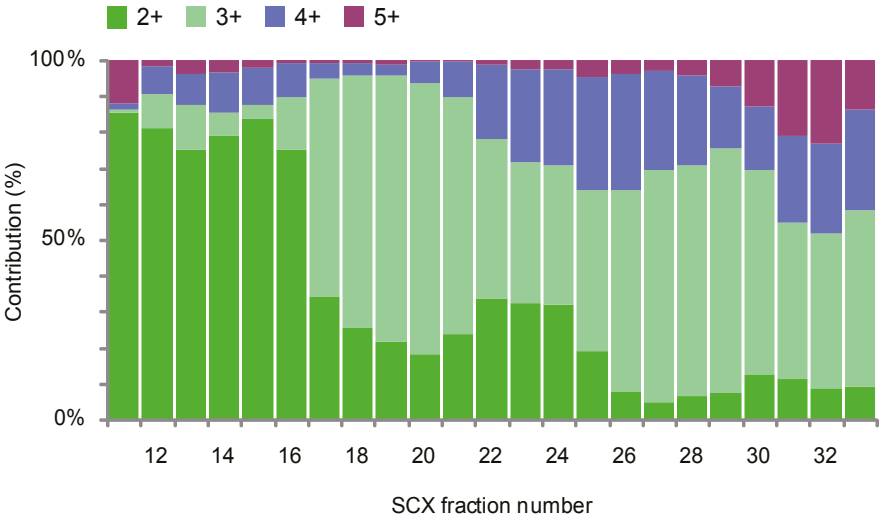
Wu, C.C., MacCoss, M.J., Howell, K.E., and Yates, J.R. (2003). A method for the comprehensive proteomic analysis of membrane proteins. Nat. Biotechnol. 21, 532–538.

Zubarev, R., Kelleher, N.L., and McLafferty, F.W. (1998). Electron capture dissociation of multiply charged protein cations. A …. J. Am. Chem. Soc 120, 3265–3266.

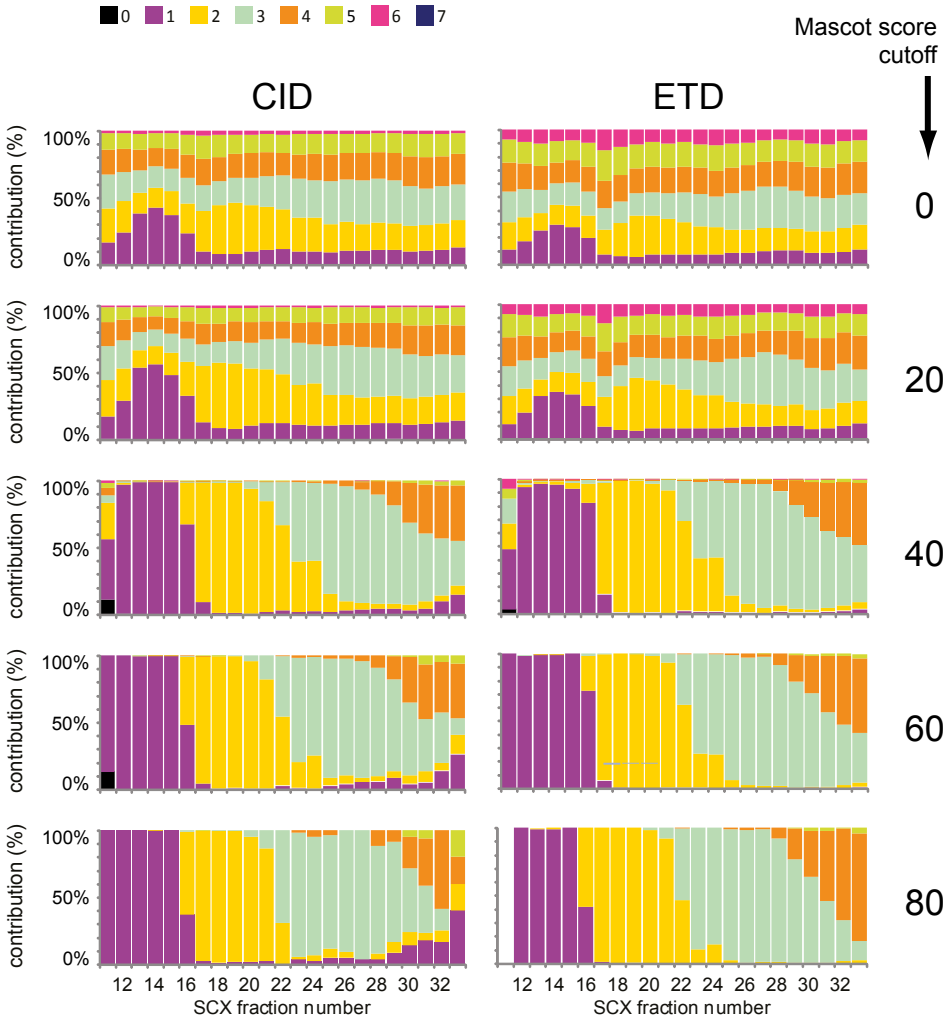
Supplementary figures



Supplementary Figure 1: **(A)** False discovery rates (FDR) as determined by a decoy database search, for every SCX fraction. CID results in blue and ETD in red. **(B)** Number of unique peptide identifications per SCX fraction. The number of peptides exclusively identified with CID are in blue, the number of peptides exclusively identified with ETD are in red, and the number of peptides identified with both activation methods are in green. Mascot cut-off placed at 40. **(C)** Same as B but with a Mascot cut-off of 60.

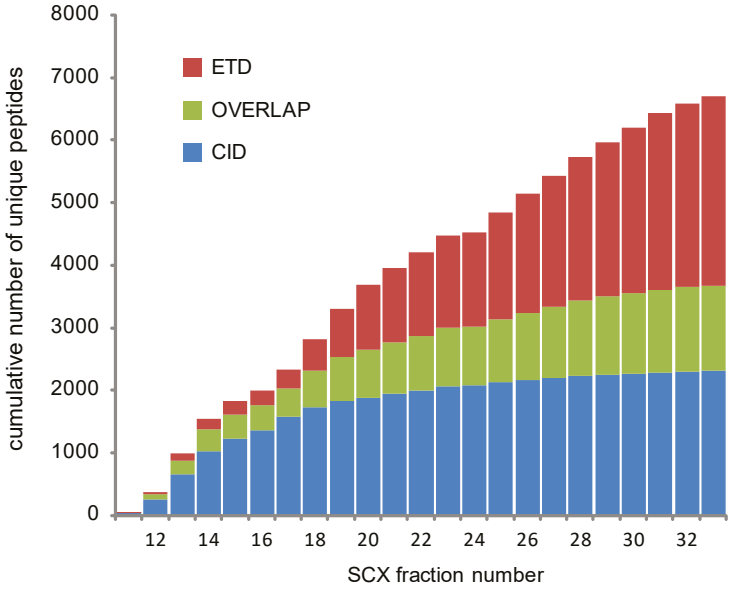


Supplementary Figure 2: Charge state peptide populations for each SCX fraction as determined by the Bioworks 'charger' preprocessor program when the 6+ and 7+ charge states are omitted.



Supplementary Figure 3: Charge state peptide populations for each SCX fraction as determined by utilising Mascot results for both CID and ETD. Each histogram represents the results using a different Mascot cutoff i.e score cut offs of 0, 20, 40, 60 and 80.

fraction	Unique peptides CID	ETD	OVERLAP	Cumulative unique peptides CID	ETD	OVERLAP
11	47	5	4	47	5	4
12	202	15	90	249	20	94
13	402	92	132	651	112	226
14	380	66	113	1031	178	339
15	199	41	42	1230	219	381
16	132	17	11	1362	236	392
17	208	76	60	1570	312	452
18	150	201	135	1720	513	587
19	112	253	119	1832	766	706
20	47	270	66	1879	1036	772
21	59	164	53	1938	1200	825
22	51	138	50	1989	1338	875
23	68	144	65	2057	1482	940
24	21	32	2	2078	1514	942
25	45	190	77	2123	1704	1019
26	33	200	70	2156	1904	1089
27	43	196	54	2199	2100	1143
28	32	196	58	2231	2296	1201
29	21	181	49	2252	2477	1250
30	16	171	44	2268	2648	1294
31	17	173	32	2285	2821	1326
32	18	111	22	2303	2932	1348
33	6	108	13	2309	3040	1361



Supplementary Figure 4: Absolute number of peptide identifications for each individual SCX fraction with a Mascot cutoff score of 60. Adjacent to the table are the sane results are represented graphically (also figure 3B).

5. An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas

Giansanti, P., Aye, T.T., van den Toorn, H., Peng, M., van Breukelen, B., and Heck, A.J.R. (2015). An Augmented Multiple-Protease-Based Human hosphopeptide Atlas. Cell Rep. 11, 1834–1843.

Although mass-spectrometry-based screens enable thousands of protein phosphorylation sites to be monitored simultaneously, they often do not cover important regulatory sites. Here, we hypothesized that this is due to the fact that nearly all large-scale phosphoproteome studies are initiated by trypsin digestion. We tested this hypothesis using multiple proteases for protein digestion prior to Ti^{4+}-IMACbased enrichment. This approach increases the size of the detectable phosphoproteome substantially and confirms the considerable tryptic bias in public repositories. We define and make available a less biased human phosphopeptide atlas of 37,771 unique phosphopeptides, correlating to 18,430 unique phosphosites, of which fewer than 1/3 were identified in more than one protease data set. We demonstrate that each protein phosphorylation site can be linked to a preferred protease, enhancing its detection by mass spectrometry (MS). For specific sites, this approach increases their detectability by more than 1,000-fold.

Introduction

Cellular signaling proceeds largely via cascades of post-translational modifications, in which reversible protein phosphorylation provides a key mechanism (Huang and White, 2008; Rigbolt and Blagoev, 2012). Site-specific protein phosphorylation can be monitored by site-specific phospho-antibodies, such as those raised against, for instance, pERK T202/Y204 and pSRC Y419. Although powerful, there are only a limited number of these antibodies available, hardly sufficient to monitor the more than 100,000 unique phosphosites present in a human cell. Other caveats in using these antibodies are their limited specificity, recognizing multiple sites in a single protein, or similar phospho-sequences in other proteins. Moreover, these approaches are difficult to multiplex for use in high-throughput assays. Mass spectrometry (MS)-based phosphoproteomics has recently surfaced as the method of choice for global and highthroughput protein analysis. Immense progress in both mass spectrometric instrumentation (Hebert et al., 2014; Michalski et al., 2011, 2012) and sample preparation and analysis (Di Palma et al., 2012; Ruprecht and Lemeer, 2014; Wisniewski et al., 2009; Yates et al., 2014) have allowed this technology to confidently identify as many as thousands of proteins and

phosphorylation sites in a single experiment and to accurately quantify changes in protein expression or post-translational modifications (PTMs). Recently, we demonstrated the high potential of a new material for phosphopeptide enrichment (Zhou et al., 2013). This affinity matrix termed titanium (IV)-IMAC (Ti^{4+}-IMAC) was shown to possess high selectivity, sensitivity, and quantification reproducibility, allowing in-depth monitoring of more than 10,000 phosphorylation events by a single-step phosphopeptide enrichment (de Graaf et al., 2014). Predominantly, large-scale phosphoproteomics analyses are based on peptides derived from the tryptic digestion of proteins in a lysate (Mallick et al., 2007; Sharma et al., 2014). Trypsin represents a valid choice because it is highly specific, very effective and, compared to the other available proteases, generates a higher number of peptides in the preferred mass range suitable for identification by MS (Guo et al., 2014; Swaney et al., 2010). As a result, the vast majority of the reported workflows in proteomics are dominated by using exclusively trypsin (Tsiatsiani and Heck, 2015; Wilhelm et al., 2014). However, potentially interesting sequences, and thus particular relevant phosphosites, will remain occluded by this approach, as the tryptic peptides generated do not always possess appropriate physicochemical properties that make them suitable for detection by LC-MS/MS. Although it has been shown that proteome coverage can be increased by using multiple alternative proteases (Bian et al., 2012; Gauci et al., 2009; Guo et al., 2014; Peng et al., 2012; Swaney et al., 2010), this approach has not systematically been investigated for the analysis of PTMs. Here, we report a systematic study using five commercially available proteases to extend the coverage of the phosphoproteome. We targeted the phosphoproteome of human Jurkat T cells stimulated by PGE2 and used Ti^{4+}-IMAC for the enrichment of the phosphopeptides from lysates digested in parallel by AspN, chymotrypsin, GluC, LysC, and trypsin, starting with just 600 mg for each digest. Using a stringently filtered cumulative data set of 37,771 unique phosphopeptides, linked to 18,430 unique phosphosites, we were able to answer the following questions: is it beneficial to do phosphoproteomics using multiple enzymes rather than extending the number of trypsin replicates? Do different proteases possess unique phosphopeptides features? Is the enzymatic efficiency affected by the presence of a phosphorylated site? Is label-free quantitative phosphoproteomics possible when using other en-

zymes than trypsin? Is there at present a detectable bias toward "tryptic" phosphosites in the very large public depositories? The final result of our work is a human phosphopeptide resource, made publicly available, that can be queried by using a simple and intuitive web interface to facilitate the identification of the most-suitable protease for both shotgun as well as targeted MRM/PRM/SWATH-based phosphoproteomics studies.

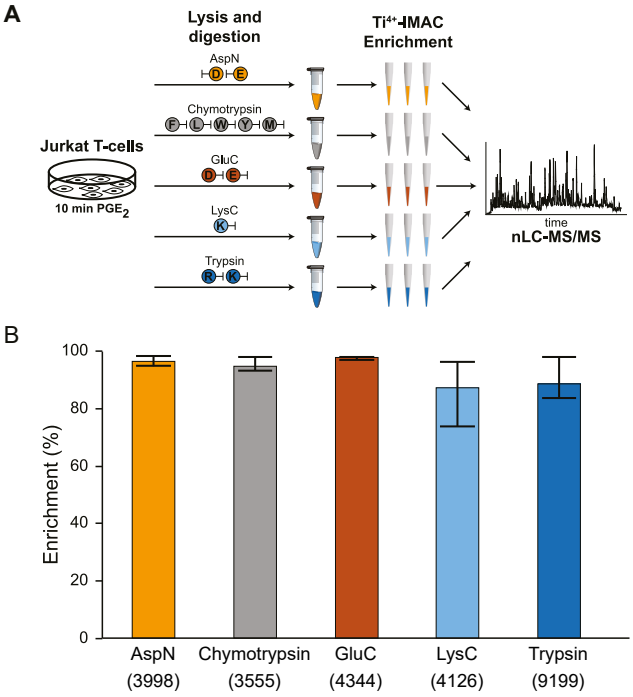


Figure 1. Generation of the Human Phosphopeptide Atlas **(A)** Jurkat T lymphocyte cells were stimulated for 10 min with PGE2, lysed, and digested in parallel with one of the five proteases: AspN, chymotrypsin, GluC, LysC, and trypsin. Each digest was divided over three Ti^{4+}-IMAC phosphopeptide enrichment columns, loading 200 mg per enrichment, resulting in 15 samples that were analyzed by nLC-MS/MS. The color scheme representing the data on each individual protease is kept consistent throughout the manuscript. These colors are AspN, orange; chymotrypsin, gray; GluC, vermillion; LysC, sky blue; and trypsin, blue. **(B)** Enrichment efficiency and total number of unique phosphosites detected. The bar chart represents the median enrichment efficiency calculated as the percentage of phosphopeptides in the six MS runs, demonstrating the high selectivity (typically above 90%) of the Ti4+-IMAC enrichment for all used proteases. For each population, the whiskers represent the minimum and maximum selectivity. The number of unique phosphosites identified in the data sets originating from each protease is reported underneath the bars.

Results and Discussion

Generation of the Data for an Augmented Human Phosphopeptide Atlas
All phosphoproteomics data presented here were obtained from human Jurkat T cells that were harvested following 10 min of PGE2 stimulation. PGE2 increases the levels of cAMP, thereby activating among other the cAMP-dependent kinase PKA. We did choose PGE2-activated T cells, as recently a rather large trypsin-based phosphoproteomics data set was made available for this system using a similar experimental approach (de Graaf et al., 2014). The activated T cells were lysed and extracted proteins were digested with each of the five enzymes, i.e., AspN, chymotrypsin, GluC, LysC, and trypsin, using identical protocols as described in detail earlier (Low et al., 2013). Three independent Ti^{4+}-IMAC enrichments were used for each of the five digests to specifically enrich for phosphopeptides, and only 200 mg of digest material were used as input per enrichment (Figure 1A). All 15 samples were analyzed in replicate by nLCMS/ MS using 150-min gradients and a decision-tree-driven peptide fragmentation scheme, selecting the appropriate activation technique being either CID or ETD, depending on the precursor charge state and m/z (Frese et al., 2011; Swaney et al., 2008). In total, more than 1.1 3 106 MS/MS spectra were acquired (Table 1), resulting in 37,771 unique phosphopeptides originating from 5,326 phosphoproteins. A complete list of all the identified peptide to spectrum matches (PSMs), (phospho)peptides, including confidence scores for identification and site localization for each enzyme, can be found in Table S1 and on the web-based resource page (http://phosphodb.hecklab.com). Among the five data sets, trypsin comprised the largest number of unique phosphopeptide identifications (13,476), followed by GluC (7,502), LysC (6,595), AspN (6,407), and chymotrypsin

Table 1. Unique Phosphopeptide and Phosphoprotein Identifications in Each of the Protease Data Sets

Protease	Trypsin	LysC	AspN	GluC	Chemotrypsin	Cumulative
No. MS/MS	248,753	224,227	204,657	208,559	225,911	1,112,107
No. PSMs	87,650	59,104	45,095	67,808	40,064	299,721
Success rate (%)	35.2	26.4	22.0	32.5	17.7	26.9
No. CIDs	64,970	24,523	19,438	28,714	22,270	159,915
No. ETDs	22,680	34,581	25,657	39,094	17,794	139,806
CID/ETD ratio	2.86	0.71	0.76	0.73	1.25	1.14
No. phosphopeptides	13,476	6,595	6,407	7,502	5,376	37,771
No. phosphoproteins	3,519	1,868	2,021	2,225	1,936	5,326

Although the number of MS/MS events and phosphopeptide enrichment efficiency are alike in all digests, the number of unique phosphopeptides detected is twice as high for trypsin as for the other proteases. ETD is highly beneficial and complementary in the identification of phosphopeptides, especially in the digest of LysC, AspN, and GluC.

(5,376). Comparison of Data Sets Originating from Different Proteolytic Digests As Ti^{4+}-IMAC has been so far only used for the enrichment of phosphopeptides from tryptic digests, we first assessed the feasibility and validity of using this material for the enrichment of non-tryptic digests. The data from three independent Ti^{4+}- IMAC enrichments clearly indicate a high selectivity (Figure 1B), ranging from on average 86% for LysC up to 98% for GluC, similar to the enrichment efficiency obtained for trypsin (90%). Thus, the selectivity for enrichment by Ti^{4+}-IMAC seems protease independent. We next examined the number of phosphorylation sites identified by each protease (Table 1; Figure 1B). Using the same amount of sample (600 mg), we recovered 9,199 unique phosphorylation sites (from 13,476 unique phosphopeptides) in the tryptic digest; whereas, around 4,000 unique phosphosites sites could be recovered from each of the other digests. All these numbers are quite favorable (considering the low amount of sample used) but also clearly show that trypsin seemingly outperforms the other enzymes in number of identified sites. Such a lower efficiency for GluC, AspN, LysC, and chymotrypsin is in agreement with what has been reported at the unmodified peptide level (Low et al., 2013; Swaney et al., 2010). However, we hypothesize that there may be multiple origins for this observation. At least part of the higher number of identifications for trypsin can be attributed to the higher identification rates (PSMs), likely due to a positive bias of search engines toward trypsin (Granholm et al., 2014). In addition, superior fragmentation of tryptic peptides, especially in CID/HCD, attributes further to the higher number of identification for trypsin (Meyer et al., 2014; Swaney et al., 2010). In line with this notion, our data show that the use of ETD fragmentation complementary to CID/HCD fragmentation is substantially more beneficial for digests of LysC, AspN, and GluC and not so much for trypsin (see Table 1). Next, we evaluated the unique phosphorylation sites per protease data set and between the five data sets. To compare the data sets, we applied stringent site assignments criteria by using the phosphoRS algorithm (Taus et al., 2011), which allowed to confidentially localize the phosphorylation event with amino acid resolution. The resulting data set reveals that, by using different proteases, we can significantly enhance our phosphopeptide atlas, as the five data sets turned out to be extremely complementary. As illustrated in A, we cumulatively identify 18,430 unique phosphosites, albeit that just

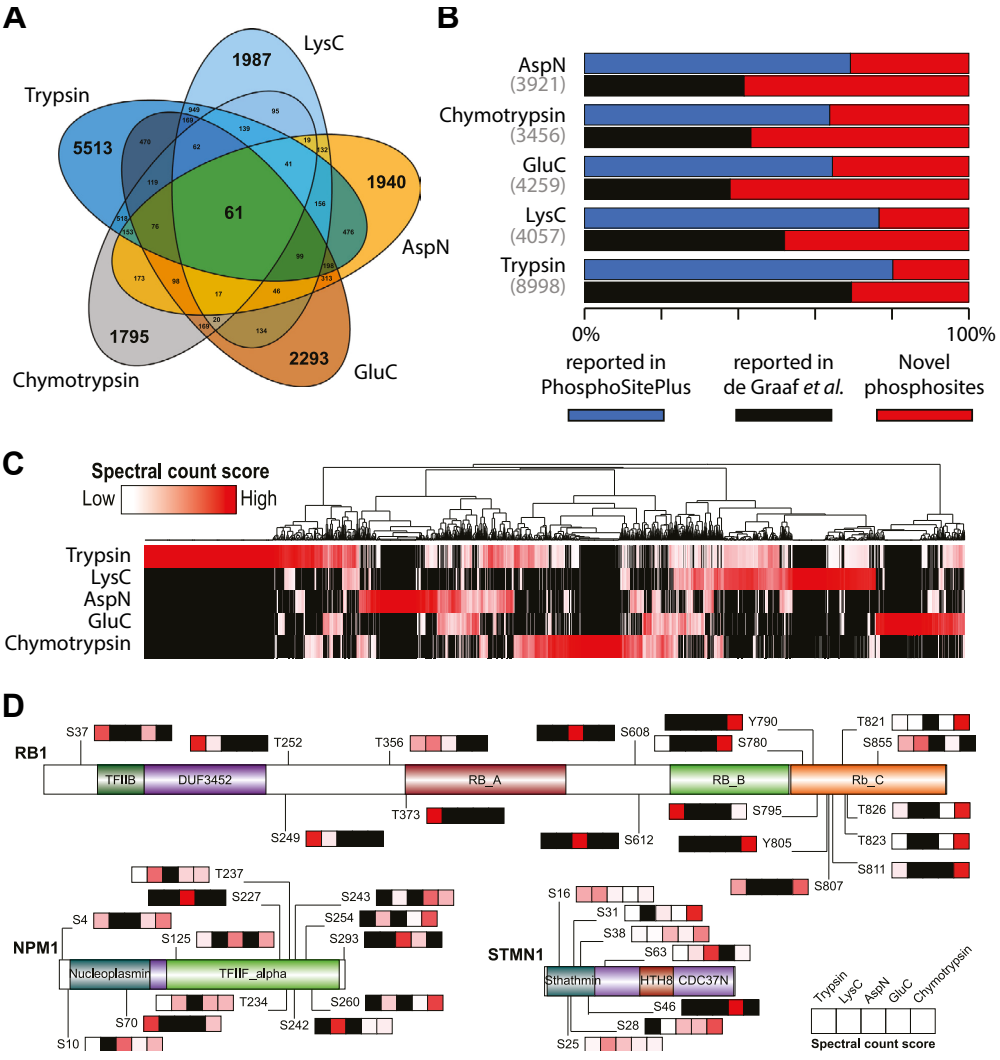
about 27% of these were identified in more than one protease data set and only a marginal fraction of 0.3% (i.e., 61) could be identified in all five data sets.

Characterization of the Identified Phosphopeptides Reveals that Phosphorylation Induces Widespread Protease Missed Cleavages

To gain more insight into the complementarity observed among the five data sets, we first evaluated the global physico-chemical characteristics of the phosphopeptides. First, we examined the length distribution of the identified phosphopeptides (Figure S1A), which revealed that they possess on average similar length distributions, with most of them having a length in between 14 and 25 amino acids. Not surprisingly, identified phosphopeptides generated by trypsin digestion are on average relatively shorter (7–20 amino acids long) compared to phosphopeptides identified after GluC digestion, which generates longer sequences (14–30 amino acids long). Although the results generally correlated with data reported in other studies for unmodified peptides (Biringer et al., 2006; Kalli andHa° kansson, 2010;Swaney et al., 2010),we noticed that there was a substantial shift in the observed average peptide length for trypsin, from_10 forunmodified peptides to_18 amino acids for phosphopeptides. This behavior has been attributed by the presence of a phosphorylation site in close proximity to the sites of proteolytic cleavage, which prevent trypsin and LysC cleavage and neighboring sites (Molina et al., 2007). To explore whether a similar phenomenon exists in the other protease data sets, we investigated the frequency of missed cleavage events in all five data sets. As depicted by the violin plots in Figure S1B, relative high number of missed cleavages is present in the phosphopeptide data sets from each proteolytic digest. If the enzymatic digestions were inefficient, one would expect unmodified peptides to show the same distribution.On the contrary, the analysis on the unmodified peptides, co-purified by the Ti^{4+}-IMAC material, reveal a more efficient digestion, similar to the one commonly observed for the same enzymes in large-scale proteome studies (Low et al., 2013). Thus, the here-generated data set reiterates that the number of missed cleavages is significantly higher for phosphorylated peptides, not only in digest generated by trypsin and LysC but also in AspN, GluC, and chymotrypsin digests. To

Figure 2 (facing page). Qualitative and Quantitative Benefit of Using Multiple Proteases in Phosphoproteomics **(A)** Venn diagrams displaying the overlap in detected unique phosphosites between the data sets generated by Trypsin, LysC, AspN, GluC, and Chymotrypsin. Nearly 3/4 of the phosphosites (i.e., 13,500) were not detected by more than one protease, indicative of the high orthogonality of the multi-protease strategy. **(B)** Benchmarking the unique and distinct phosphorylation sites detected in each digest to the human phosphorylation sites reported in the comprehensive PhosphoSitePlus database depository (153,900 entries) and the related large Jurkat (phospho)proteome data set reported earlier by de Graaf et al. (2014; 16,200 entries). The here-identified phosphopeptide sites in the AspN, GluC, and chymotrypsin digests are clearly underrepresented both in the public depository as well as in the largest reported Jurkat cell (tryptic) phosphoproteome. These comparisons clearly reveal a substantial tryptic bias in public depositories. **(C)** Heatmap, based on spectral count scores, illustrating the contribution of each protease in the detection of a particular phosphosite. Black color means not detected. Phosphosites were further grouped using hierarchical clustering (distance metric was Euclidean correlation and linkage method was average). **(D)** Domain structures for three representative phosphoproteins (RB1, STMN1, and NPM1), bearing multiple phosphosites, whereby by using bars is presented how well they are detectable in the different digests. Black color means not detected. Residue numbers of the identified phosphorylation sites are indicated.

gain closer insights into the influence of phosphorylation on digestion, we probed the composition of the amino acid sequence surrounding the missed cleavage site for the phosphopeptides detected in each data set (Figure S1C). For each data set, the residue corresponding to an uncleaved site was centered and the distance to the nearest localized phosphoresidue, within ±5 positions, was calculated. If phosphorylation has no influence on the cleavage, an even distribution of the phosphoresidue should be observed around the missed cleaved bond. In accordance with previous studies (Dickhut et al., 2014; Gershon, 2014), this analysis shows for trypsin and LysC a phosphoresidue in the positions +1, +2, and +3 selectively hampers cleavage. For AspN and GluC, phosphorylation impedes cleavage mostly when located at position_2, whereas for chymotrypsin, a phosphorylation at the +1 position hampers protease activity. The negative correlation between cleavage efficiencyand nearby phosphorylation events is thus generally protease independent, albeit that specific protease-dependent rules govern the most-affecting sites. We also determined the charge distribution of the phosphopeptides in each data set. As illustrated in Figure S1D, trypsin digestion generates, predominantly, doubly and triply protonated peptides. In contrast, LysC, AspN and GluC generated



substantially more triply and higher charge state phosphopeptides. Notably, even though the charge and length properties of the chymotryptic peptides are very similar to the ones generated by trypsin, the success rate in identification for chymotrypsin was a factor 2 lower (Table 1). This can be partly explained by the low specificity of chymotrypsin, which may cleave at fivedifferent amino acids, posing a bigger challenge to current search engines. Despite this lower efficiency, the added value of using chymotrypsin is illustrated by the 1,795 uniquely identified phosphosites (Figure 1A). We think this can be explained by the rather unique capabil-

ity of chymotrypsin to cleave peptide bonds consisting of amino acids with hydrophobic (L and M) and aromatic large side chains (F, W, and Y). Therefore, we hypothesize that, by using chymotrypsin, more peptides from hydrophobic protein regions can be retrieved, often missed in tryptic digests. Tryptic Bias in Public Phosphopeptide Depositories We evaluated the here-identified 18,430 phosphorylation sites against two very large public data depositories (Figure 2B). First, we matched the phosphosites identified in the data sets of each protease against one of the largest manually curated database repositories: PhosphoSitePlus (Hornbeck et al., 2012; ~153,900 human phosphorylation sites), which includes results from large-scale phosphoproteomics experiments but also from lower-throughput assays based on immune purification using phospho-specific antibodies, cumulating data from different cellular and/or tissue origins. We observed that our multiple-protease- based data sets confidently identified 6,032 phosphorylation sites not yet reported in PhosphoSitePlus. Although our trypsin data set yielded also a number of not yet reported sites, more than 70% of the novel sites came from the data sets using the other proteases. Based on the fact that they have similar sizes, a more-fair benchmark we hypothesized would be to compare the current data sets to the recently (by us) reported exhaustive trypsinbased phosphoproteomics data set acquired from the same Jurkat T lymphocyte cells by using the same enrichment strategy, albeit by using 54 distinctive Ti^{4+}-IMAC-based enrichments (~16,200 phosphorylation sites; de Graaf et al., 2014). In that earlier experiment, we likely approached the maximum in detection of tryptic phosphosites by doing 54 replicates across six time points, instead of the three used here. Strikingly, this analysis (Figure 2B) revealed that many phosphosites found by using the alternative proteases were not detected earlier in the exhaustive trypsin-based experiments. In some of these data sets, this accumulated to close to 2/3 of the data set, i.e., AspN (59%), chymotrypsin (57%), and GluC (62%). From these data, we conclude that, although trypsin performs very well, it provides a biased representation of the phosphoproteome. This bias can be substantially reduced, making use of the orthogonality of the alternative proteases applied here. Quantitative Assessment of Protease Bias for Individual Phosphoproteins: Impact on Studying Cellular Signaling In absence of any protease bias, the five proteases should equally contribute to the detection of a par-

ticular phosphosite on a given protein, especially for the _3,300 phosphosites for which high numbers of spectra (R10) were successfully and unambiguously matched to a peptide sequence. However, when we evaluated each protease contribution by using a spectral count score (see Data Analysis section), we observed massive differences (Figures 2C and S2). Many phosphosites were either highly overrepresented or underrepresented in particular protease data sets. Consequently, certain protein phosphorylation events are considerably better detectable by using one or two proteases over the others. To further illustrate this phenomenon, we focused on a group of key signaling phosphoproteins, e.g., a list of reported well-known oncogenes (Table S2). In signal transduction, reversible protein phosphorylation often represents an activating or deactivating switch for protein activity, which can play an important role in the pathogenesis of human cancers (Chong et al., 2008). Therefore, the identification and characterization of the phosphorylation events associated to oncogenic signaling are particularly important. Our data clearly show that, even though trypsin is an efficient and robust protease that generates the highest number of (phospho) peptides detectable by MS experiments, it may not always be the optimal choice for specific important regulatory sites. We further zoom in on three illustrative examples from the full list of proteins given in Figure S2, namely, the retinoblastoma protein (RB1), stathmin (STMN1), and nucleophosmin (NPM1) (Figures 2D and S3). Complementary, the web-based database set up to make our data publicly available (http://phosphodb.hecklab.com) provides similar analysis on all the here-reported phosphoproteins. The retinoblastoma protein is an important tumor suppressor protein that is dysfunctional in several major cancers (Murphree and Benedict, 1984). The hypo-phosphorylated form interacts with and sequesters the E2F1 transcription factor, leading to cell cycle arrest. On the contrary, the hyper-phosphorylated form is unable to interact with E2F1 and, therefore, unable to restrict progression from the G1 to the S phase of the cell cycle. The main regulators of RB1 activity are several members of the cyclin-dependent kinases (CDKs) family, which can phosphorylate RB1 at several different sites, such as S249, S252, S807, S811, T821, and T826 (Knudsen and Wang, 1997). Evidently, in ideal phosphoproteomics experiments, all these sites should be monitored. For three of these sites, our data clearly reveal that chymotrypsin is the pro-

tease that would substantially facilitate their detection by MS (Figures 2D and S3). Another example is provided by STMN1, a protein involved in the biogenesis and remodeling of the cellular microtubule cytoskeleton. STMN1 activity depends on phosphorylation of at least four key serine residues (S16, S25, S38, and S63) by different kinases (Santamarı´a et al., 2009). Our data show that, whereas S16, S25, and S38 could be detected by each protease, S63 clearly possesses a strong preference in detection by using AspN. This suggests that S63 would be underrepresented in conventional tryptic-based phosphoproteomics approaches. Indeed, when we evaluated these four sites in the PhosphoSite-Plus database, we discovered that S16, S25, and S38 were detected by over 100 large-scale mass-spectrometry-based experiments, whereas S63 has been reported in just 50 data sets. Moreover, in the latter case, these observations originated mainly from experiments in which a targeted low-throughput approach had been used, i.e., by using motif-specific antibodies directed against kinase motifs (e.g., PKA, PKC, and PKD). Also, STMN1 S31 phosphorylation was readily detected with over hundreds of PSMs in the chymotrypsin and AspN data sets, but only six PSMs were detected in the much-larger tryptic data set. Validating our trypsin bias hypothesis, this site turned out to be also underrepresented in PhosphoSitePlus. As third example, we highlight nucleophosmin. NPM1 is overexpressed, mutated, and chromosomally translocated in many tumor types (Falini et al., 2007). During the different phases of the cell cycle NPM1 can be phosphorylated at multiple sites by PLK or CDK kinases, modifications that have been proposed to trigger or abolish NMP1 functions. In our data, we detect a high number of PSMs for S4, S10, S70, S125, T234, T237, and S242 and lower levels of PSMs for S260, S227, S293, S243, and S254 (Table S2). However, of these S10, T234, T237, and S242 are hardly detectable in the tryptic data set, whereas these sites have high numbers of PSMs in the AspN and chymotrypsin data sets (Figures 2D and S3). GluC digestion seems to be really beneficial for T234 and T237, whereas S70 is only detected with high PSMs in the data from the tryptic and chymotryptic digests. In summary, our data on nucleophosmin reveal that each site in this protein has a preference in detectability correlated with different proteases (Figure S3). Moreover, only by using all five proteases, we are able to map all known important regulatory phosphosites in this protein. These three proteins illustrate

the protease complementarity, which holds true for all the other phosphoproteins detected and listed in Table S2. Clearly, when the phosphorylation status of a given protein needs to be monitored, the use of complementary proteases is highly beneficial. Therefore, targeted proteomics studies should be expanded to include non-tryptic peptides, as intensities likely increase by factors between 10 and 10,000. Presently, a number of factors limit the usefulness of these non-tryptic phosphopeptides for systems biology studies. First, the public available repositories are nearly exclusively based on tryptic peptides, underrepresenting many potential interesting sites. Second, spectra are not always accessible to assess the quality of phosphopeptide identification and the accuracy of the phosphorylation site assignment. Moreover, the data are mostly presented as lists of phosphosites, whereas information at the phosphopeptide level might provide more useful information for further analysis such as targeted MRM/PRM/SWATH. To enhance the utility of our augmented phospho-

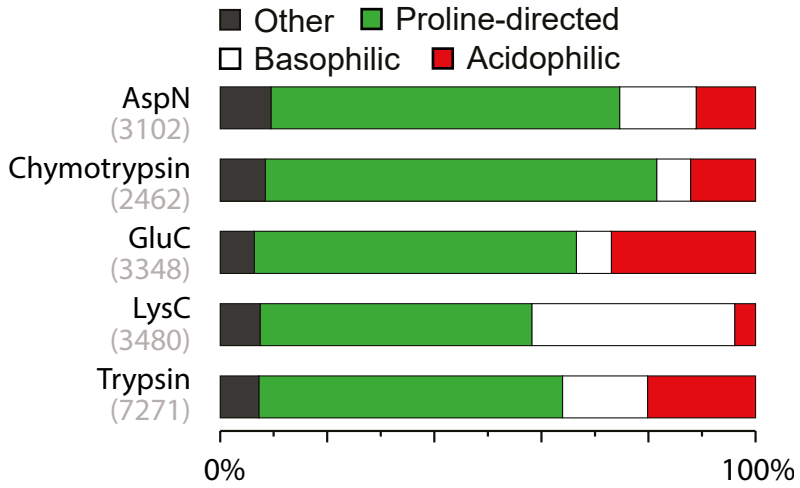


Figure 3. Distribution of the Sequence Motifs Extracted from the Protease-Linked Phosphorylation Data Sets. The motif-x algorithm revealed more than 130 unique phosphorylation motifs in the phosphopeptides data sets. They could generally be classified into proline-directed, basophilic, and acidophilic motifs (http://www.hprd.org). The contribution of each class of motifs is given, and the total number of phosphosites that could be assigned to these motifs is shown in brackets. Most notably, whereas for trypsin, AspN, and chymotrypsin the ratio between acidophilic and basophilic motifs is close to 1:1, GluC and LysC show a clear bias toward the acidophilic and the basophilic motifs, respectively. Additionally, compared to the other proteases, chymotrypsin seems to be biased toward proline-directed motifs.

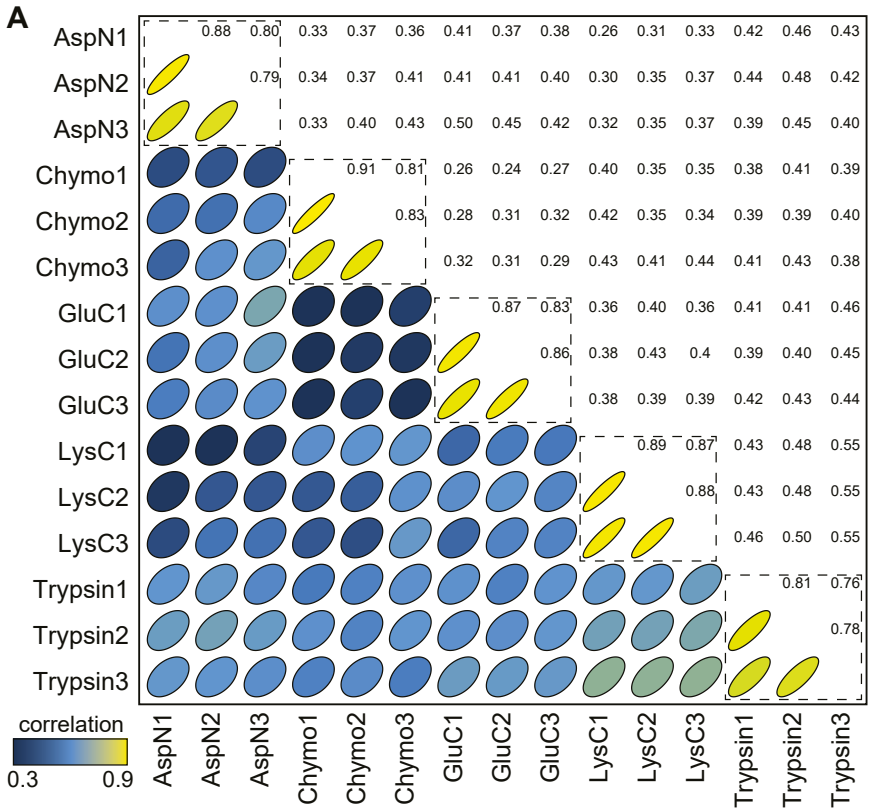
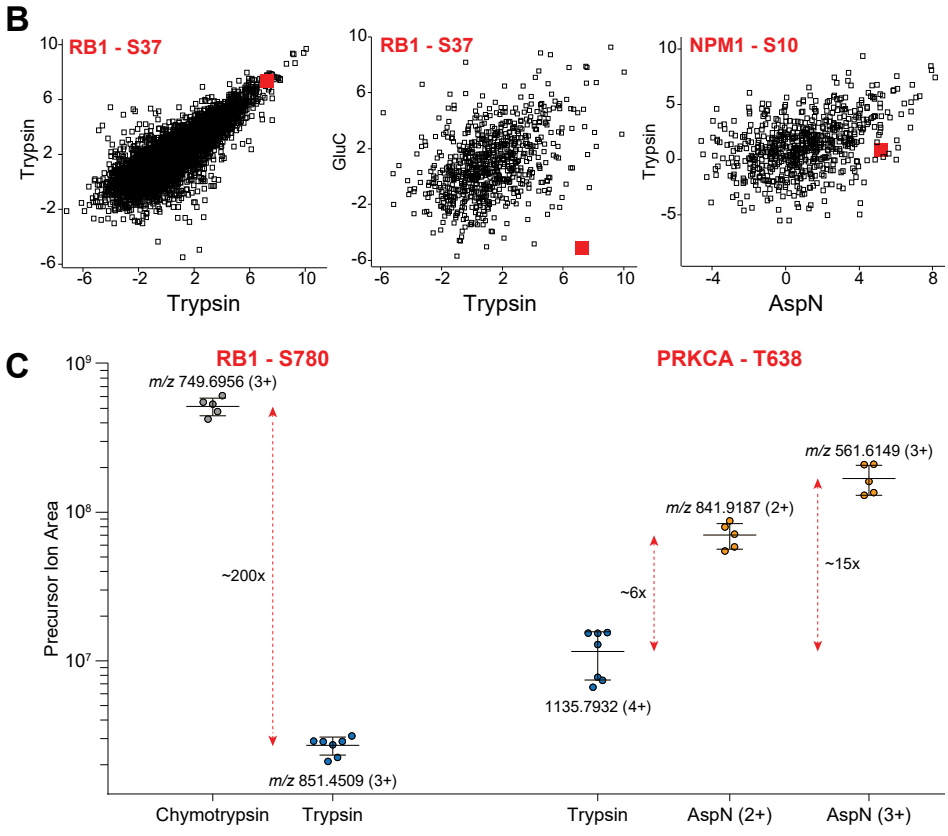


Figure 4. Label-free and Targeted Phosphoproteomics Using Complementary Proteases **(A)** Pearson correlation matrix of all performed experiments reveal a high correlation in phosphosite intensity when data sets are obtained following digestion by the same protease (dashed squares; r > 0.8) but a low correlation between data sets originating from different proteases (r 0.25–0.55). (B) Illustrative scatter plots depicting phosphositebearing peptide intensities (log base 2) in digests obtained by different proteases. The plots are extracted from Figure S5. Two phosphosites are highlighted in red, showing a clear bias in intensity toward detection by trypsin (RB1-S37) or chymotrypsin (NPM1-S10) in this label-free intensity based analysis. (C) Also, a targeted phosphoproteomics experiments on specific phosphorylation sites shows a clear bias in average precursor intensity (error bars, ±SD) toward detection by chymotrypsin (RB1-S608) and AspN (PRKCA-T638).

peptide atlas, we provide our data as a searchable resource for targeted phosphoproteomics experiments (http://phosphodb.hecklab.com). The users can browse through thephosphorylation sites on their proteins of interest, applycustom filters, visually identify which protease would be the



most suitable for the detection of the sites of interest, and export the necessary parameters to setting up the MRM/PRM-like assays. Our resource provides a first draft of a less-biased human phosphopeptide atlas that may be further expanded by importing data from other proteases, other cell lines, and tissue. Moreover, we believe that this resource, which also includes all raw data and MS/MS spectra interpretations, may help bioinformaticians in their efforts to develop and improve computational prediction and analysis tools for non-tryptic peptides in phosphoproteomics analysis, which we here identified as being one of the current likely bottlenecks. Computational Analysis on the Identified Phosphorylation Sites. To assess a potential preference for certain kinase motifs in the five phosphopeptides data sets, we subjected the hereidentified phosphorylation sites first to the motif-x algorithm (Schwartz and Gygi, 2005). A total of 132 distinct motifs could be defined among all the five proteases data

sets (Table S3). By convention, we could classify these kinase motifs into four well-defined groups: acidic, basophilic, proline-directed, and "other" motifs (Figure 3). This analysis revealed that the sequences bearing the [pS/pT]P motif, target of the very large family of proline-directed kinases including, among others, CDKs and MAPKs (Lu et al., 2002), were almost evenly spread across all the five proteases data sets, with just AspN and chymotrypsin possibly displaying a slight overrepresentation. Considering next the substrates of basophilic and acidophilic kinases, a clear bias is observed in between the data sets, with especially LysC and GluC being specific outliers (Figure 3). Motifs in the basophilic group convolute around the PKA/Akt/PKC [R/K]X[R/K]XX[pS/pT] consensus motif (Pearce et al., 2010). Many of the arginine-containing sequences were retrieved from the tryptic digests, whereas lysine-rich motifs were mainly present in the LysC data set, making these two proteases the most suitable when the interest lays in the analysis of basophilic kinases substrates. Likewise, we hypothesized that AspN and GluC would be the protease of choice in detecting substrates of acidophilic kinases (e.g., CK1, CK2, and PLKs), which mostly target serine and threonine residues flanked by acidic residues (Amanchy et al., 2007). Our analysis revealed that only GluC showed a clear bias toward detecting peptides bearing an acidophilic motif. All the identified phosphorylation sites were further subjected to the NetworKIN algorithm (Horn et al., 2014) to predict the kinases involved (Table S3). This analysis predicted that phosphorylation of the acidophilic motifs was mostly associated to the kinases CK1, CK2, and SGK1. These three kinases were mainly overrepresented in the trypsin data set. In agreement with above basophilic kinases, members of the AGC and CAMK groups were predominantly enriched in the trypsin and LysC data sets. Interestingly, the NetworKIN analysis also revealed a bias toward LysC for almost all members of the GRK family (Figure S4). GRKs are activated downstream of GPCRs, which we here activated by PGE2, and function to phosphorylate and control the activity of these receptors. Unlike most AGC kinases, for GRKs phosphorylation, the consensus motifs are not clearly defined yet. It has been reported that some members prefer acidic residues flanking the phospho-acceptor sites whereas other members favor basic residues (Pitcher et al., 1998). Indeed, a closer inspection of the phosphosites detected in our data sets confirmed this binary characteristic, with lysine,

glutamic acid, or both residues in close proximity to the phosphorylated serine/threonine. A Protease Bias in Label-free Quantification of Phosphorylation Sites Due to their highly dynamic spatial regulation, it is crucial to perform quantitative studies on protein phosphorylation to fully understand the signaling networks controlling cellular fate. Many different strategies have been developed to accomplish this task (Macek et al., 2009). Recently, others and we demonstrated the possibility of performing in-depth and highthroughput reproducible label-free phosphoproteome quantification (Courcelles et al., 2013; de Graaf et al., 2014; Montoya et al., 2011; Soderblom et al., 2011). Here, we assessed the feasibility of obtaining similar quantitative data when proteases other than trypsin are used, making it possible to profile phosphosites that would be easily missed in a conventional trypsinbased workflow. Such an analysis would provide insight into the reproducibility of the enrichment protocol but also the reproducibility in digestion by the different proteases. To minimize any potential bias of the different algorithms used to perform identification and quantification, we processed the here-obtained data through the same computational pipeline as previously described (de Graaf et al., 2014). The data from three independent enrichments within each of the distinct protease data sets show a high quantitative reproducibility across all phosphosites, with a correlation typically around 0.85 (Figures 4A and S5). Therefore, we conclude that quantitative phosphoproteomics, including label-free quantification, is possible, not only by using trypsin, as shown before (de Graaf et al., 2014; Montoya et al., 2011), but also by using any of the other four alternative proteases used here. However, when we compare data generated by using more than one protease, the correlation is low (r ~0.25–0.55). These lower correlations further endorse the previously demonstrated substantial protease bias in data sets generated in yeast on non-modified peptides (Peng et al., 2012). In Figure 4B, a zoom in of some of these correlation plots is given, highlighting a few illustrative phosphosites in which one protease clearly outperforms the other, resulting sometimes in a more than 1,000-fold higher phospho-site bearing peptide intensity. As an example, S37 on RB1 is easily detected as a high abundant peptide following trypsin digestion. In the GluC digest, its intensity is much lower, affecting the accuracy of the peak detection. Similarly, the tryptic peptide bearing S10 on NPM1 is low abundant, whereas the AspN

peptide bearing this S10 site is about 4-fold more intense. We believe that these results support the observation that the best peptides are not necessarily tryptic (Peng et al., 2012), highlighting the valuable contribution of non-tryptic peptides also in quantitative phosphoproteomics analysis. To further demonstrate that some phosphorylation sites can be hardly observable in tryptic digests, we conducted a targeted proteomics experiment making use of a well-defined inclusion list (Jaffe et al., 2008; Schmidt et al., 2008). With this approach, the mass spectrometer is specifically instructed to sequence only pre-selected ion species, overcoming the stochastic undersampling that may occur during LC-MS/MS analysis. As a proof of principle, we selected two phosphorylation sites with a well-known biological significance, S780 on RB1, which is one of the residues modulating the activity of RB1 during the cell cycle, and T638 on PRKCA, one of three residues needed to be phosphorylated for full activation of this kinase. We hypothesized that the two phosphosites would be easier to assay using chymotrypsin and AspN, respectively, rather than trypsin because of the observed differences in spectral count values for the candidate phosphopeptides. Equal starting amounts of Jurkat cells lysate were digested using AspN (five replicates), chymotrypsin (five replicates), or trypsin (seven replicates). Each digest was subjected to phosphopeptide enrichment as described above, and resulting phosphopeptides were then analyzed by using the inclusion-list-based approach. As depicted in Figure 4C, a difference in peak area of ~200-fold was observed for S780 on RB1. Alike for T638 on PRKCA, the use of AspN led to an _20-fold more-intense phosphosite-bearing peptide when compared to its tryptic counterpart. To demonstrate that this difference is not caused by different amount of sample loaded on the column during the nLC-MS/MS analysis, we compared the total ion current (TIC) of all the inclusion-list-based acquisitions (Figure S6). This analysis revealed that similar intensity of the TICs was measured across the different digests, demonstrating that equal amounts were loaded of the phosphopeptides. We conclude thus that the biases revealed by our spectral-counting-based methods are also detectable and present when adopting a targeted proteomics approach.

Conclusions

Although high-throughput phosphoproteomics approaches can nowadays identify several thousands of unique phosphosites, it has become apparent that a large portion of regulatory important phosphorylation events remain elusive. These issues are partly due to the fact that many sites are simply inaccessible or very hard to detect following digestion with trypsin, hampering MS-based studies. In this work, we demonstrated that the use of multiple proteases is beneficial for large-scale phosphoproteomics analysis, exposing many regulatory relevant phosphosites on key signaling proteins, which would be occluded when only using trypsin. Notably, we also substantially increased the phosphoproteome coverage, ultimately showing that this workflow enables reproducible, quantitative, and complementary phosphoproteomic analysis. As a result, we make publicly available a human phosphopeptide atlas of more than 37,771 unique phosphopeptides, correlating to over 18,000 unique phosphosites. Researchers interested in particular phosphorylation events on targeted proteins may use the here-described web tool to pick the appropriate protease that will lead to successful detection of their sites of interest and extract the necessary parameters to construct targeted assays for phosphoproteomics studies. Our extensive resource also allows us to conclude several important technical issues, notably, (1) Ti^{4+}-IMAC phosphopeptide enrichment is equally efficient and selective independent of the protease used; (2) cumulative evidence from all protease data sets demonstrates unambiguously that protein phosphorylation reduces the cleavage efficiency near those sites for each protease used; (3) label-free quantitative phosphoproteomics is possible, however, only when comparing data generated by one protease; and (4) there is a clear bias toward trypsin in the currently available and widely used phosphopeptide and phosphosite databases. We expect our data will be a valuable resource for researcher in the signaling and proteomics field and may assist in encouraging people to perform, next to trypsin-based experiments, additional experiments with one or more complementary proteases.

Experimental Procedures

Sample Preparation and MS Analysis

Jurkat T lymphoma cells were resuspended at a final concentration of 1 or 2 3 106 cells/ml with 10 mM PGE2 in RPMI and incubated for 10 min. After treatment, Jurkat cells were lysed in 50 mM ammonium bicarbonate (pH 8.0), 8 M urea, 1 mM sodium orthovanadate, complete EDTA-free protease inhibitor mixture, and phosSTOP phosphatase inhibitor mixture (both Roche) lysis buffer. Digested proteins were subjected to phosphopeptide enrichment using Ti^{4+}-IMAC beads (Zhou et al., 2013), and enriched phosphopeptides were identified on a LTQ-Orbitrap Elite (Thermo Scientific) or Orbitrap Fusion (Thermo Scientific) using a decision-tree-based ion trap CID or ETD fragmentation. Further details are given in the Supplemental Experimental Procedures.

Data Analysis

The MS data were processed using Proteome Discoverer 1.4 (Thermo Scientific) and searched with the MS-GF+ search tool (Kim and Pevzner, 2014; Kim et al., 2010) against a Swissprot Homo sapiens database. The phosphorylation site localization of the identified phosphopeptides was performed using the phosphoRS algorithm 3.1 (Taus et al., 2011). For label-free analysis, raw data were processed with MaxQuant version 1.3.0.5 (Cox and Mann, 2008). For inclusion-list-based experiments, the data analysis was done manually. Extracted ion chromatograms (XICs) for a given parent mass were extracted with Xcalibur 3.0.63 (Thermo Scientific) with a mass tolerance of ±20 ppm. To evaluate the detectability of a given phosphopeptide, we made use of a spectral counting score (SCS) (Old et al., 2005; Zhang et al., 2006). The SCS for a given phosphorylation site was calculated as follows: (1) the amount of PSMs of that phosphosite in a given protease data set was divided by the total spectral counts obtained by that protease and then (2) the obtained value was normalized to 100% on the sum of the five values obtained from each protease. Further details are given in the Supplemental Experimental Procedures.

Accession Numbers

The MS proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository (Vizcaíno et al., 2013) with the data set identifier PXD001428.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, six figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.05.029.

Author contributions

P.G. and A.J.R.H. conceived the idea for this study. P.G. and T.T.A. performed the phosphoproteomics experiments. P.G analyzed and interpreted the data supported by T.T.A., M.P., B.v.B., and A.J.R.H. H.v.d.T. and B.v.B. developed the tools for phosphosite localization by the phosphoRS algorithm and built the web-based phosphopeptide atlas depository (http://phosphodb.hecklab.com/). P.G. and A.J.R.H. wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

We thank Simone Lemeer and Maarten Altelaar for critical evaluation of the manuscript. Part of this research was performed within the framework of the PRIME-XS project, grant number 262067, funded by the European Union 7th Framework Program, and the Netherlands Organization for Scientific Research (NWO) supported large-scale proteomics facility Proteins@Work (project 184.032.201) embedded in the Netherlands Proteomics Centre. Received: December 1, 2014 Revised: March 27, 2015 Accepted: May 17, 2015 Published: June 11, 2015

References

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S.G., and Pandey, A. (2007). A curated compendium of phosphorylation motifs. Nat. Biotechnol. 25, 285–286.

Bian, Y., Ye, M., Song, C., Cheng, K., Wang, C., Wei, X., Zhu, J., Chen, R., Wang, F., and Zou, H. (2012). Improve the coverage for the analysis of phosphoproteome of HeLa cells by a tandem digestion approach. J. Proteome Res. 11, 2828–2837.

Biringer, R.G., Amato, H., Harrington, M.G., Fonteh, A.N., Riggins, J.N., and Hu¨hmer, A.F. (2006). Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS. Brief. Funct. Genomics Proteomics 5, 144–153. Chong, P.K., Lee, H., Kong, J.W., Loh, M.C., Wong, C.H., and Lim, Y.P. (2008). Phosphoproteomics, oncogenic signaling and cancer research. Proteomics 8,4370–4382.

Courcelles, M., Frémin, C., Voisin, L., Lemieux, S., Meloche, S., and Thibault,P. (2013). Phosphoproteome dynamics reveal novel ERK1/2 MAP kinase substrates with broad spectrum of functions. Mol. Syst. Biol. 9, 669.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

de Graaf, E.L., Giansanti, P., Altelaar, A.F.M., and Heck, A.J.R. (2014). Singlestep enrichment by Ti^{4+}-IMAC and label-free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. Mol. Cell. Proteomics 13, 2426–2434.

Di Palma, S., Hennrich, M.L., Heck, A.J., and Mohammed, S. (2012). Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. J. Proteomics 75, 3791–3813.

Dickhut, C., Feldmann, I., Lambert, J., and Zahedi, R.P. (2014). Impact of digestion conditions on phosphoproteomics. J. Proteome Res. 13, 2761–2770.

Falini, B., Nicoletti, I., Bolli, N., Martelli, M.P., Liso, A., Gorello, P., Mandelli, F., Mecucci, C., and Martelli, M.F. (2007). Translocations and mutations involving the nucleophosmin (NPM1) gene in lymphomas and leukemias. Haematologica 92, 519–532.

Frese, C.K., Altelaar, A.F.M., Hennrich, M.L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A.J.R., and Mohammed, S. (2011). Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. J. Proteome Res. 10, 2377–2388.

Gauci, S., Helbig, A.O., Slijper, M., Krijgsveld, J., Heck, A.J.R., and Mohammed, S. (2009). Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. Anal. Chem. 81, 4493–4501.

Gershon, P.D. (2014). Cleaved and missed sites for trypsin, lys-C, and lys-N can be predicted with high confidence on the basis of sequence context. J. Proteome Res. 13, 702–709.

Granholm, V., Kim, S., Navarro, J.C.F., Sjo¨ lund, E., Smith, R.D., and Ka¨ ll, L. (2014). Fast and accurate database searches with MS-GF+Percolator. J. Proteome Res. 13, 890–897.

Guo, X., Trudgian, D.C., Lemoff, A., Yadavalli, S., and Mirzaei, H. (2014). Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. Mol. Cell. Proteomics 13, 1573–1584.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2014). The one hour yeast proteome. Mol. Cell. Proteomics 13, 339–347.

Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. Nat. Methods 11, 603–604.

Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 40, D261–D270.

Huang, P.H., and White, F.M. (2008). Phosphoproteomics: unraveling the signaling web. Mol. Cell 31, 777–781. Jaffe, J.D., Keshishian, H., Chang, B., Addona, T.A., Gillette, M.A., and Carr, S.A. (2008). Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification. Mol. Cell. Proteomics 7, 1952–1962.

Kalli, A., and Håkansson, K. (2010). Electron capture dissociation of highly charged proteolytic peptides from Lys N, Lys C and Glu C digestion. Mol. Biosyst. 6, 1668–1681.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277.

Kim, S., Mischerikow, N., Bandeira, N., Navarro, J.D., Wich, L., Mohammed, S., Heck, A.J., and Pevzner, P.A. (2010). The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Mol. Cell. Proteomics 9, 2840–2852.

Knudsen, E.S., and Wang, J.Y. (1997). Dual mechanisms for the inhibition of E2F binding to RB by cyclin-dependent kinase-mediated RB phosphorylation. Mol. Cell. Biol. 17, 5771–5783.

Low, T.Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hu¨ bner, N., van Breukelen, B., Mohammed, S., et al. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 5, 1469–1478.

Lu, K.P., Liou, Y.C., and Zhou, X.Z. (2002). Pinning down proline-directed phosphorylation signaling. Trends Cell Biol. 12, 164–172.

Macek, B., Mann, M., and Olsen, J.V. (2009). Global and site-specific quantitative phosphoproteomics: principles and applications. Annu. Rev. Pharmacol. Toxicol. 49, 199–221.

Mallick, P., Schirle, M., Chen, S.S., Flory, M.R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., et al. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. Nat. Biotechnol. 25, 125–131.

Meyer, J.G., Kim, S., Maltby, D.A., Ghassemian, M., Bandeira, N., and Komives, E.A. (2014). Expanding proteome coverage with orthogonal-specificity a-lytic proteases. Mol. Cell. Proteomics 13, 823–835.

Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011). Mass spectrometrybased proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol. Cell. Proteomics 10, M111.011015.

Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., et al. (2012). 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, O111.013698.

Molina, H., Horn, D.M., Tang, N., Mathivanan, S., and Pandey, A. (2007). Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. Proc. Natl. Acad. Sci. USA 104, 2199–2204.

Montoya, A., Beltran, L., Casado, P., Rodrı´guez-Prados, J.C., and Cutillas, P.R. (2011). Characterization of a TiO2 enrichment method for label-free quantitative phosphoproteomics. Methods 54, 370–378.

Murphree, A.L., and Benedict, W.F. (1984). Retinoblastoma: clues to human oncogenesis. Science 223, 1028–1033.

Old, W.M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K.G., Mendoza, A., Sevinsky, J.R., Resing, K.A., and Ahn, N.G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol. Cell. Proteomics 4, 1487–1502.

Pearce, L.R., Komander, D., and Alessi, D.R. (2010). The nuts and bolts of AGC protein kinases. Nat. Rev. Mol. Cell Biol. 11, 9–22.

Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A.J. (2012). Protease bias in absolute protein quantitation. Nat. Methods 9, 524–525.

Pitcher, J.A., Freedman, N.J., and Lefkowitz, R.J. (1998). G protein-coupled receptor kinases. Annu. Rev. Biochem. 67, 653–692.

Rigbolt, K.T., and Blagoev, B. (2012). Quantitative phosphoproteomics to characterize signaling networks. Semin. Cell Dev. Biol. 23, 863–871.

Ruprecht, B., and Lemeer, S. (2014). Proteomic analysis of phosphorylation in cancer. Expert Rev. Proteomics 11, 259–267.

Santamarı´a, E., Mora, M.I., Mun˜ oz, J., Sa´ nchez-Quiles, V., Ferna´ ndez-Irigoyen, J., Prieto, J., and Corrales, F.J. (2009). Regulation of stathmin phosphorylation in mouse liver progenitor-29 cells during proteasome inhibition. Proteomics 9, 4495–4506.

Schmidt, A., Gehlenborg, N., Bodenmiller, B., Mueller, L.N., Campbell, D., Mueller, M., Aebersold, R., and Domon, B. (2008). An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. Mol. Cell. Proteomics 7, 2138–2150.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. 23, 1391–1398.

Sharma, K., D'Souza, R.C., Tyanova, S., Schaab, C., Wisniewski, J.R., Cox, J., and Mann, M. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Rep. 8, 1583–1594.

Soderblom, E.J., Philipp, M., Thompson, J.W., Caron, M.G., and Moseley, M.A. (2011). Quantitative label-free phosphoproteomics strategy for multifaceted experimental designs. Anal. Chem. 83, 3758–3764.

Swaney, D.L., McAlister, G.C., and Coon, J.J. (2008). Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat. Methods 5, 959–964.

Swaney, D.L., Wenger, C.D., and Coon, J.J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J. Proteome Res. 9, 1323–1329.

Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., and Mechtler, K. (2011). Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. 10, 5354–5362.

Tsiatsiani, L., and Heck, A.J.R. (2015). Proteomics beyond trypsin. FEBS J., Published online March 30, 2015. http://dx.doi.org/10.1111/febs.13287.

, J.A., Côté , R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J., et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 41, D1063–D1069.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz,M., Savitski, M.M., Ziegler, E., Butzmann, L.,Gessulat, S.,Marx, H., et al. (2014).Mass spectrometry-based draft of the human proteome. Nature 509, 582–587.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods 6, 359–362.

Yates, J.R., 3rd, Mohammed, S., and Heck, A.J. (2014). Phosphoproteomics. Anal. Chem. 86, 1313.

Zhang, B., VerBerkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L., and Samatova, N.F. (2006). Detecting differential and correlated protein expression in label-free shotgun proteomics. J. Proteome Res. 5, 2909–2918.

Zhou, H., Ye, M., Dong, J., Corradini, E., Cristobal, A., Heck, A.J.R., Zou, H., and Mohammed, S. (2013). Robust phosphoproteome enrichment using monodisperse microsphere-based immobilized titanium (IV) ion affinity chromatography. Nat. Protoc. 8, 461–480.

Supplemental information

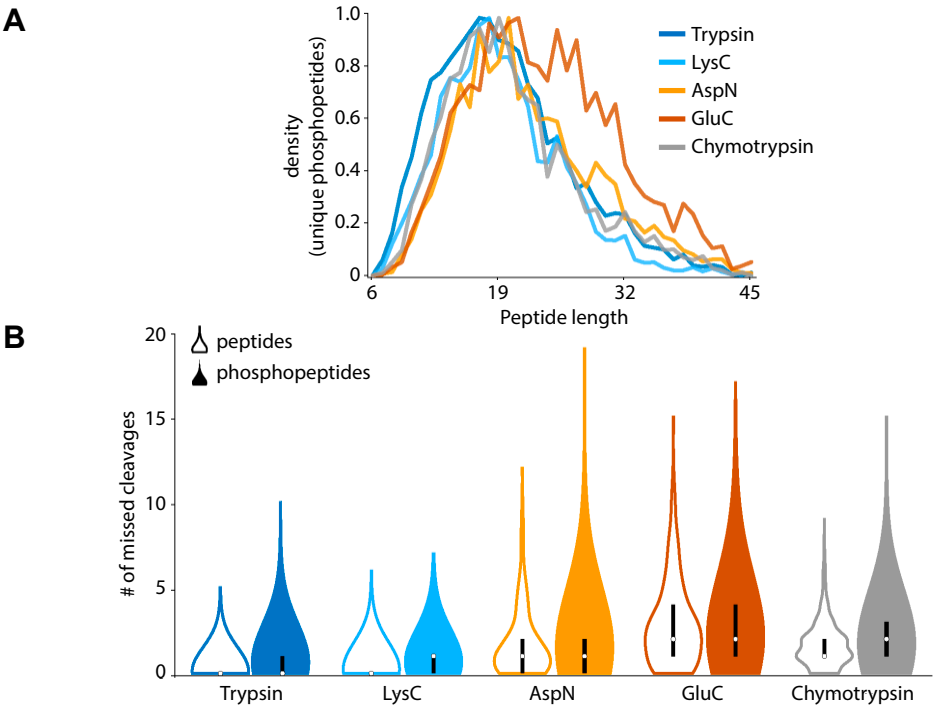
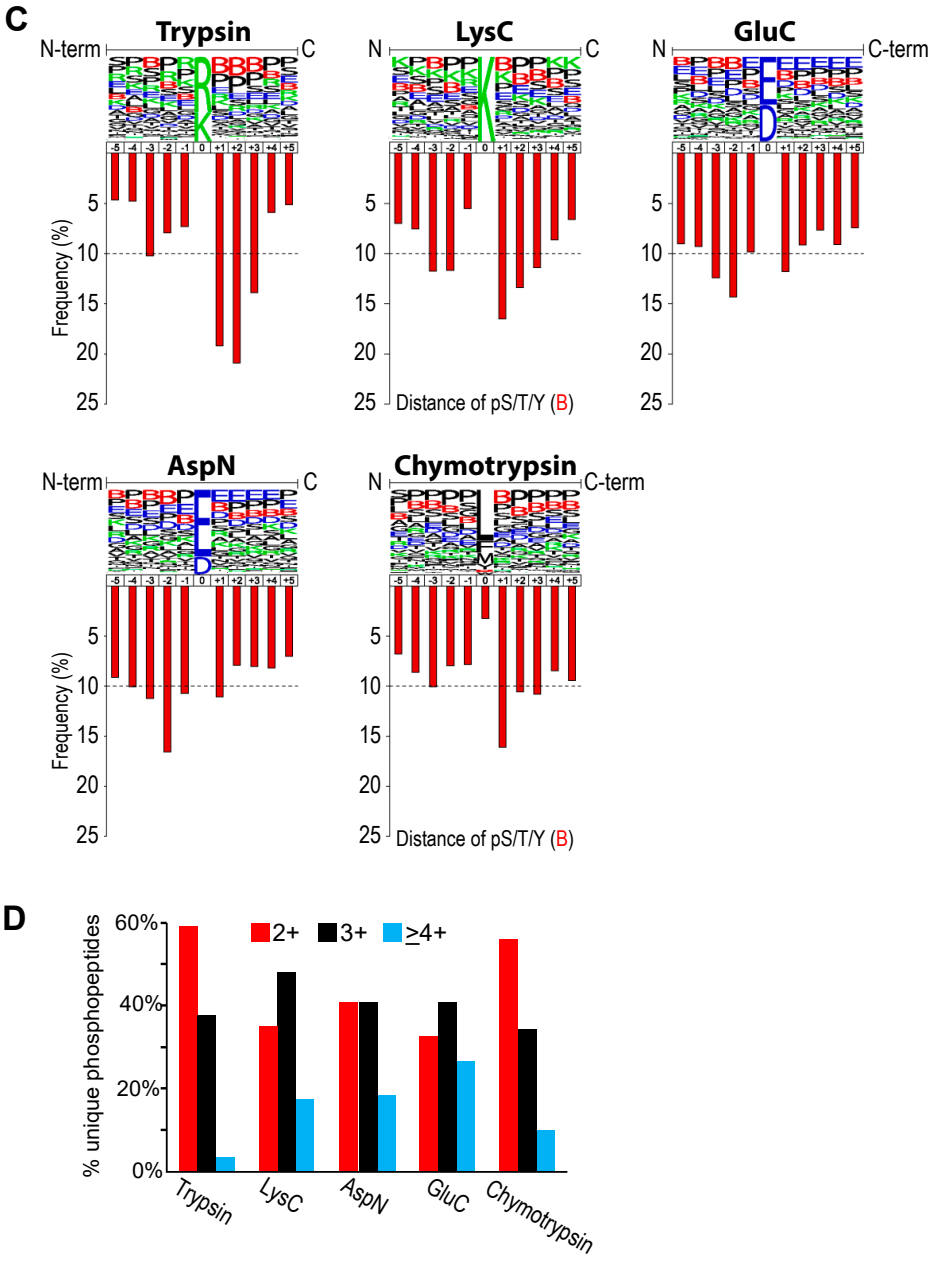


Figure S1, Related to Figure 1. Global physico-chemical properties of the phosphopeptides. (A) Length distribution of the phosphopeptides detected in the digests of the five different proteases. For 4 out of 5 proteases most of the identified phosphopeptides are about 14–25 amino acids long, whereas GluC tends to generate somewhat longer phosphopeptides. (B) Violin plot illustrating a different distribution in the number of missed proteolytic events between peptides and phosphopeptides as detected in the digests of the 5 different proteases, revealing that phosphorylation leads to an enhanced number of miss-cleavages for all proteases. (C) Missed cleavage analysis in the context of the close-proximity of a phosphorylated site in the five different data sets. For each protease, the composition of the sequence surrounding the missed cleavage site was evaluated and displayed as a residue frequency logo over the distance range -5/+5 residues. The phosphorylated residue is annotated as B (pS, pT, or pY). A specific prevalence of the site of B is clear in each dataset, with for trypsin, LysC and chymotrypsin this being +1 and +2, adjacent to the missed cleavage, and for GluC and AspN a phosphorylation at the -2 site mostly hampers cleavage. (D) Distribution of the doubly (red), triply (black), quadruply and higher charge state (light blue) phosphopeptide ions identified in the data sets of each of the 5 digests. Data bars are normalized to the total numbers of phosphopeptides within each digest.


C
Trypsin
N-term
C
LysC
N
C
GluC
N
C-term
AspN
N-term
C
Chymotrypsin
N
C-term
Frequency (%)
5
10
15
20
25
Distance of pS/T/Y (B)
D
% unique phosphopeptides
60%
40%
20%
0%
2+
3+
≥4+
Trypsin
LysC
AspN
GluC
Chymotrypsin

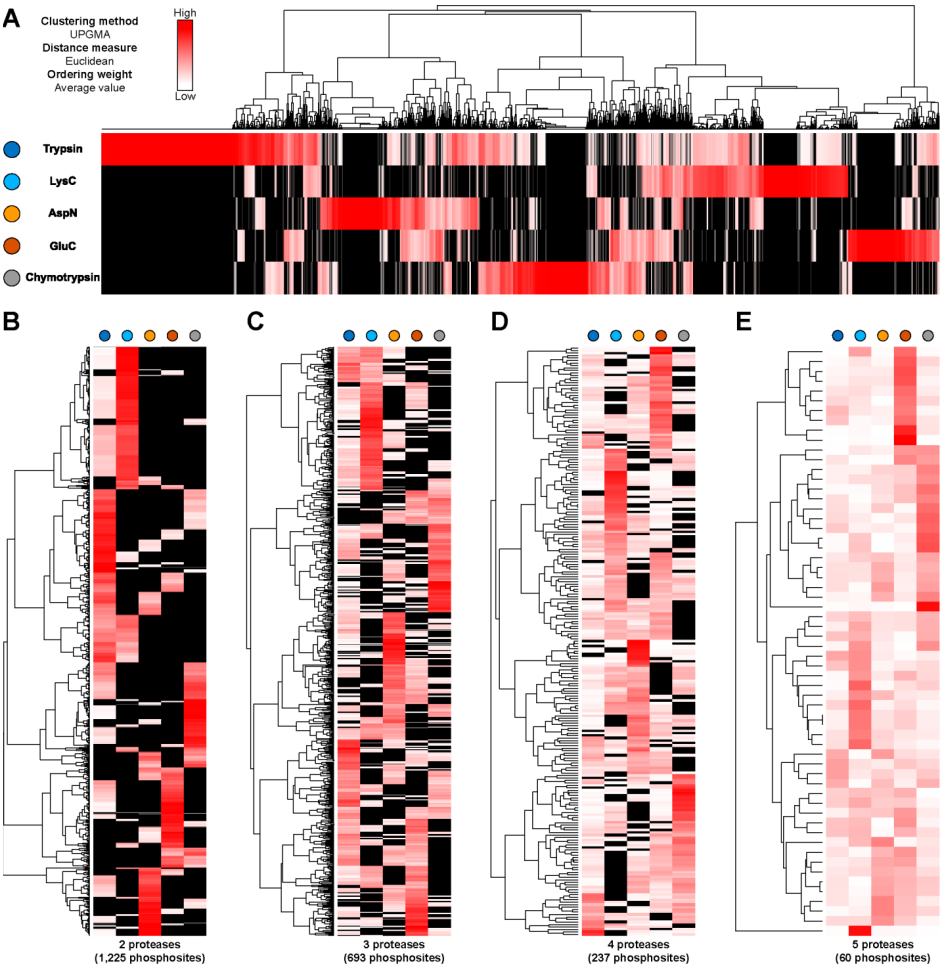


Figure S2, Related to Figure 2. Spectral count score heat map illustrating the contribution of each protease in the detection of a particular phosphosite. **(A)** Heat maps of all the phosphorylation sites identified with a least 10 PSMs by at least one protease, **(B)** two proteases, **(C)** three proteases, **(D)** four proteases, and **(E)** five proteases. The numbers of sites in each heat map is reported in brackets.

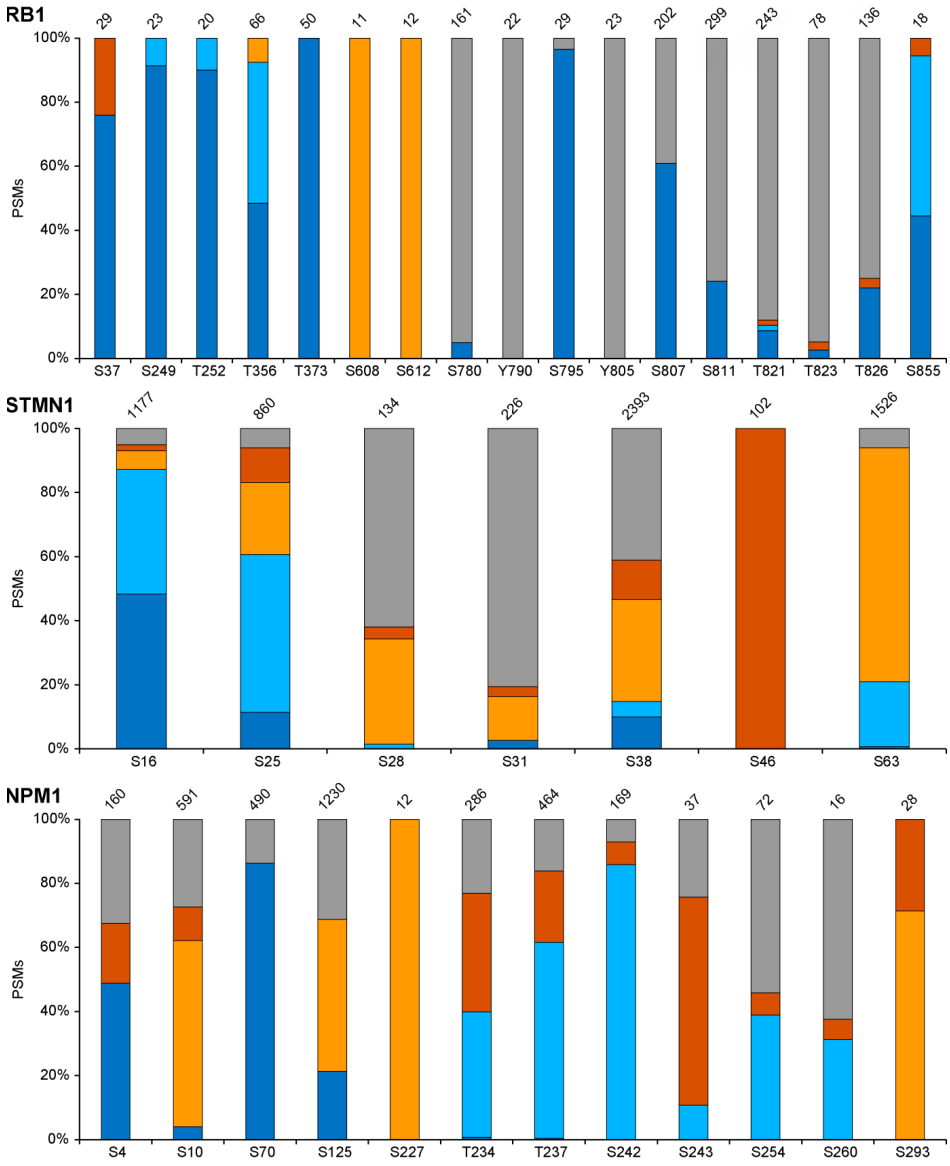


Figure S3, Related to Figure 2. Protease dependent detectability of various sites in phosphoproteins. The x-axis indicates the detected sites, at the top the total number of PSMs is indicated. The stack bars indicate how these PSMs are distributed over all protease data sets.

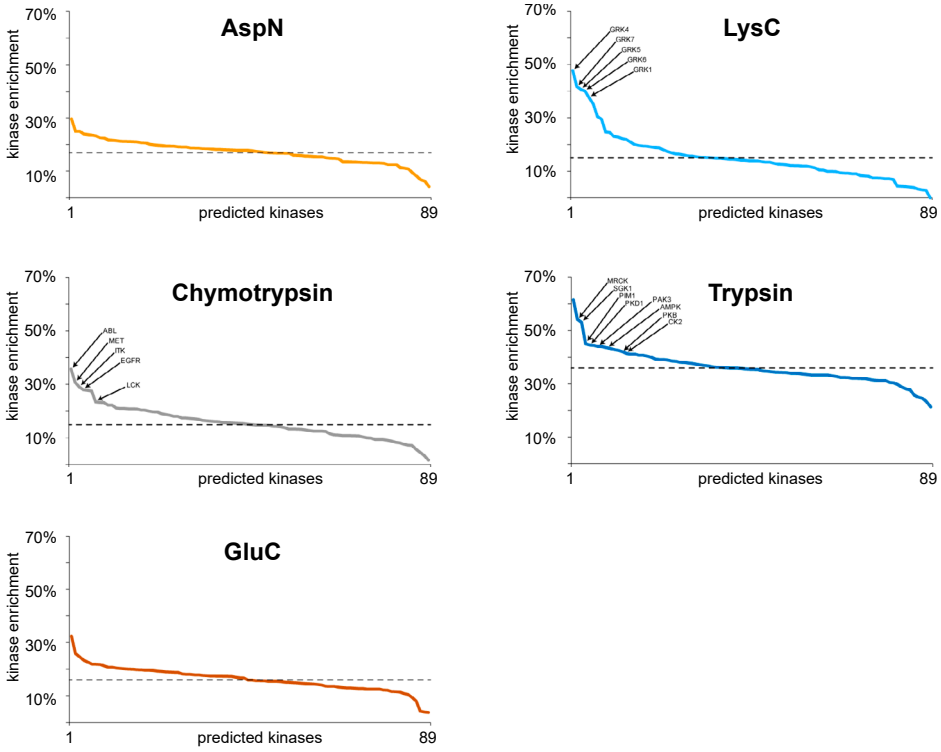


Figure S4, Related to Figure 3. Upstream kinases predictions by NetworKIN. NetworKIN predicted in total 89 kinases to be responsible for the phosphorylation of the here identified phosphosites. The plots show how these predictions are distributed over all the five protease-linked data sets. Dash lines represent the average distribution within each data set. Predictions belonging to different catalytic subunits of the same kinase are averaged. Annotated are kinases that lead to substantially more predicted sites in specific protease-linked datasets, with for example the family of GRK-kinases being specifically enriched in the LysC dataset. This NetworKIN analysis further confirms that there is a bias in the datasets.

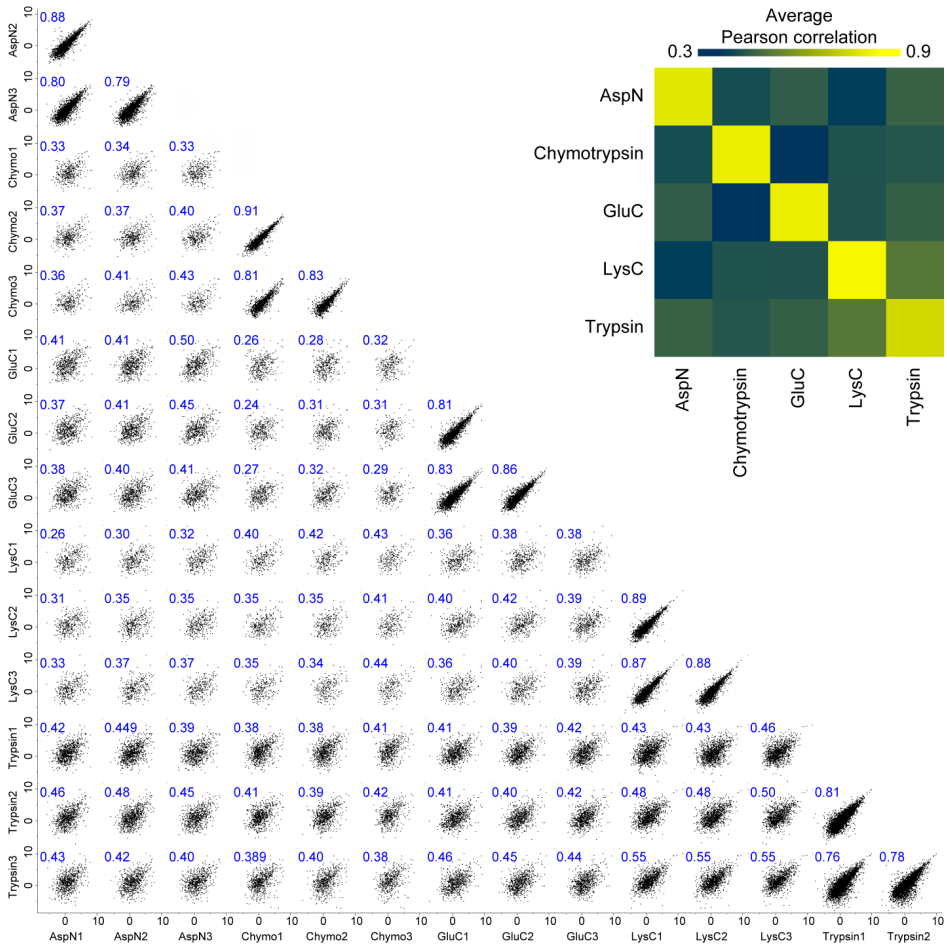


Figure S5, Related to Figure 4A and 4B. Intensity Pearson correlations for all the phosphosite quantifications in the five digests. The phosphosite intensities (log base 2) are plotted for all the enrichment replicas.

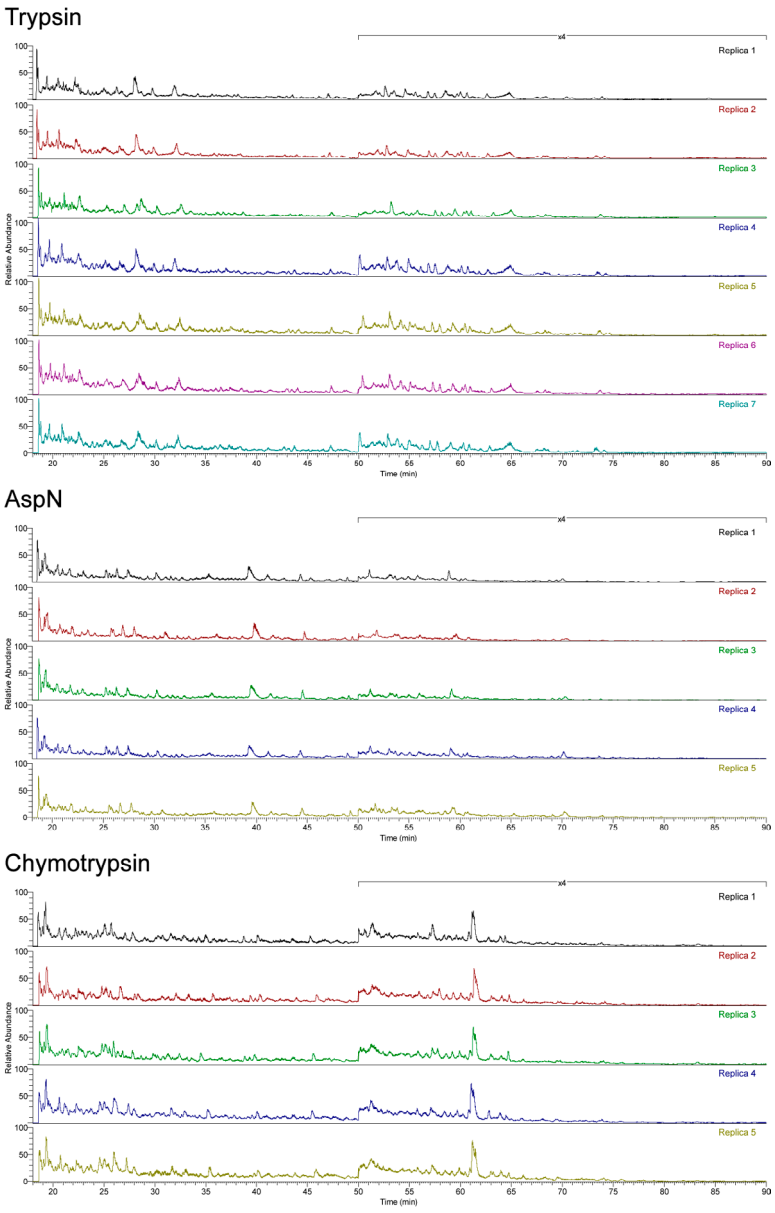


Figure S6, Related to Figure 4C. Total ion chromatograms (TIC) in the inclusion-list based experiments. The chromatograms (18-90 min) show a similar level of TIC, indicating that equal amounts of phosphopeptide samples had been loaded on the column. Intensity was normalized to the highest TIC, corresponding to 1e10. From 50 to 90 min the TIC intensities have been enhanced by a factor 4 for a better visualization.

Extended experimental procedures

Cell Culture and Digest Preparation

Jurkat T lymphoma cells were grown in RPMI 1640 medium supplemented with 10% fetal bovine serum and penicillin/streptomycin (Lonza). Before PGE2 stimulation, cells were centrifuged for 1 min at 1500 · g, growth medium was removed and the cells were resuspended at a final concentration of 1-2 · 10^6 cells/mL with 10 µM PGE2 in RMPI and incubated for 10 min. After treatment Jurkat cells were washed twice with PBS and harvested. Cell lysis was performed on ice by sonication in buffer containing 50 mM ammonium bicarbonate (pH 8.0), 8 M urea, 1 mM sodium orthovanadate, complete EDTA-free protease inhibitor mixture and phosSTOP phosphatase inhibitor mixture (both Roche). Cell debris were then removed by centrifugation at 20,000 · g for 15 min at 4 °C. The total protein concentration was measured using a Bradford Assay (BioRad) and then split into 1 mg aliquots for enzymatic digestion.
Proteins were reduced with DTT at a final concentration of 4 mM at 56°C for 25 min; subsequently samples were alkylated with iodoacetamide at a final concentration of 8 mM at RT for 30 min in the dark. For proteolytic digestion, the urea concentration was diluted to 1 M before addition of proteases (AspN, chymotrypsin, GluC, trypsin (Promega) and LysC (Wako)). An enzyme:substrate ratio of 1:100 was used. The digestion was quenched by acidification to 5% of formic acid (FA).The digests were desalted using Sep-Pak C18 cartridges (Waters), dried in vacuo and stored at -80 °C for further use.

Phosphopeptides enrichment by Ti^{4+}-IMAC

Phosphopeptides enrichment was performed essentially as previously described (Zhou et al., 2013). Briefly, the Ti4+-IMAC beads (500 µg of beads/200 µL pipet tip) were loaded onto GELoader tips (Eppendorf) using a C8 plug and in parallel spin tip enrichment was used. The Ti4+-IMAC columns were conditioned using 50 µL of loading buffer consisting of 6% trifluoroacetic acid (TFA) in 80% acetonitrile (ACN) and centrifugation at 200 · g for 10 min. The protein digests were dissolved in the loading

buffer and split in aliquots corresponding to 200 µg of cell lysates. The aliquots were transferred to the spin tips and centrifuged at 100 · g for 30 min. The columns were sequentially washed with 50 µL of washing buffer 1 (50% ACN, 0.5% TFA containing 200 mM NaCl) followed by additional washing with 50 µL of 0.1% TFA in 50% ACN, each centrifuged at 170 · g for 15 min. The bound peptides were eluted into a new tube (already containing 35 µL of 10% formic acid) with 20 µL of 10% ammonia by centrifugation at 100 · g for 20 min. A final elution was performed with 5 µL of 2% FA in 80% ACN at 100 · g for 10 min. The collected eluate was further acidified by adding 3 µL of 100% FA prior to nLC-MS analysis.

Reverse phase chromatography and mass spectrometry

Peptides were subjected to reversed phase nLC-MS/MS analysis using a Proxeon EASY-nLC 1000 (Thermo Scientific) with an analytical column heater (40°C) and a LTQ-Orbitrap Elite (Thermo Scientific). Peptides were first trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) at a maximum pressure of 800 bar with 100% solvent A (0.1% FA in water) before being separated on the analytical column (Agilent Poroshell 120 EC-C18, 2.7 µm, 40 cm x 50 µm). Peptides were chromatographically separated by a 150 min gradient from 7% to 30% solvent B (0.1% FA in ACN) at a flow rate of 100 nL/min. The total measurement time for each sample was 180 min. The eluent was sprayed via a distal coated fused silica emitter (360 µm o.d., 20 µm i.d., 10 µm tip i.d.; constructed in-house) butt-connected to the analytical column. The electrospray voltage was set to 1.7 kV. The mass spectrometer was operated in a data-dependent mode to automatically switch between MS and MS/MS. Briefly, survey full-scan MS spectra were acquired in the Orbitrap analyzer, scanning from m/z 350 to m/z 1500 at a resolution of 60,000 at m/z 400 using an AGC setting of 1e6 ions. Charge state screening was enabled and precursors with either unknown or 1+ charge states were excluded. After the survey scan the 20 most intense precursors were selected for subsequent decision tree-based ion trap CID or ETD fragmentation (Frese et al., 2011; Swaney et al., 2008).
The normalized collision energy for CID was set to 35% and supplemental activation for ETD and dynamic exclusion were enabled (exclusion size list 500, exclusion duration 60 s).
Samples for the targeted experiments were subjected to reversed phase

nLC-MS/MS analysis using Agilent 1290 Infinity System (Agilent Technologies) and an Orbitrap Fusion (Thermo Scientific) as described previously (Cristobal et al., 2012). Peptides were first trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) at 5 µl/min with 100% solvent A (0.1% FA in water) before being separated on the analytical column (Agilent Poroshell 120 EC-C18, 2.7 µm, 50 cm x 75 µm). Peptides were chromatographically separated by a 100 min gradient from 13% to 44% solvent B (0.1% FA in 80% ACN) at a flow rate of 350 nL/min. The total measurement time for each sample was 120 min. The electrospray voltage was set to 2.0 kV. The mass spectrometer was operated in a data-dependent mode to automatically switch between MS and MS/MS as described above.
After the survey scan, only precursors from the inclusion list were targeted for MS/MS spectrum acquisition over the course of the experiment via a decision tree-based ion trap CID or ETD fragmentation. Inclusion lists were generated by querying in our phosphoDB database the m/z and charge state of the precursor ion of all the phosphopeptides carrying the phosphorylation site. Phosphopeptides ASTRPPTLpSPIPHIPRSPY (chymotrypsin) and TNILQYASTRPPTLpSPIPHIPR (trypsin) for S780 on RB and, DKFFTRGQPVLpTPP (AspN) and GQPVLpTPPDQLVIANIDQSDFEGFSYVNPQFVHPILQSAV (trypsin) for T638 on PRKCA were selected. Charge state screening was enabled for selection of precursors, and the m/z tolerance around targeted precursors was ±20 ppm. The normalized collision energy for CID was set to 35% and supplemental activation for ETD and dynamic exclusion were enabled (exclusion duration 60 s).

Data analysis

Raw data were converted from their native raw file format to the mgf or mzML file format using Proteome Discoverer version 1.4 (Thermo Scientific). Subsequently the data was searched against a Swissprot Homo sapiens database version 2012_09 (40,992 sequences) and, separately, against the corresponding reversed decoy database using the MS-GF+ search tool, version 9881 (Kim and Pevzner, 2014; Kim et al., 2010). The database search was performed with the following parameters: mass tolerance of ±20 ppm for precursor masses and appropriate settings for activation technique and fragmentation spectrum mass accuracy. The enzymatic parameters were set

to allow fully enzymatic termini for each peptide, for the respective enzyme. Cysteine carbamidomethylation was used as a fixed modification and methionine oxidation,
protein N-terminal acetylation and serine, threonine and tyrosine phosphorylation were set as variable modifications. The false discovery rate was set to 1% at the PSMs level. The minimum and maximum peptide lengths allowed were 6 and 45 amino acids, respectively. The phosphorylation site localization of the identified phosphopeptides was performed using the phosphoRS algorithm 3.1 (Taus et al., 2011). A site localization probability (pRS) of at least 0.75 was used as threshold for the phosphoresidue localization. Finally, the mzIdentML output files were converted to plain text files by the mzidLibrary tool 1.6 (Ghali et al., 2013).
For inclusion list-based experiments the data analysis was done manually. Extracted ion chromatograms (XICs) for a given parent mass were extracted with Xcalibur 3.0.63 (Thermo Scientific) with a mass tolerance of ±20 ppm. Gaussian smoothing was applied with a numbers of points set to 5, and Genesis was use as peak detection algorithm. Confirmation of the presence of a phosphopeptide was based on the successful identification by database search. When no phosphopeptides were identified, the peak with highest area belonging to the extracted mass was selected for comparison. To evaluate the detectability of a given phosphopeptide we made use of a spectral counting score (SCS) (Old et al., 2005; Zhang et al., 2006). The SCS for a given phosphorylation site was calculated as follows: (1) the amount of PSMs of that phosphosite in a given protease data set was divided by the total spectral counts obtained by that protease, then (2) the obtained value was normalized to 100% on the sum of the five values obtained from each protease. To identify significant phosphorylation motifs in each of the five data sets, the high confident localized phosphorylation sites were further analyzed using motif-x (Schwartz and Gygi, 2005). As parameters we used a minimum occurrence of 20, a significance threshold of 1e-6 taking the entire human proteome (IPI Human Proteome) as background reference set. Upstream kinases responsible for the observed phosphorylation sites were also evaluated by using the NetworKIN 3.0 kinase prediction algorithm (Horn et al., 2014). A score threshold of 2 was applied, and only the top 10 predictions were used. Kinases with less than 10 predictions were discharged.

Label-free quantification

For label-free analysis, raw data were processed with MaxQuant version 1.3.0.5 (Cox and Mann, 2008). The database search was performed in Andromeda search engine with the following parameters: an initial mass tolerance of ±20 ppm and a final mass tolerance of ±6 ppm for precursor masses, ±0.6 Da for CID and ETD ion trap fragment ions, allowing two missed cleavages for trypsin, three for LysC and chymotrypsin, five for GluC and AspN. Cysteine carbamidomethylation was used as a fixed modification and methionine oxidation, protein N-terminal acetylation and serine, threonine and tyrosine phosphorylation as variable modifications. The false discovery rate was set to 1% for peptides, proteins and phosphosites, the minimum peptide length allowed was six amino acids and a minimum Andromeda peptide score of 60 was required. The match between runs feature was enabled. A site localization probability of at least 0.75 and a score difference of at least 5 were used as threshold for the localization of phosphorylated residues.

Supplemental references

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

Cristobal, A., Hennrich, M.L., Giansanti, P., Goerdayal, S.S., Heck, A.J.R., and Mohammed, S. (2012). In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. Analyst 137, 3541–3548.

Frese, C.K., Altelaar, A.F.M., Hennrich, M.L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A.J.R., and Mohammed, S. (2011). Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. J. Proteome Res. 10, 2377–2388.

Ghali, F., Krishna, R., Lukasse, P., Martinez-Bartolome, S., Reisinger, F., Hermjakob, H., Vizcaino, J.A., and Jones, A.R. (2013). Tools (viewer, library and validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. Mol. Cell. Proteomics 12, 3026–3035.

Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. Nat. Methods 11, 603–604.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277.

Kim, S., Mischerikow, N., Bandeira, N., Navarro, J.D., Wich, L., Mohammed, S., Heck, A.J., and Pevzner, P.A. (2010). The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Mol. Cell. Proteomics 9, 2840–2852.

Old, W.M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K.G., Mendoza, A., Sevinsky, J.R., Resing, K.A., and Ahn, N.G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol. Cell. Proteomics 4, 1487–1502.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. 23, 1391–1398.

Swaney, D.L., McAlister, G.C., and Coon, J.J. (2008). Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat. Methods 5, 959–964.

Taus, T., Kocher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., and Mechtler, K. (2011). Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. 10, 5354–5362.

Zhang, B., VerBerkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L., and Samatova, N.F. (2006). Detecting differential and correlated protein expression in labelfree shotgun proteomics. J. Proteome Res. 5, 2909–2918.

Zhou, H., Ye, M., Dong, J., Corradini, E., Cristobal, A., Heck, A.J.R., Zou, H., and Mohammed, S. (2013). Robust phosphoproteome enrichment using monodisperse microsphere-based immobilized titanium (IV) ion affinity chromatography. Nat. Protoc. 8, 461–480.

6. Deep proteome profiling of Trichoplax adhaerens reveals remarkable features at the origin of metazoan multicellularity

Ringrose, J.H.J., van den Toorn, H.W.P., Eitel, M., Post, H., Neerincx, P., Schierwater, B., Altelaar, A.F.M., and Heck, A.J.R. (2013). Deep proteome profiling of Trichoplax adhaerens reveals remarkable features at the origin of metazoan multicellularity. Nat. Commun. 4, 1408.

Genome sequencing of arguably the simplest known animal, *Trichoplax adhaerens****, uncovered a rich array of transcription factor and signalling pathway genes. While the existence of such genes allows speculation about the presence of complex regulatory events, it does not reveal the level of actual protein expression and functionalization through posttranslational modifications. Using high-resolution mass spectrometry, we semi-quantified 6,516 predicted proteins, revealing evidence of horizontal gene transfer and the presence at the protein level of nodes important in animal signalling pathways. Moreover, our data demonstrate a remarkably high activity of tyrosine phosphorylation, in line with the hypothesized burst of tyrosine regulated signalling at the instance of animal multicellularity. Combined, this*** *Trichoplax* ***proteomics dataset offers significant new insight into the mechanisms underlying the emergence of metazoan multicellularity and provides a resource for interested researchers.***

Introduction

Recently the genomes of several phyla located near the base of the metazoan phylogenetic tree have been sequenced, including Porifera, Ctenophora, Cnidaria and Placozoa (Srivastava et al., 2008). The placozoan genome is by far the smallest of these and has been regarded as the best living surrogate for the hypothetical Cnidaria-Bilateria ancestor genome or even metazoan genome in general (Schierwater and Kuhn, 1998; Schierwater et al., 2009). The placozoan Trichoplax adhaerens is morphologically the simplest of all animals, lacking a body axis, basal lamina and extracellular matrix (ECM), and containing only 5 somatic cell types (Guidi et al., 2011). It can be found in tropical and subtropical sea waters and appears as a flat disc of 2-3mm diameter consisting of two epithelial layers with a loose layer of fiber cells in between (Schierwater, 2005). Trichoplax reproduces in vitro by fission and budding, and although in vitro the egg stadium does not develop into an embryonic stage beyond 64-128 cells, there are clear indications for a bisexual reproduction cycle, which left its signature in the DNA (Eitel et al., 2011; Signorovitch et al., 2005). The Trichoplax genome contains 11,500 genes and interestingly include important genes characteristic of more complex bilaterian animals such as developmental

signalling pathways, neuroendocrine processes, and extracellular matrix proteins (Srivastava et al., 2008). The available genome, however, does not reveal which proteins are expressed, to what level and whether proteins are functionally regulated by posttranslational modifications (PTMs). Using high-resolution mass spectrometry based proteomics we monitor for the first time which Trichoplax genes are actually translated and expressed. Moreover, as the functionality of proteins is to a large extent determined by posttranslational modifications (PTMs), which can only be studied at the protein level, we look into more detail at some important PTMs such as phosphorylation and acetylation. In summary, here we show that studying the proteome of Trichoplax, one of the most ancient extant multicellular animals, may provide significant insight into the mechanisms underlying the emergence of metazoan multicellularity.

Results

We used 2,800 hand-picked animals of Trichoplax adhaerens, and combined two independent enzymatic digestions, using trypsin and Lys-N, with strong cation exchange (SCX) peptide fractionation, nano- reversed-phase liquid chromatography, and high resolution mass spectrometry (LC-MS) to confidently identify 6,516 proteins at a false discovery rate of less than 1% (Supplementary Table S1). This first extensive catalogue of proteins expressed by Trichoplax constitutes 57% of all predicted proteins, and reveals similar qualitative features as the published in-depth proteome of C. elegans (Supplementary Fig. S1). The high quality dataset allows determination of individual protein expression abundance as described previously (Aye et al., 2010), which covers over four orders of magnitude (Fig. 1). Conformation of the expression abundance was obtained via a bio-duplicate experiment, performed 6 months later using a new batch of hand-picked animals (supplementary Fig. S2). Detailed inspection of the expression abundance of Trichoplax proteins proves a real treasure trove, strengthening evolutionary insights through the addition of an extra layer of information including the occurrence of posttranslational events and confirmation of hypothesized gene expression. For instance, our data contain clear evidence for the abundant expression of "apicortin", a unique protein with a putative cytoskeletal role shared only

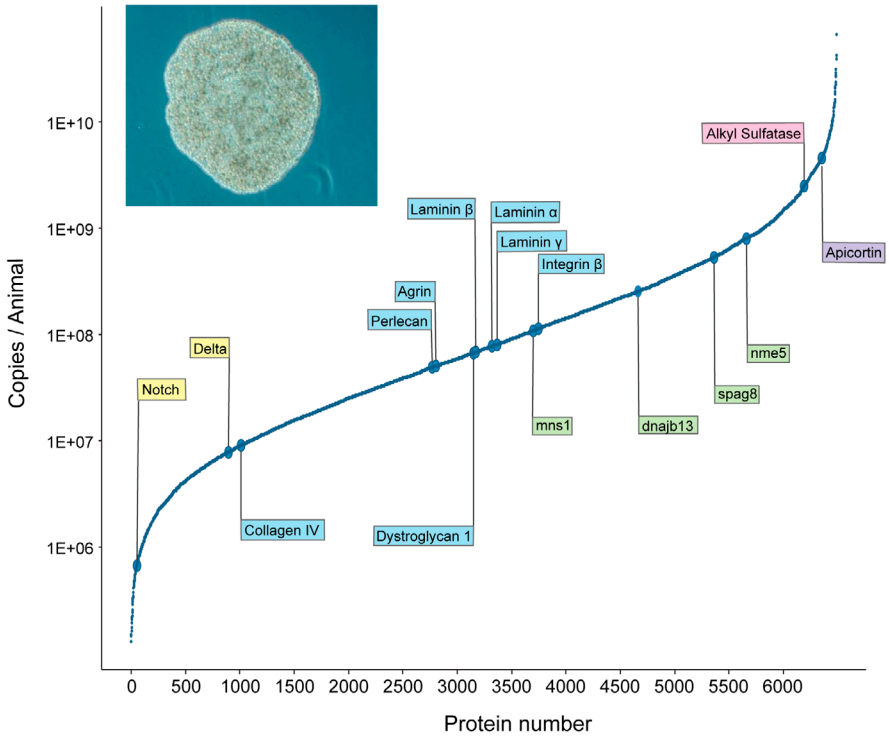


Figure 1 An in-depth quantitative view of the *Trichoplax* proteome. Protein copy numbers per animal of all identified proteins were estimated based on spectral counts as described in the methods section. Selected proteins are highlighted and colored by functional grouping as annotated in the KEGG (Kanehisa 2012) reference pathway maps: yellow for proteins belonging to the notch signaling pathway, blue for ECM and adhesion related proteins, green for male germ line markers, and pink and purple for alkyl sulfatase and apicortin. Inset photograph: *Trichoplax* animal in culture.

by apicomplexan parasites and Trichoplax (Orosz, 2009). And although there has been indirect evidence for a sexual lifecycle in Trichoplax, five conserved sperm markers characteristic for different stages in spermatogenesis have only recently been identified (Eitel et al., 2011), four of which we can now confirm as being expressed abundantly on the protein level. Our data also provide evidence for the expression of proteins annotated by the KEGG database as part of signalling pathways important for animal development and patterning, including Delta and Notch proteins from the Notch pathway and some downstream proteins from the Notch,

Wnt and TGF-β pathways (Fig. 1 and Supplementary Fig. S3). The placozoan genome encodes orthologs for many typical bilaterian ECM proteins. However, a peculiarity among the Metazoa is that an ECM of any kind, including a basal lamina, has eluded detection in adult Trichoplax, raising the possibility for expression of the ECM in other hitherto unknown developmental or life cycle stages. In contrast, the proteomics data confirm the presence of proteins involved in ECM and ECM receptor interactions including integrin-β, laminins, collagen IV, perlecan, agrin, and dystroglycan, although there is a possibility that these proteins have alternative functions and/ or organization. Recently, it was argued that a classical type cadherin in complex with two armadillo-type catenins is a key element in the origin of metazoan multicellularity (Hulpiau and van Roy, 2011). Here, we only found proof for expression of the flamingo type cadherin and not of the suggested classical cadherin. The expression of the two armadillo type catenins p120ctn and β-catenin could not be confirmed by our proteomics data.

Although not specifically targeted, in depth sequencing the Trichoplax proteome additionally allowed us to detect numerous protein post-translational modifications, including widespread N-acetylation, Lysine-acetylation and phosphorylation. To focus on the latter, we detected, using an FDR < 1%, 2177 unique phosphosites by SCX. This number is similar to what we expect if an identical amount of mammalian sample is analysed using similar SCX-based proteomic strategies(Zhai et al., 2008). At the same time our quantitative proteomics data allowed us to define a semi-quantitative kinome tree for Trichoplax, depicted in Fig. 2, in which detection and abundance in our proteome dataset are shown for each kinase present in the Trichoplax genome (Supplementary Table S2). Many Trichoplax kinases show high homology with human kinases (Supplementary Fig. S4 and S5), and abundant kinases in our dataset are homologues to, for instance, PKC, MAPK, CamK, AKT, CK2, and Src. Not surprisingly, motifs of abundant kinases such as CK2, MAPK and Src family kinases were also abundantly recognized in the Trichoplax phosphopeptide dataset. Strikingly, combining the data of three independent SCX phospho-datasets (Fig. 3 and Supplementary Table S3) serine accounted for 1432 (66%) of the phosphosites, threonine for 555 (25%), and tyrosine for 190 (9%). The latter number of ~9% (respectively, 8.9, 7.3 and 9.8 % in the three

independent experiments) is extremely high and consisted of 166 unique proteins for which 95 human orthologs could be found, almost all of which (95%) have been reported to be phosphorylated on tyrosine

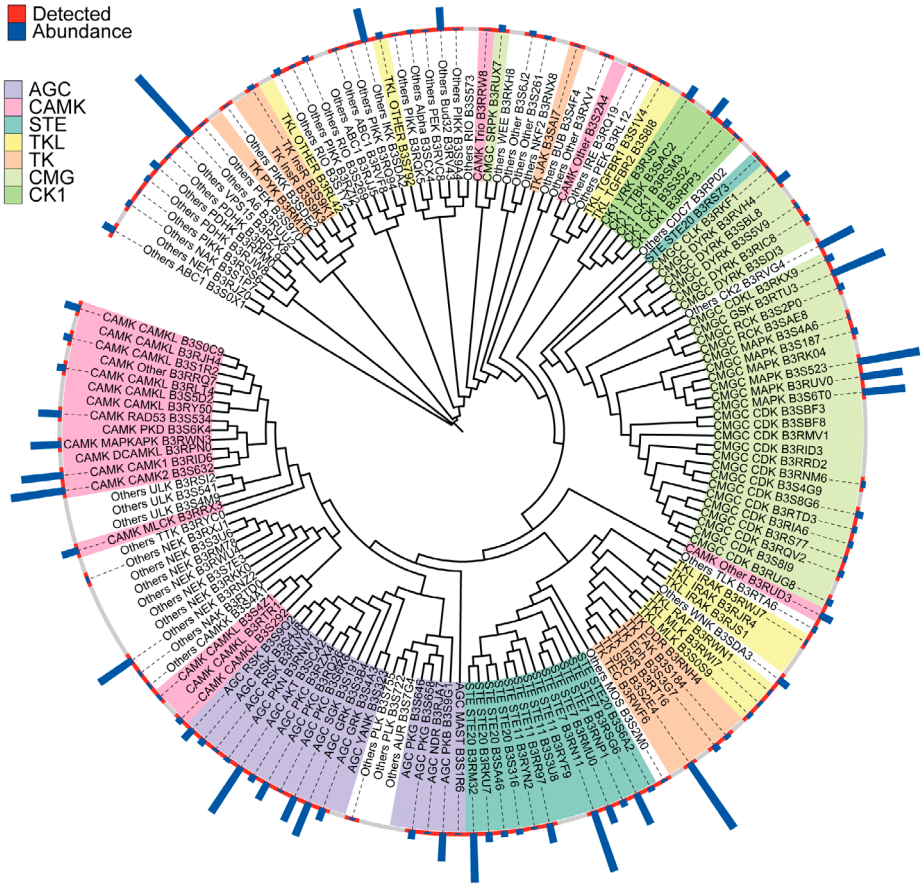


Figure 2 A quantitative view of the *Trichoplax* kinome. All identified kinases were ordered by phylogenetic distances of the kinase sequences. FASTA sequences of all proteins designated by the KEGG (Kanehisa 2012) database as protein kinase (http://www.genome.jp/kegg-bin/get_htext?tad01001) were aligned using ClustalX2.1 (Larkin et al., 2007) using default parameters for multiple alignment and bootstrapping. For visualization a phylogenetic tree was calculated with the neighbour-joining algorithm, exported and loaded into the Interactive Tree of Life tool (Letunic and Bork, 2007). Kinase families are coloured as defined by the KEGG (Kanehisa 2012) database. On the outer rim the colour red indicates that the kinase is detected, blue bars indicate the relative abundances in protein copy number per animal. Kinases are named according to their respective group, followed by family and finally their Uniprot accession number.

residues as well (Supplementary Table S4)(Hornbeck et al., 2011). Using similar experimental approaches, ~2% tyrosine phosphorylation is generally reported in dozens of phosphoproteomics studies on higher organisms ranging from C. elegans to humans (Huttlin et al., 2010; Zhai et al., 2008; Zhou et al., 2011; Zielinska et al., 2009).

Discussion

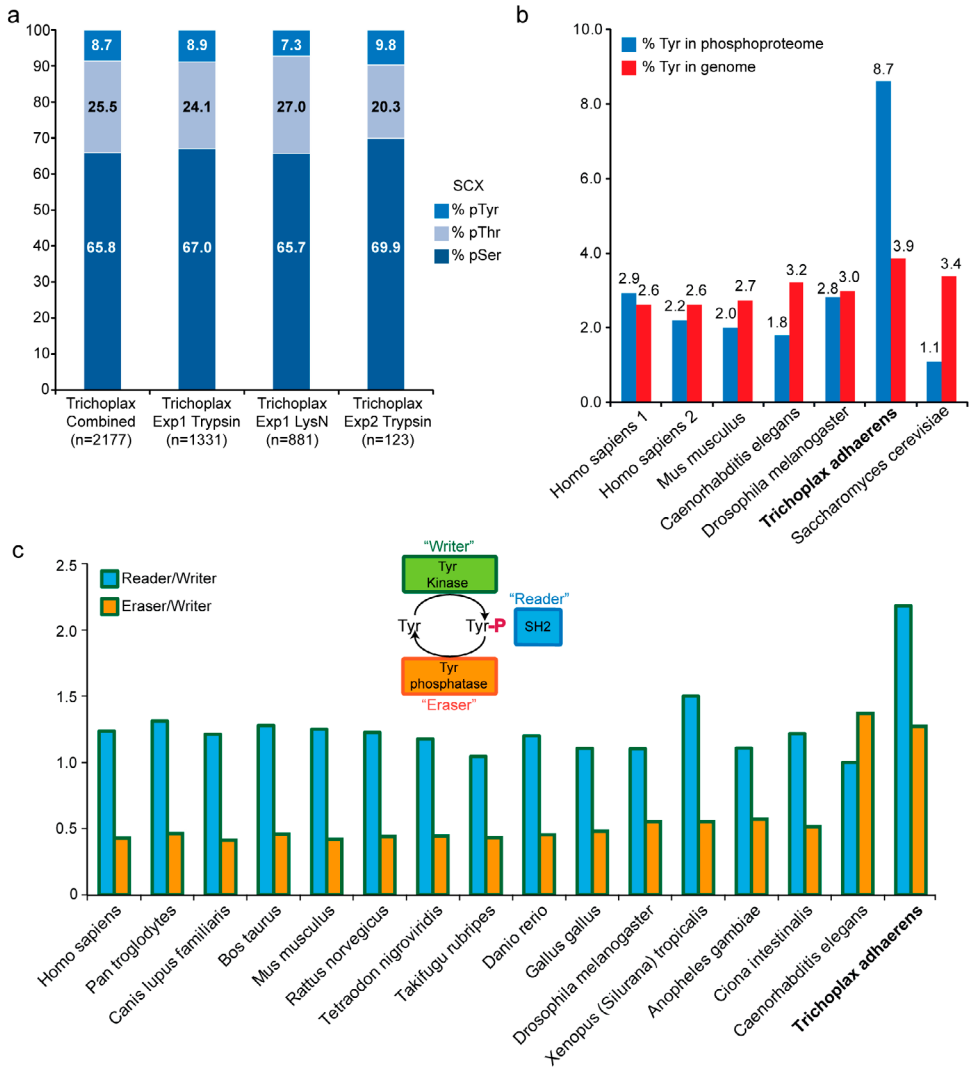
The study of such a simple and probably most ancient extant multicellular animal as Trichoplax is highly significant to gain insight into the emergence of metazoan multi cellularity, early animal evolution and subsequent metazoan diversity. Additionally, it can be used as a simple multicellular model system to shed light on the origins and functioning of many biological processes essential in higher metazoa, like cellular differentiation, cell-cell communication, development and basic animal patterning. Therefore, it was particularly interesting to observe expression of proteins important in processes thought to be characteristic of more complex bilaterian animals such as genes involved in developmental signalling pathways, i.e. Notch, Wnt and TGF-β pathways (Supplementary Fig. S3). Also, although nerve cells and ECM seemed to be absent in Trichoplax, proteins involved with neuro-endocrine processes as well as putative ECM proteins and other ECM components were observed.

Ranking all proteins based on number of spectral counts we detected a protein that is homologous to the Saccharomyces cerevisiae BSD1 alkyl-sulfatase among the top 10 proteins in all SCX experiments (Supplementary Table S1 and Fig. 1 and S2). Interestingly, this yeast protein has been identified as being acquired by horizontal gene transfer from proteobacteria (Hall et al., 2005). The very abundant Trichoplax BSD1 enzyme belongs to a group of less well-characterized alkylsulfatases known from yeast, bacteria, and very few higher animals (e.g. pea aphid and vase tunicate) that feed on chloroplast rich diets(Remm et al., 2001). We hypothesize that the presence of this enzyme in Trichoplax is linked to its unique feeding mode and its food-source. It has been suggested that the first animals that appeared were "grazers" feeding on the cyanobacterial and algal mats of the oceans (Gehling, 1999). In this context Trichoplax has been associated with fossils of Dickinsonia, whose motile feeding mode most closely

Figure 3 A burst of Tyrosine phosphorylation in Trichoplax. **(a)** Replicate measurements of tyrosine phosphorylation by SCX, with in the first bar the combined result of three independent SCX experiments, second bar the first trypsin experiment, third bar the LysN experiment and in the fourth bar an additional trypsin repeat experiment (n=total number of unique phosphosites (S, T, Y)). **(b)** Compared to other organisms Trichoplax exhibits a high percentage (3.9%) of tyrosine amino acids in its genome (red bars). The percentage of detected tyrosine phosphosites in Trichoplax phosphoproteomics datasets is 4-5 fold higher than detected in large-scale phosphoproteomics datasets in for other organisms, including H. sapiens (Rigbolt et al., 2011; Zhou et al., 2011), M. musculus (Huttlin et al., 2010), D. melanogaster (Zhai et al., 2008) and S. cerevisiae (Zielinska et al., 2009) (blue bars, supplementary table S4). Only species for which comprehensive protein phosphorylation data is also available, obtained using alike protocols, are shown. **(c)** Trichoplax contains a relative high number of readers (SH2 phosphotyrosine recognition domains) and erasers (tyrosine phosphatases), compared to writers (tyrosine kinases) involved in tyrosine signalling. Tyrosine-kinase domains, SH2 domains and protein tyrosine phosphatase domains were detected using HMM models from SMART (Letunic et al., 2009) using the online SMART tool (http://smart.embl-heidelberg.de/). Inset shows phosphotyrosine signalling as a tripartite system comprising tyrosine kinase (writer), tyrosine phosphatase (eraser) and SH2 domain (reader).

resembled that of Trichoplax (Garrett, 1970). It moves over and on top of its food, which is then digested externally by uptake of the released nutrients. Trichoplax still uses this very ancient feeding mode(Sperling and Vinther, 2010). The food of the first animals consisted of phototrophic organisms containing chloroplasts, forming a source of sulfolipids and other long chain alkyl-sulfates (Dembitsky and Srebnik, 2002; Haines, 1973). The high abundance of the BSD1 enzyme, that is able to scavenge sulfate from organically-bound sources, is an important advantage for an extremely primitive animal without a gut and its associated flora.

Most unanticipated was the observation of the relatively high number of tyrosine phosphorylation sites (i.e. 8.9, 7.3 and 9.8% of the observed phosphopeptides in the three individual SCX datasets, Fig 3a), especially when taking into account that no common tyrosine phosphorylation enrichment procedures (Boersema et al., 2010) were used. To further independently confirm these high levels of tyrosine phosphorylation, we performed phosphopeptide enrichment experiments using Ti-IMAC affinity beads (Zhou et al., 2011). We performed these experiment, in duplo, on similar protein amounts of Trichoplax and human HeLa cell lysates. Although, the sample amount was limited we identified around 1,000 phosphopeptides in each of the four experiments (Supplementary Fig. S6). Using the Hela



cells as a control revealed that the percentage of tyrosine phosphorylated peptides observed, using Ti4+-IMAC, is reduced by two-fold when compared to the SCX experiment (i.e. from ~3 to 1.5 % pTyr, see Fig. 3b). It is well-known that by applying certain PTM enrichment strategies one may introduce potential biases towards certain physicochemical properties of the targeted modification. In this specific case, our prior experience with Ti4+-IMAC elucidated that it favours serine and also threonine phosphorylation over tyrosine phosphorylation, as further evidenced here.

Notwithstanding, the relative ratio of tyrosine phosphorylated peptides detected in the Trichoplax Ti4+-IMAC phosphoproteomics experiments is still >two-fold higher than detected in the human HeLa cell datasets, confirming independently that pTyr is relatively more present in Trichoplax.

Why then does Trichoplax have such a high relative frequency of phosphorylation on tyrosine residues? Tyrosine phosphorylation provides a molecular system for transmitting cellular regulatory information that appeared ~ 600 million years ago, close to the appearance of Trichoplax, and has been associated with the advent of multicellularity (Lim and Pawson, 2010; Pincus et al., 2008; Tan et al., 2009). The basic repertoire of metazoan tyrosine kinases already existed before the advent of metazoan multicellularity, before the divergence of filasterians from metazoa and choanoflagellates. However, at the onset of metazoan multicellularity probably recruitment of receptor tyrosine kinases as a communication tool between cells led to huge diversifications of these kinases between pre-metazoan and metazoan lineages (Suga et al., 2012). The current view is that tyrosine phosphatases and Src Homology 2 (SH2) domains had already evolved in earlier organisms, prior to the appearance of dedicated protein tyrosine kinases. With that view in mind, we performed genome-wide analysis of the tyrosine content and the number of tyrosine kinases ("writers"), SH2 domains ("readers"), and phosphatase domain ("erasers") as predicted by SMART (Letunic et al., 2009), including genomes of 16 different species (Fig. 3c and Supplementary Table S5). Of all species analysed, Trichoplax contains the highest percentage of tyrosine amino acids in the predicted proteins (i.e. 3.9%) whereas these numbers drop to ~2.6% in mammals (Fig. 3b). As predicted by SMART Trichoplax contains 11 tyrosine kinases, 24 SH2 domain containing proteins and 14 phosphatases. Most strikingly, the ratio of reader to writer is exceptionally high for Trichoplax, (Fig. 3c) indicating that the substrate-to-enzyme ratio is very favourable, possibly forming the basis for the relatively high phosphotyrosine count in our data. Together these data provide strong experimental support of the concept of a sudden burst in tyrosine phosphorylation signalling at the beginning of metazoan multicellularity, followed by a gradual streamlining of phosphotyrosine signalling after the appearance of tyrosine kinases by reducing the number of possible deleterious phosphorylation sites as

tyrosine kinase numbers increase (Lim and Pawson, 2010; Pincus et al., 2008; Tan et al., 2009, 2011).

Our in-depth proteomics data also allows improvements to be made to the Trichoplax genome annotation and gene models by matching mass spectra directly onto the genome sequences (Supplementary Fig. S7). Furthermore, besides the discussed highly remarkable features of the Trichoplax proteome, the presented dataset also contains information on N-acetylation, Lys-acetylation, and the abundance of proteins involved in many different pathways, making the dataset a resource for researchers interested in the mechanisms of the origin and diversification of metazoan multi-cellularity.

Acknowledgements

This work was in part supported by the PRIME-XS project, Grant Agreement Number 262067, funded by the European Union Seventh Framework Program and by the German Science Foundation (Schi-277/20-3 and 26-1). The Netherlands Proteomics Centre, embedded in The Netherlands Genomics Initiative is acknowledged for funding. ME acknowledges funding by the Evangelisches Studienwerk e.V. Villigst and the German Academic Exchange service (DAAD).

Methods:

Animal culture and harvesting.
Animals of the so-called "Grell" (Schierwater, 2005) clone were cultured in artificial seawater (ASW) at 23 °C and at an LD regime of 16:8h. Animals were fed ad libitum on the green alga Pyrenomonas helgolandii, but starved for 24h prior to protein extraction in order to avoid contamination with the food(Eitel et al., 2011; Schierwater, 2005). Approximately 2,800 individual animals were transferred to sterile 6-well plates and washed three times each on two successive days with sterile ASW. Afterwards animals were transferred to a sterile 1.5 ml eppendorf tube (approximately 400 animals per tube), pelleted using a table centrifuge and washed with cold ASW (4°C). For the replicate experiment approximately 1,000 animals were harvested under the same conditions, delivering approximately

160 mg of protein.

Sample preparation.

After the washing procedure, animal pellets were lysed in 8 M Urea, 50 mM ammonium bicarbonate and EDTA-free protease inhibitor cocktail (Sigma). After homogenization, lysates were cleared by centrifugation at 13.000g for 20 mins at 4 °C and supernatant was snap-frozen until digestion. After reductive alkylation of cysteine residues using 2 mM DTT and 5mM iodoacetamide, 180 µg protein from approximately 1400 animals was digested with 2 µg Lys-C (Roche Diagnostics, Ingelheim, Germany) in 500 µl 8M Urea 50 mM ammonium bicarbonate for 4 hrs at 37 °C, followed by digestion with 4 µg trypsin (Roche Diagnostics, Ingelheim, Germany) in 2 M Urea and 50 mM ammonium bicarbonate at 37 °C for 16 hrs. Consecutively, 225 µg protein was digested with 2.5 µg Lys-N (U-Protein Express, Utrecht, The Netherlands) in 8 M Urea, 50 mM ammonium bicarbonate for 4 hrs at 37 °C followed by digestion with 3.5 µg Lys-N for 16 hrs at 37 °C in 4 M Urea, 50 mM ammonium bicarbonate. Peptides were desalted using 1 ml Sep Pack C18 columns (Waters), and separated by strong cation exchange chromatography (SCX) using a Zorbax BioSCX-Series II column (0.8-mm inner diameter, 50 mm length, 3.5 µm). SCX solvent A consists of 0.05 % formic acid in 20 % ACN, while solvent B was 0.05 % formic acid, 0.5 M NaCl in 20 % ACN. The SCX salt gradient was as follows: 0–0.06 min (0–2 % B); 0.06–10.06 min (2–3 % B); 10.06–20.06 min (3–8 % B); 20.06–30 min (8–20 % B); 30–40 min (20–40 % B); 40–46 min (40–90 % B); 46–50 min (90 % B). Fractionated peptides were dried and resuspended in 10% formic acid. Thirty-seven fractions from both digests were then each analyzed twice by reversed-phase LC–MS/MS.

Phosphopeptide Enrichment

Ti^{4+}-IMAC material was prepared and used essentially as previously described by us (21, 48). Affinity material was loaded onto Gel-loader tip microcolumns using a C8 plug and ~1–2cm length of material. The columns were pre-equilibrated with 2 × 30 mL of Ti-IMAC loading buffer (80% ACN, 6% TFA). Next, samples were, resuspended in 60 mL loading

buffer and loaded onto the equilibrated gel-loader tip microcolumns. Columns were sequentially washed with 60 mL of loading buffer, followed by washing with 60 mL of 50% ACN/0.5% TFA containing 200 mM NaCl and additional washing by 60 mL of 50% ACN/0.1% TFA. The bound peptides were eluted by 20 mL of 5%ammonia, followed by a second elution with 80% ACN/6%FA, into 20 mL of 10% formic acid and then stored at -20 °C for LC-MS analysis.

LC-MS/MS.

Peptide fractions were analyzed using an Agilent 1100-Series LC system coupled to an LTQ-Orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany). The LC system was equipped with a 20 mm Aqua C18 (Phenomenex, Torrance, CA) trapping column (packed inhouse, i.d., 50 μm; resin 5 μm) and a 400 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH, Ammerbuch, Germany) analytical column (packed in-house, i.d. 50 μm; resin 3 μm). Trapping was performed at 5 μL min-1 for 10 min, and elution was achieved with a gradient of 0–13% B in 0.1 min, 13–28% B in 107 min, 28–50% B in 35 min, 50-100% B for 2 min. The flow rate was passively split from 0.35 mL min-1 to 50 nL min-1. Nanospray was achieved using a coated fused silica emitter (New Objective, Cambridge, MA) (o.d., 360 μm; i.d., 20 μm, tip i.d. 10 μm) biased to 1.7 kV. The mass spectrometer was operated in data dependent mode to switch between MS and MS/MS. The five most intense ions were selected for fragmentation in the linear ion trap using collisionally induced dissociation at a target value of 30,000.

Database search and validation.

Spectra were processed with Maxquant software to generate peak lists which were then analyzed with Mascot search engine version 2.3.02 (Matrix Science, London, UK) using a concatenated forward/reverse database of the Trichoplax Triad1-best-proteins-fasta sequences, setting carbamidomethyl (C) as fixed and oxidation (M) and acetylation (protein N-term) as variable modifications. Maximum 2 missed cleavages were allowed, peptide tolerance was set to 50 ppm and MS/MS tolerance to 0.6 Da. For phosphorylation analysis, carbamidomethyl (C) was set as fixed, and phospho (STY) and oxidation (M) were set as variable modifications with

maximum 1 missed cleavage allowed. For lysine acetylation analysis, carbamidomethyl (C) was set as fixed, and acetyl (K) and oxidation (M) were set as variable modifications with maximum 2 missed cleavages allowed. Search results were filtered with Rockerbox (van den Toorn et al., 2011) to an FDR of 1% using the concatenated database decoy method.

Protein abundance calculations.

Numbers of identified spectra were used to calculate protein abundances (Aye et al., 2010). Briefly, abundance factors were calculated from the number of identified spectra of a particular protein, divided by its molecular weight, multiplied by 10^6 for clarity. The abundance factor of a particular protein as a fraction of the total sum of all abundance factors was multiplied by the total amount of protein material used in each experiment and from this the numbers of protein copy numbers were calculated by dividing by the protein molecular weights. This was divided by the number of animals used to get the number of protein copies per animal.

References

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S.G., and Pandey, A. (2007). A curated compendium of phosphorylation motifs. Nat Biotechnol *25*, 285–286.

Aye, T.T., Scholten, A., Taouatas, N., Varro, A., Van Veen, T.A.B., Vos, M.A., and Heck, A.J.R. (2010). Proteome-wide protein concentrations in the human heart. Mol. Biosyst. *6*, 1917–1927.

Boersema, P.J., Foong, L.Y., Ding, V.M., Lemeer, S., van Breukelen, B., Philp, R., Boekhorst, J., Snel, B., den Hertog, J., Choo, A.B., et al. (2010). In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. Mol. Cell. Proteomics *9*, 84–99.

Dembitsky, V.M., and Srebnik, M. (2002). Natural halogenated fatty acids: their analogues and derivatives. Prog Lipid Res *41*, 315–367.

Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res *39*, D261-7.

Eitel, M., Guidi, L., Hadrys, H., Balsamo, M., and Schierwater, B. (2011). New Insights into Placozoan Sexual Reproduction and Development. PLoS One *6*.

Garrett, P. (1970). Phanerozoic stromatolites: noncompetitive ecologic restriction by grazing and burrowing animals. Science (80-.). *169*, 171–173.

Gehling, J.G. (1999). Microbial mats in terminal Proterozoic siliciclastics: Ediacaran death masks. Palaios *14*, 40–57.

Gnad, F., de Godoy, L.M.F., Cox, J., Neuhauser, N., Ren, S., Olsen, J. V, and Mann, M. (2009). High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. Proteomics *9*, 4642.

Gnad, F., Gunawardena, J., and Mann, M. (2011). PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res *39*, D253-60.

Guidi, L., Eitel, M., Cesarini, E., Schierwater, B., and Balsamo, M. (2011). Ultrastructural analyses support different morphological lineages in the phylum placozoa Grell, 1971. J. Morphol. *272*, 371–378.

Haines, T.H. (1973). Halogen- and sulfur-containing lipids of Ochromonas. Annu Rev Microbiol *27*, 403–411.

Hall, C., Brachat, S., and Dietrich, F.S. (2005). Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. Eukaryot Cell *4*, 1102–1115.

Hornbeck, P. V, Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2011). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res *40*, D261-70.

Hulpiau, P., and van Roy, F. (2011). New insights into the evolution of metazoan cadherins. Mol Biol Evol *28*, 647–657.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villen, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. Nucleic Acids Res *37*, D229-32.

Lim, W.A., and Pawson, T. (2010). Phosphotyrosine signaling: evolving a new cellular communication system. Cell *142*, 661–667.

Orosz, F. (2009). Apicortin, a unique protein, with a putative cytoskeletal role, shared only by apicomplexan parasites and the placozoan Trichoplax adhaerens. Infect Genet Evol *9*, 1275–1286.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L.L., Köstler, T., Messina, D.N., et al. (2009). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res *38*, D196-203.

Pincus, D., Letunic, I., Bork, P., and Lim, W.A. (2008). Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. Proc. Natl. Acad. Sci. U. S. A. *105*, 9680–9684.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol *314*, 1041–1052.

Rigbolt, K.T.G., Prokhorova, T.A., Akimov, V., Henningsen, J., Johansen, P.T., Kratchmarova, I., Kassem, M., Mann, M., Olsen, J. V, and Blagoev, B. (2011). System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. Sci. Signal. *4*, rs3.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotech *29*, 24.

Roos, F.F., Jacob, R., Grossmann, J., Fischer, B., Buhmann, J.M., Gruissem, W., Baginsky, S., and Widmayer, P. (2007). PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. Bioinformatics *23*, 3016.

Schierwater, B. (2005). My favorite animal,Trichoplax adhaerens. BioEssays *27*, 1294–1302.

Schierwater, B., and Kuhn, K. (1998). Homology of Hox genes and the zootype concept in early metazoan evolution. Mol. Phylogenet. Evol. *9*, 375–381.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Dellaporta, S.L., Kolokotronis, S.-O., and DeSalle, R. (2009). Concatenated Analysis Sheds Light on Early Metazoan Evolution and Fuels a Modern "Urmetazoon" Hypothesis. PLoS Biol. *7*, e1000020.

Signorovitch, A.Y., Dellaporta, S.L., and Buss, L.W. (2005). Molecular signatures for sex in the Placozoa. Proc. Natl. Acad. Sci. U. S. A. *102*, 15518–15522.

Sperling, E.A., and Vinther, J. (2010). A placozoan affinity for Dickinsonia and the evolution of late Proterozoic metazoan feeding modes. Evol Dev *12*, 201–209.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., et al. (2008). The Trichoplax genome and the nature of placozoans. Nature *454*, 955.

Suga, H., Dacre, M., de Mendoza, A., Shalchian-Tabrizi, K., Manning, G., and Ruiz-Trillo, I. (2012). Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci. Signal. *5*, ra35.

Tan, C.S., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009). Positive selection of tyrosine loss in metazoan evolution. Science (80-.). *325*, 1686–1688.

Tan, C.S.H., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2011). Response to Comment on " Positive Metazoan Evolution ." Science (80-.). *917*.

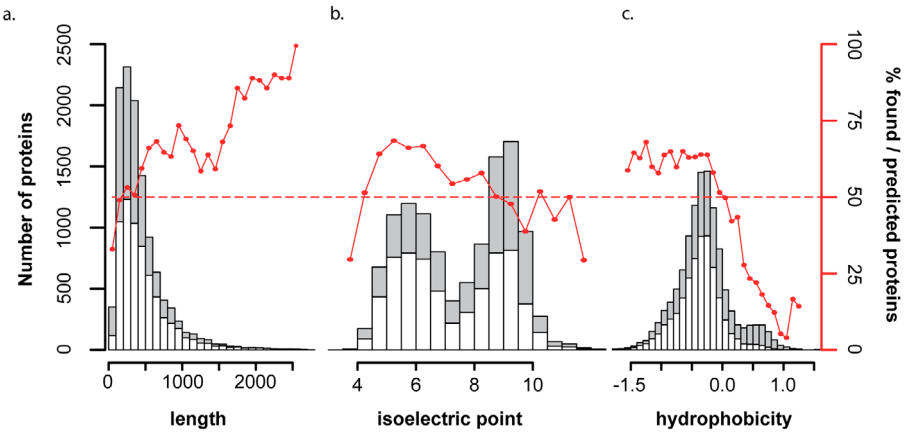
van den Toorn, H.W.P., Muñoz, J., Mohammed, S., Raijmakers, R., and Heck, A.J.R. (2011). RockerBox: analysis and filtering of massive proteomics search results. J. Proteome Res. *10*, 1420–1424.

Zhai, B., Villen, J., Beausoleil, S.A., Mintseris, J., and Gygi, S.P. (2008). Phosphoproteome analysis of Drosophila melanogaster embryos. J Proteome Res *7*, 1675–1682.

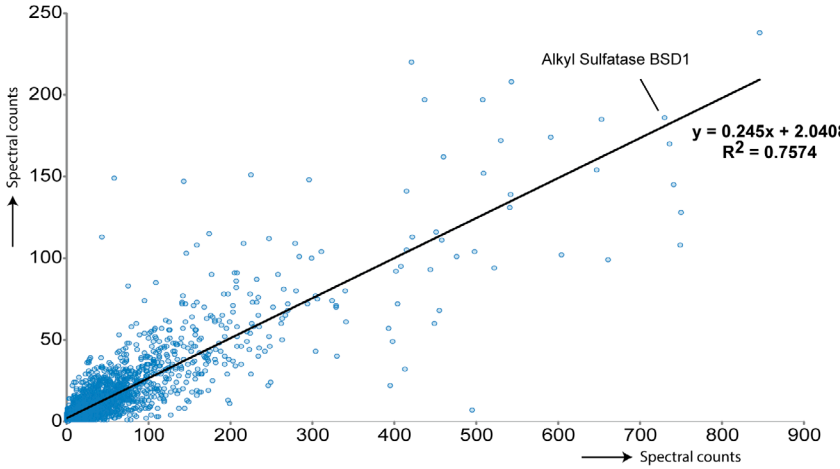
Zhou, H., Low, T.Y., Hennrich, M.L., van der Toorn, H., Schwend, T., Zou, H., Mohammed, S., and Heck, A.J.R. (2011). Enhancing the identification of phosphopeptides from putative basophilic kinase substrates using Ti (IV) based IMAC enrichment. Mol. Cell. Proteomics *10*, M110.006452.

Zielinska, D.F., Gnad, F., Jedrusik-Bode, M., Wisniewski, J.R., and Mann, M. (2009). Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. J Proteome Res *8*, 4039–4049.

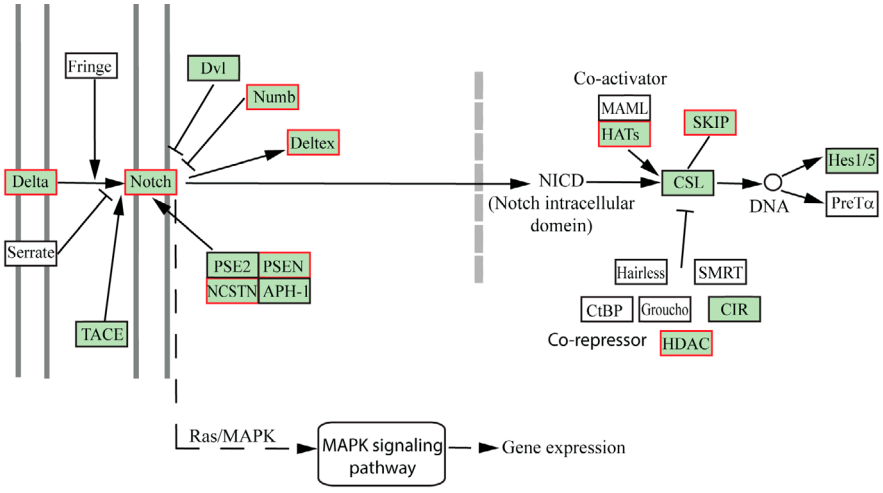
Supplementary figures



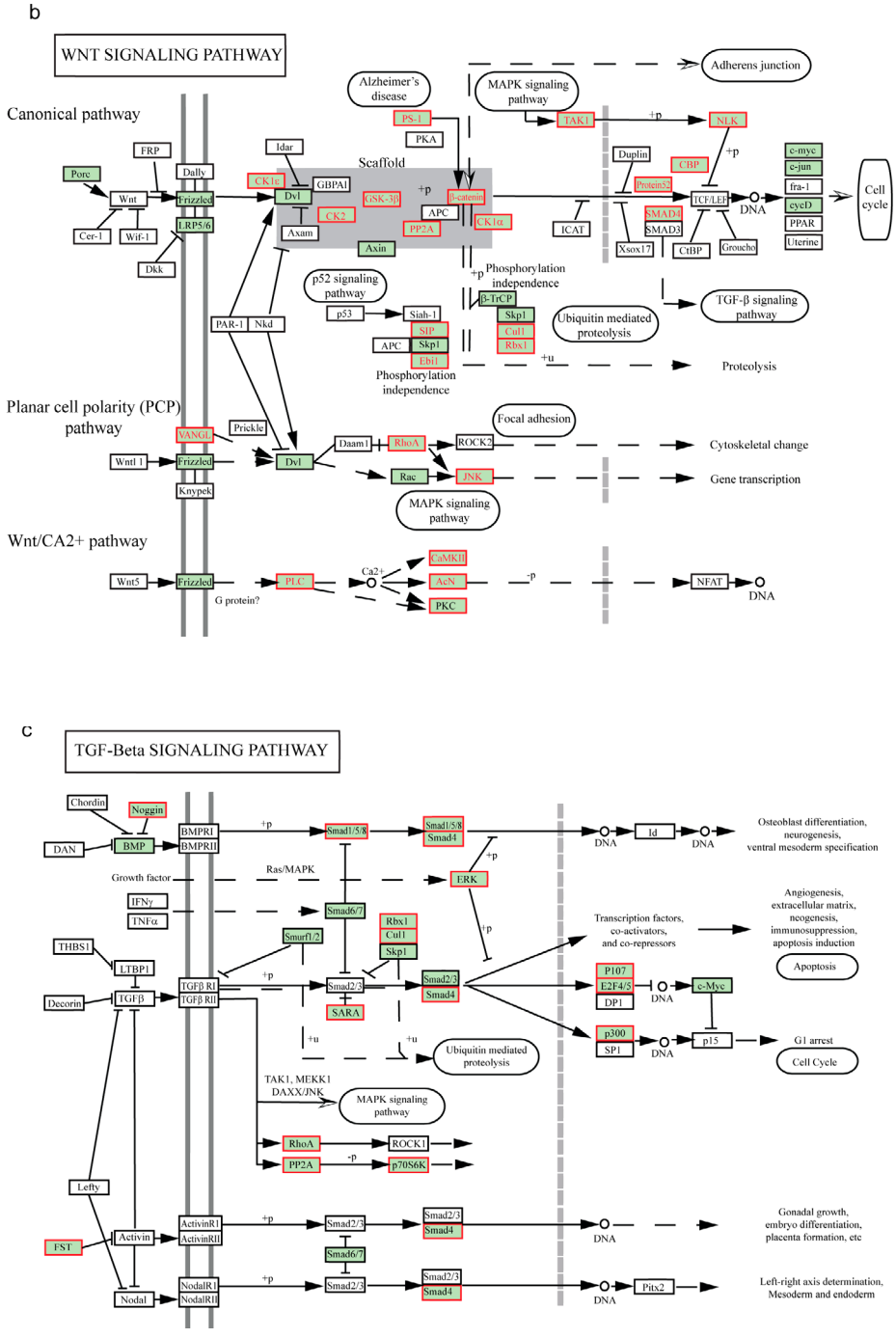
Supplementary Figure S1 Bias analysis of all identified *Trichoplax* proteins. Properties for all proteins in SwissProt were calculated with BioPython 1.58 module ProtParam and plotted as a histogram (grey bars, corresponding to the left vertical axis). The analysis was repeated for all significantly detected proteins (white bars overlapping the grey bars). The proportion of the known and detected proteins for each histogram bar was calculated and depicted as red dots (rightmost axis), connected with lines for clarity. **(A)** There is a clear preference for longer proteins for detection by bottom-up mass spectrometry which is in accordance with the notion that longer proteins yield more peptides, hence are more likely to be detected. **(B)** The Isoelectric point does not have a clear influence on mass spectrometry coverage which is around 50%. **(C)** A clear preference for hydrophilic proteins exists in our analysis.

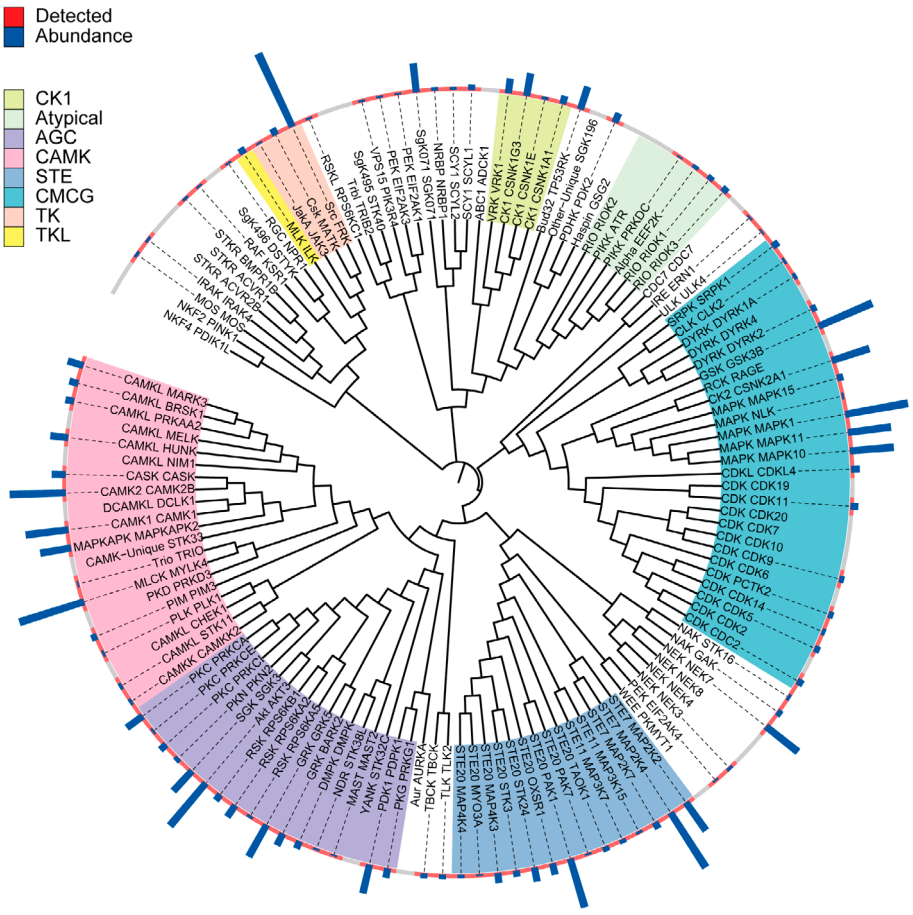


Supplementary Figure S2 Correlation between two LC-MS experiments of trypsin digested *Trichoplax*. The plot shows the correlation between number of spectral counts between the two experiments of 3839 proteins that are in common between the two replicate experiments.

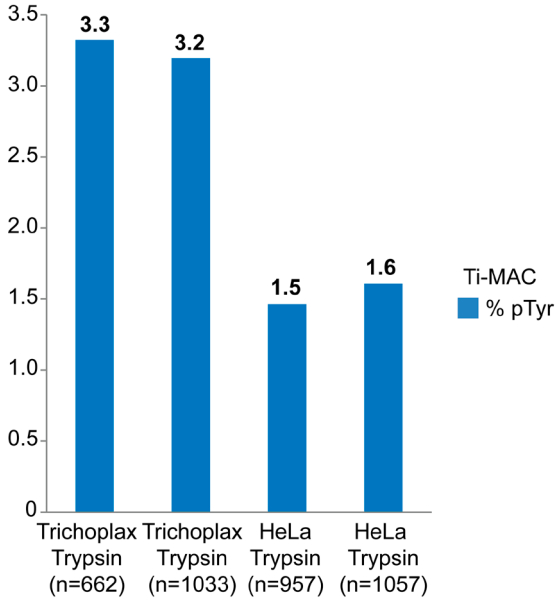


Supplementary Figure S3 KEGG Pathways. Our protein identifications are mapped onto KEGG (van den Toorn et al., 2011) reference pathway maps at http://www.genome.jp/kegg/kegg2.html. Color green indicates the annotation of a Trichoplax ortholog in KEGG, red indicates identification of the Trichoplax ortholog in our dataset. **(A)** Notch signaling pathway map-04330, **(B)** Wnt signaling pathway map-04310, and **(C)** TGF-β signaling pathway map-04350.

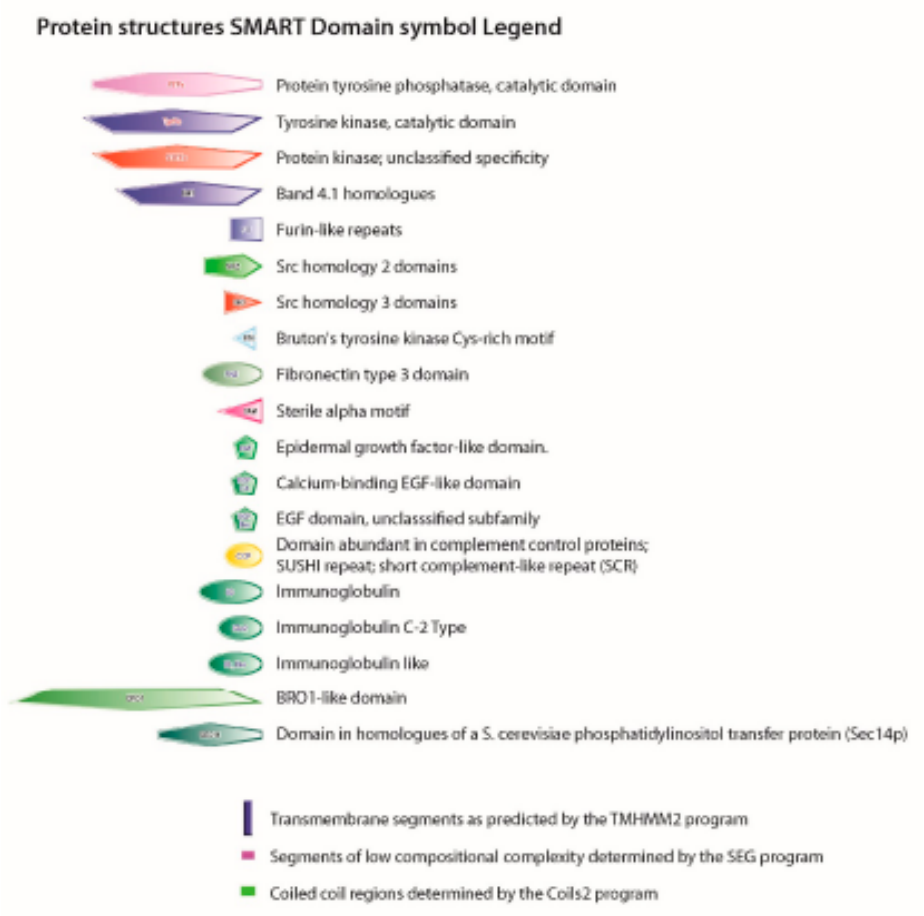

b
WNT SIGNALING PATHWAY
Canonical pathway
Alzheimer's disease
MAPK signaling pathway
Adherens junction
PS-1
PKA
TAK1
NLK
Scaffold
FRP
Idar
Porc
Dally
Wnt
Frizzled
LRP5/6
Cer-1
Wif-1
Dkk
CK1ε
GBPAI
Dvl
CK2
Axam
Axin
GSK-3β
β-catenin
APC
PP2A
CK1α
ICAT
Duplin
CBP
Protein22
TCF/LEF
DNA
SMAD4
SMAD3
Xsox17
CtBP
Groucho
c-myc
c-jun
fra-1
cycD
PPAR
Uterine
Cell cycle
p52 signaling pathway
p53
Siah-1
SIP
APC
Skp1
Ebi1
β-TrCP
Skp1
Cul1
Rbx1
Phosphorylation independence
Ubiquitin mediated proteolysis
TGF-β signaling pathway
Proteolysis
PAR-1
Nkd
Planar cell polarity (PCP) pathway
VANGL
Prickle
Wntl 1
Frizzled
Knypek
Dvl
Daam1
RhoA
ROCK2
Rac
JNK
Focal adhesion
Cytoskeletal change
Gene transcription
MAPK signaling pathway
Wnt/CA2+ pathway
Wnt5
Frizzled
G protein?
PLC
Ca2+
CaMKII
AcN
PKC
NFAT
DNA
c
TGF-Beta SIGNALING PATHWAY
Chordin
Noggin
DAN
BMP
BMPRI
BMPRII
Smad1/5/8
Smad1/5/8
Smad4
DNA
Id
DNA
Osteoblast differentiation, neurogenesis, ventral mesoderm specification
Growth factor
Ras/MAPK
ERK
IFNγ
TNFα
Smad6/7
Smurf1/2
Rbx1
Cul1
Skp1
Transcription factors, co-activators, and co-repressors
Angiogenesis, extracellular matrix, neogenesis, immunosuppression, apoptosis induction
Apoptosis
THBS1
LTBP1
Decorin
TGFβ
TGFβ RI
TGFβ RII
Smad2/3
Smad2/3
Smad4
SARA
P107
E2F4/5
DP1
DNA
c-Myc
p300
SP1
DNA
p15
G1 arrest
Cell Cycle
Ubiquitin mediated proteolysis
TAK1, MEKK1
DAXX/JNK
MAPK signaling pathway
RhoA
ROCK1
PP2A
p70S6K
Lefty
FST
Activin
ActivinRI
ActivinRII
Smad2/3
Smad2/3
Smad4
Smad6/7
DNA
Gonadal growth, embryo differentiation, placenta formation, etc
NodalRI
NodalRII
Nodal
Smad2/3
Smad2/3
Smad4
DNA
Pitx2
Left-right axis determination, Mesoderm and endoderm



Supplementary Figure S4 Expression of *Trichoplax* kinases with Human orthologs. Human protein kinase orthologs for all Trichoplax protein kinases were determined by Inparanoid (Ostlund et al., 2009). FASTA sequences of the kinase domains of the human orthologs were retrieved from the KinBase resource (http://kinase.com/kinbase/) and aligned using ClustalX2.1 (Larkin et al., 2007) using default parameters for multiple alignment and bootstrapping. A phylogenetic tree was calculated with the neighbor-joining algorithm, exported and loaded into the Interactive Tree of Life tool (Letunic and Bork, 2007). Kinase families are colored by grouping into kinase families as defined in the KinBase resource (http://kinase.com/kinbase/). On the outer rim the color red indicates that the *Trichoplax* kinase is detected, blue bars indicate the relative abundances as copies / animal. Human ortholog names are listed by their grouping into kinase groups and families.



Supplementary Figure S6 Percentage of tyrosine phosphorylation in Ti-IMAC enriched phosphopeptide datasets. First two bars represent Ti-IMAC enrichment experiments on *Trichoplax* and the second two bars represent Ti-IMAC enrichment on human HeLa cells. Although, the fold difference is smaller the percentage of detected tyrosine phosphosites in *Trichoplax* is still substantially higher than in human HeLa cells. (n=total # of unique phosphopeptides identified per Ti-IMAC experiment).



Supplementary Figure S5 Trichoplax adhaerens SMART predicted kinase and phosphatase structures An overview of the kinase and phosphatase domains in the Trichoplax genome, as predicted by the SMART algorithm (Letunic et al., 2009). Since the "best genes" gene prediction included in the SMART database mostly reflects only the catalytic domain of the kinases, longer predicted protein sequences surrounding the SMART annotated genes were obtained from the JGI genome browser (http://genome.jgi-psf.org), mostly based on the fgenesh algorithm. In one instance of a kinase domain (B3RT13), the fgenesh algorithm predicted two separate open reading frames very close together. Assuming the algorithm incorrectly omitted a splice site, a complete protein structure encompassing fibronectin type 3 domains, a transmembrane domain and a kinase domain is likely to be encoded, similar to two other predicted genes highly homologous to insulin receptor-related proteins. SMART annotates several predicted protease domains as inactive. We have reported them here since no assays have been performed to confirm this. The protein "B3S1F1" contains two equal scoring Immunoglobin C2-type domains of different sizes, which were both collapsed into the same cartoon for legibility.

Trichoplax adhaerens SMART predicted kinase structures

Uniprot Accession JGI Identifier - genomic location

If used: alternative gene model containing the SMART domains are shown in gray.

B3RT13 TRIADDRAFT_4397 - scaffold_3:7169187-7171461 [fgeneshTA2_pg.C_scaffold_3000807 + fgeneshTA2_pg.C_scaffold_3000806]

B3RT14 TRIADDRAFT_12208 - scaffold_3:7184204-7186415 [fgeneshTA2_pg.C_scaffold_3000808]

B3RT16 TRIADDRAFT_22855 - scaffold_3:7199488-7201856 [fgeneshTA2_pg.C_scaffold_3000810]

B3S184 TRIADDRAFT_27547 - scaffold_7:3969419-3970848 [fgeneshTA2_pg.C_scaffold_7000452]

B3S2E6 TRIADDRAFT_28106 - scaffold_7:1985055-1985870 [fgeneshTA2_pg.C_scaffold_7000214]

B3RWF6 TRIADDRAFT_24853 - scaffold_5:6179281-6182174 [fgeneshTA2_pg.C_scaffold_5000625]

B3S5N3 TRIADDRAFT_50752 - scaffold_11:2348062-2349975

B3S2E4 TRIADDRAFT_50482 - scaffold_7:1972362-1973243 [fgeneshTA2_pg.C_scaffold_7000212]

B3S3G7 TRIADDRAFT_12344 - scaffold_8:2553643-2557133 [fgeneshTA2_pg.C_scaffold_8000263]

B3RL37 TRIADDRAFT_51865 - scaffold_1:5469567-5480863

B3RWV1 TRIADDRAFT_25090 - scaffold_5:7454667-7456059 [fgeneshTA2_pg.C_scaffold_5000780]

Trichoplax adhaerens SMART predicted phosphatase structures

Uniprot Accession JGI Identifier - genomic location

If used: alternative gene model containing the SMART domains are shown in gray.

B3S834 TRIADDRAFT_60390 - scaffold_15:571260-579048

B3S1F8 TRIADDRAFT_27423 - scaffold_7:344953-348330 [fgeneshTA2_pg.C_scaffold_7000033] *

B3RRB0 TRIADDRAFT_54170 - scaffold_3:1467464-1476407

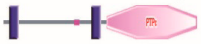
B3SCC0 TRIADDRAFT_61918 - scaffold_31:393183-399332

B3S630 TRIADDRAFT_64201 - scaffold_12:35087-54047 *

B3RQJ8 TRIADDRAFT_63802 / 55015 - scaffold_3:8821914-8831157

B3RQN0 TRIADDRAFT_55048 - scaffold_3:9172847-9176657

B3SEH5 TRIADDRAFT_34519 - scaffold_281:6517-8156 [fgeneshTA2_pg.C_scaffold_281000002]



B3RIR1 TRIADDRAFT_52495 - scaffold_1:10167403-10175316

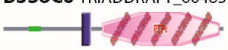
B3RQJ9 TRIADDRAFT_55016 - scaffold_3:8831643-8836222

B3S1F1 TRIADDRAFT_27501 - scaffold_7:258692-276916 [fgeneshTA2_pg.C_scaffold_7000025]

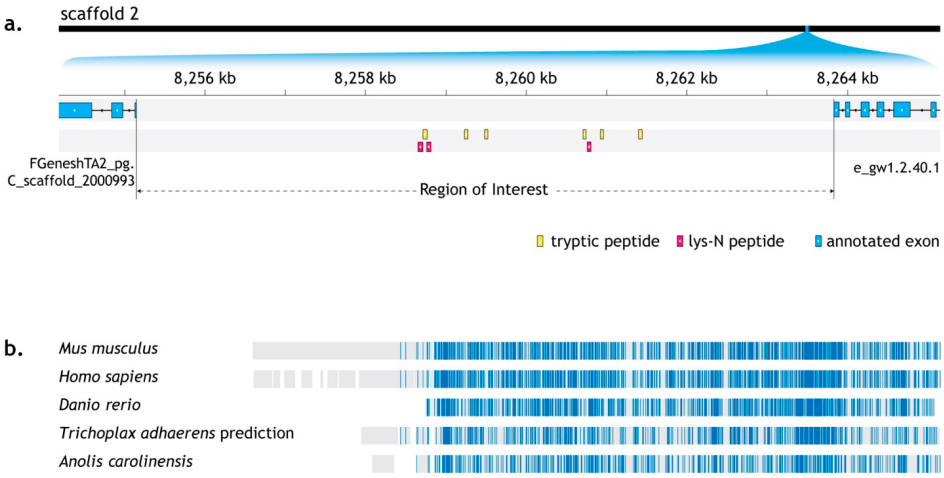
B3S1N2 TRIADDRAFT_57823 - scaffold_7:418035-423971

B3RJC0 TRIADDRAFT_51484 - scaffold_1:2150632-2158903

B3S8C0 TRIADDRAFT_60483 - scaffold_15:1439404-1445841



All protein structure cartoons have the same scale, except those marked with * which had to be scaled down to fit on the page.



Supplementary Figure S7 Discovery of a putative CCDC40 homologue. **(A)** Graphical representation of the analysis of peptides mapped to Trichoplax scaffold 2. We searched both the LysN and Trypsin datasets against all 1415 scaffolds with estimated false discovery rate of 1% using the Pepsplice (Roos et al., 2007) search engine, allowing for the presence of introns. In the Integrative Genomics Viewer (IGV (Robinson et al., 2011)) version 2.0.3 a cluster of both tryptic (yellow) and lys-N (red) digested peptides were found in a region where no gene model (blue) was mapped. The nucleotide sequence of the unmapped DNA was extracted using the IGV viewer ("region of interest") and used in a BLASTX search against the Swissprot database (www.uniprot.org). A gene prediction for this fragment was made using GeneMark.hmm, based on Caenorhabditis elegans ES-3.0 hidden markov model. **(B)** The predicted protein was aligned with Mus musculus, Homo sapiens, Danio Rerio and Anolis carolinensis sequences, obtained from the GeneCards database using ClustalX. An overview of the alignment is shown in, showing a high sequence similarity over most of the predicted protein sequence.

7. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis

Low, T.Y., Van Heesch, S., Van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., Van Breukelen, B., Mohammed, S., et al. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 5, 1469–1478.

Quantitative and qualitative protein characteristics are regulated at the genomic, transcriptomic and post-transcriptional levels. Here, we comprehensively interrogate their interplay by performing an in-depth transcriptome analysis and ultra-deep proteomics analysis of liver tissues from two rat strains. First, de novo transcriptome assemblies based on RNA-Seq were created for the BN-Lx (Brown Norway) and SHR (Spontaneously Hypertensive Rat) strains. This data was combined with strain-specific whole genome sequencing data and used to extend the publicly available rat protein reference database (32,791 entries) with novel protein predictions (44,993 GENSCANs), strain-specific protein variants (10,493 missense variants in 6,189 proteins), predicted transcript isoforms (2,545), and RNA-editing events (196). Synergistically, we performed ultra-deep proteomics analysis using two-dimensional ultrahigh performance chromatography with five proteases to obtain a bias-free coverage of the entire proteome. A total of 12,989 rat liver proteins were identified (FDR of 0%)- the deepest and most stringent rat proteome ever reported. Our data provides evidence for 792 GENSCAN predictions and allele-specific isoforms for 112 genes at the protein level. Furthermore, we detected 54 and 337 peptides confirming novel RNA editing and RNA splicing events, respectively, with some of these being strain-specific. Quantitative analysis revealed a very strong correlation between either transcriptome or proteome data from both strains ($r > 0.94$) indicating limited control by genetic variation. In contrast, a much weaker correlation was observed between quantitative RNA and protein data within a strain ($r \sim 0.44$). Finally, through integrative data analysis we were able to link a genomic variant in the promoter of the most differentially expressed gene at both the RNA and protein level to the hypertension phenotype of the SHR rat strain. This gene, Cyp17a1 is a member of the cytochrome P450 (CYP450) superfamily and has previously been identified as a top hit in GWAS studies for human hypertension. These results demonstrate both the power of and need for integrative genomic, transcriptomic and proteomic data analysis to start understanding (genetic) control of molecular dynamics and phenotypic diversity in a system wide manner.

Introduction

Mass spectrometry-based proteomics (MS) and next generation sequencing (NGS) are rapidly maturing techniques, each independently enabling comprehensive measurements of gene-products at a system level (Altelaar et al., 2013; Cox and Mann, 2011; Soon et al., 2013)). Although MS and NGS are highly complementary, they are rarely applied integratively in large-scale studies (Ning et al., 2012). State-of-the-art MS approaches can currently identify over ten thousand proteins in a single experiment (Munoz et al., 2011; Nagaraj et al., 2011), which brings the analysis of complete proteomes within reach (Ahrens et al., 2010; Cox and Mann, 2011). However, as long as non-customary protein databases that are derived from (typically incomplete) reference genome assemblies and annotations remain the sole source used for MS spectra matching (Perkins et al., 1999; Yates et al., 1995) true completeness will not be reached. For example, protein isoforms arising from genetic polymorphisms, post-transcriptional events such as RNA-editing and post-translational modifications are routinely missed (Jensen, 2004; Uhlen and Ponten, 2005).

Recent advances in NGS techniques, including whole genome sequencing and total RNA-sequencing allow for the generation of near-complete inventories of genetic variation in a system and its transcribed repertoire (Ozsolak and Milos, 2011). However, from such analysis, the effects on the proteins cannot be predicted with high confidence. For example, the consequence of a single nucleotide variant (SNV) on the coding capacity of a transcript can be predicted accurately, but not the potential effects on the stability of the corresponding protein. Systematic comparison of RNA-Seq data with genomic data reveals another layer of complexity – it has now been convincingly demonstrated that certain transcripts are modified by post-transcriptional editing, primarily by targeted A to I deamination (Farajollahi and Maas, 2010). It is expected that all these types of variation will not only affect the composition and function of a protein, but may also influence expression levels. However, the additional layers of translation control may dampen or completely abolish such effects (Kleinman and Majewski, 2012; Lin et al., 2012; Pickrell et al., 2012).

An integrative analysis of different data modalities, ideally from samples from a single source, is required for correctly deciphering the effects of genomic and transcriptomic variation on molecular processes and cellular

functioning. An example of such data integration is the use of proteomic data derived from MS in combination with complete genome data to improve gene annotation (Jaffe et al., 2004; Renuse et al., 2011). This approach has so far been sparsely performed and mainly in organisms with smaller genomes (Merrihew et al., 2008; Payne et al., 2010; Venter et al., 2011). On the other hand, integrative investigations of messenger RNA levels and the proteins they encode reveal only modest correlations, implying an unresolved level of complexity in regulation of expression (Nesvizhskii et al., 2006; Ning et al., 2012; Schwanhäusser et al., 2011; de Sousa Abreu et al., 2009; Vogel and Marcotte, 2012).

For this study, we selected two rat inbred strains BN-*Lx*/Cub (BN-*Lx*) and SHR/OlaIpcv (SHR) (Printz et al., 2003), representing widely studied, renewable and genetically homogeneous resources. Both strains have previously been extensively characterized at the genomic (Atanur et al., 2010; Gibbs et al., 2004) and phenotypic level (Hubner et al., 2005; Johnson et al., 2009; Pravenec and Kurtz, 2010; Pravenec et al., 2004; Printz et al., 2003; Simonis et al., 2012a). The BN-*Lx* strain is derived from, and thus very closely related to, the Brown Norway (BN) strain. The latter strain was used for creating the rat reference genome assembly (Gibbs et al., 2004), and is commonly used as the protein reference dataset in rat proteomics studies. The spontaneously hypertensive rat (SHR) is more diverged from the BN reference and is a widely used disease model for hypertension studies. Whereas several blood pressure quantitative trait loci (QTLs) have been mapped to the SHR genome, no functional variants driving elevated blood pressure have been identified to date. Here, we combine in-depth genomic, transcriptomic and proteomic analyses from inbred rats of two different genetic backgrounds using the same sets of rat liver tissues (Figure 1A). We determine quantitative and qualitative molecular dynamics at different functional levels and achieve a new level of proteome completeness by adding variation information derived from the whole genome sequencing and RNA sequencing data. These data allow us to apply a genome-wide genetical-genomics approach (Jansen and Nap, 2001) to start understanding multi-level systems regulation and to identify candidate genes that are potentially involved in the hypertension phenotype of the SHR rat.

Results & Discussion

Extension of the rat protein database

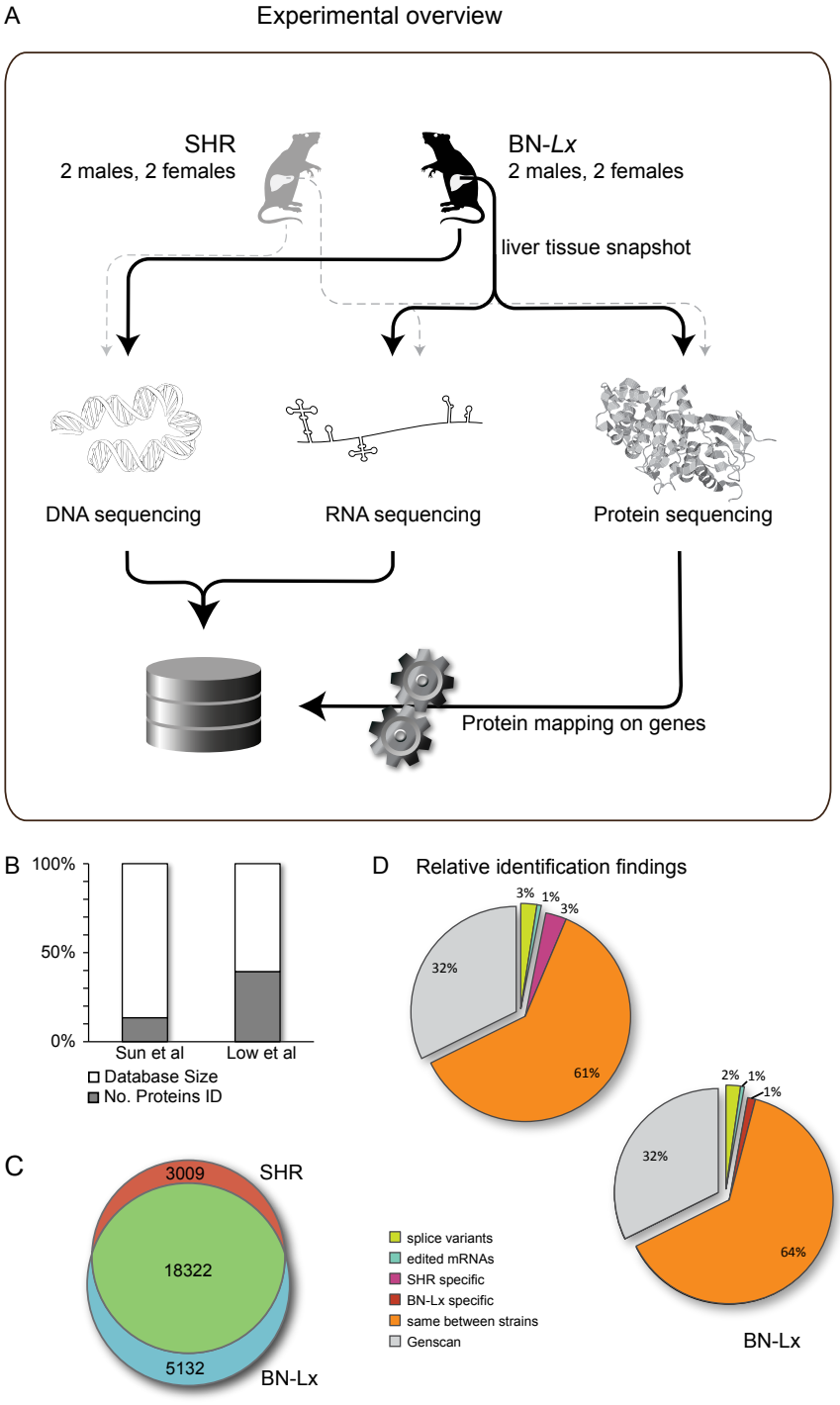
In proteomics, tandem mass spectra (of peptides) are typically annotated by searching against *in sillico* generated spectra based on a publicly available genome database. For rat, a genebuild based on reference genome assembly (Gibbs et al., 2004) is commonly used as the protein reference dataset. To create a sample-specific reference database for MS peptide searching, we extended the existing RefSeq-based peptide database by incorporating strain-specific peptides and predicted peptides. We first obtained all strain-specific genetic variation of the BN-*Lx* and SHR strains including single nucleotide variants (SNVs) and in-frame indels.

Most genomic SNVs are located in the non-coding sequences and only the less frequent non-synonymous variants in coding regions give rise to altered amino acid sequences (Hurst et al., 2009; Su et al., 2011; Valentine et al., 2006). We collected 10,493 non-synonymous variants from recently generated high-coverage whole genome sequencing data of the BN-*Lx* and SHR genomes (Atanur et al., 2010; Simonis et al., 2012b), which are predicted to affect 6,187 protein isoforms derived from 4,566 genes. Furthermore, to be able to detect *in silico* gene predictions using the proteomics data as evidence (Volkening et al., 2012), we added 44,993 GENSCAN gene predictions to our rat database (Burge and Karlin, 1997) .

Next, we generated RNA-seq data (Table S1) from liver tissue extracted from both rat strains (two males and two females per strain). This data was used to catalog post-transcriptional modifications such as alternative splicing and RNA editing. To this end, paired-end sequencing data was used to generate *de novo* transcriptome assemblies for each strain. In total, we identified 2,545 novel transcript splicing events affecting 1,015 genes. While the majority of them (1,687) were detected in both strains, the other 220 and 638 events were specific to BN-*Lx* and SHR rats, respectively (Table S2). Independent RT-PCR-based Sanger sequencing confirmed 62.3% (43 / 69) of a randomly sampled subset as new transcript isoforms (Table S3A). In addition, the same transcriptome data provided evidence for expression of 2,903 GENSCAN predictions (Table S4).

Figure 1 (facing page). Integrated proteomics, genomics and transcriptomics to improve sample-specific protein identification. **(A)** Schematic representation of the integrated genome and proteome analysis of BN-Lx and SHR rat liver **(B)** Bar plot showing the percentage of the current reference database that is covered by the experimentally derived proteomes, in respect to recent other proteomics efforts. For BN-Lx and SHR, 39.4% of the ENSEMBL database is covered (12,989 out of 32,971 entries; release 3.4.63). The human liver proteome generated by the Chinese Human Liver Proteome Profiling Consortium cover only 13.5% of the IPI human database (version 3.07; 7,050 out of 50,225 entries). **(C)** Diagram that displays identified proteins specific to BN-Lx (red), SHR specific proteins (blue) and proteins shared between both strains (green). **(D)** Relative contribution (%) of each additional layer of genomics- and transcriptomics-derived protein variants, based on identification by unique peptides only.

The *de novo* assembled transcriptome data also allows for characterization of transcriptome at nucleotide resolution. As the two rat strains used in these experiments are fully inbred, observed changes at the transcript level are unlikely to be allele-specific variation and can thus be attributed to RNA-Seq errors, mapping errors or, most interestingly, as a result of RNA editing (Farajollahi and Maas, 2010). We find a total of 799 canonical (A to I or C to T) RNA editing variants (Table S5) of which 176 and 354 were specifically observed in BN-*Lx* and SHR, respectively. As expected, a large proportion of edits reside in the non-coding UTR parts of transcripts or do not change coding capacity of a transcript, yet might be affecting RNA secondary structure, stability or miRNA binding. Only 196 edits were non-synonymous and therefore included in our protein database as potentially detectable by mass spectrometry. Of a subset of 169 candidates editing events tested by independent RT-PCR based amplicon resequencing of amplicons, most (104) showed reads corresponding to expected edited transcripts and another 12 likely represent germline variants that missed detection during genomes re-sequencing. (Table S3B).
All peptide variants and isoforms derived from genome and transcriptome variation and all novel predicted peptides based on GENSCAN and *de novo* transcriptome assembly data were appended to the ENSEMBL rat database to create our customized RAT_COMBINED database, that was used for all subsequent proteomic analyses.


A
Experimental overview
SHR
2 males, 2 females
BN-Lx
2 males, 2 females
liver tissue snapshot
DNA sequencing
RNA sequencing
Protein sequencing
Protein mapping on genes
B
100%
50%
0%
Sun et al
Low et al
Database Size
No. Proteins ID
C
3009
SHR
18322
5132
BN-Lx
D
Relative identification findings
3%
1%
3%
32%
61%
2%
1%
1%
32%
64%
BN-Lx
splice variants
edited mRNAs
SHR specific
BN-Lx specific
same between strains
Genscan

Ultra-deep proteomics analysis

To obtain direct experimental evidence for predicted peptide variants and isoforms, we generated ultra-deep MSMS data for each liver sample. We used the same tissue samples as used for generating the RNA-seq data. We proteolyzed equal amounts of each liver sample using five orthogonal proteases separately (trypsin, LysC, GluC, AspN and chymotrypsin) and analyzed 36 SCX fractions per digest, cumulating in 180 LC-MSMS runs for each strain. By using multiple proteases, not only does the identification and sequence coverage of each protein increase but also the chance of capturing evidence for new predicted peptides/proteins and consequences of RNA editing (Mohammed et al., 2008; Peng et al., 2012; Swaney et al., 2010). To further ensure comprehensive coverage, we performed a two-stage analysis with two different but complementary algorithms for spectra-to-peptide assignment. First, Mascot (Perkins et al., 1999) was used for database searching. Next, remaining unassigned peak lists were processed further with PEAKS (Ma et al., 2003), which incorporates a proprietary *de novo* sequencing algorithm. We obtained ~12 million tandem MS spectra. This enormous volume of data allowed us to apply a FDR filter of 0% (q=0), and still identify approximately 2 million peptide-spectral-matches (PSMs), corresponding to ~200,000 non-redundant peptides (Table S6). By benchmarking BN-*Lx* and SHR livers proteome against the human liver proteome reference dataset (Sun et al., 2010), the most extensive liver proteome so far generated by the Chinese Human Liver Proteome Profiling Consortium, we found that our dataset covers 39.6% of the rat database (12,989 out of 32,971 entries in ENSEMBL release 3.4.63) while the human liver proteome generated by the Chinese Human Liver Proteome Profiling Consortium covers 13.5% of the IPI human database (version 3.07; 50,225 entries), illustrating the comprehensiveness of our data (Figure 1B). Combining both BN-*Lx* and SHR datasets against genome- and transcriptome-informed protein database, we obtained peptide evidence for 27,038 database entries (RAT_COMBINED) of which 18,769 are shared between BN-*Lx* and SHR (Figure 1C, Table S7). As expected, identified peptides are evenly distributed over the rat chromosomes, concordant to the distribution of genes and transcripts Figure S1). The median coverage of all proteins is 15.6%. (Figure S2A), with roughly equal contributions from each protease dataset. The unprecedented high number of identified proteins using a 0% FDR confirms that this rat proteomics data set forms

a very high quality catalogue of rat liver protein orthologs, which might also serve as a valuable resource to complement the human liver proteome project (He, 2005) and the chromosome-centric human proteome project (Paik et al., 2012).

Identification of novel proteins and protein isoforms

Some 2000 peptides (Table S8A) generated the first experimental evidence for 792 *in silico* predicted GENSCAN proteins (Table S8B). For 786 (99%) of those, mRNA expression levels were sufficiently high to also find support in the RNA sequencing data. Fifty of them show best reciprocal hits with known mouse proteins, and another 32 with known human proteins (Table S8C). Furthermore, we detect N-terminally acetylated peptides for 69 of these 792 proteins, with, as expected, A, M, S and T as their N-terminal residues (Table S8D) (Dormeyer et al., 2007; Starheim et al., 2012). These N-terminal peptides further validate these putative genes by confirming their translational start sites. Since both the RNA-Seq and proteomics data were very deep and thoroughly analyzed, we hypothesize that most of the other predicted GENSCAN genes (44,201; 98.2%) are non-functional or expressed very lowly in liver.

Our proteomics data also provides support for 337 peptides exclusively matching 120 novel transcript splicing events (0% FDR) that were previously not annotated (Figure 1D, Table S9 and Table S10B). From all novel protein and splice isoform identifications, 119 and 48 respectively were unique for BN-*Lx* and 45 and 32 were specific to SHR.

Detection of short expressed proteins (<100 amino acids)

The identification of novel genes and their corresponding proteins is biased by the size of the protein. Generally, short proteins (<100 amino acids) are easily missed in proteomics experiments because they are underrepresented in the genome annotation due to an arbitrary cut-off of 100 amino acids (Carninci et al., 2005; Dinger et al., 2008; Maeda et al., 2006). Although computational analysis (Frith et al., 2006) and ribosome profiling (Ingolia et al., 2011) indicate the existence of thousands of short-ORFs (sORFs) that encode short expressed proteins (SEPs) (Slavoff et al., 2013), relatively few are confirmed by proteomics (Kastenmayer et al., 2006; Oyama et al.,

2007; Slavoff et al., 2013; Yang et al., 2011). SEPs have been implicated in plant and animal development (Galindo et al., 2007; Hashimoto et al., 2001; Oelkers et al., 2008), stressing the need for their accurate detection. Within ENSEMBL release 3.4.63, peptides derived from sORFs constitute only 5.13% (1,683) of all entries. However, the relative contribution of sORFs to the *in silico* predicted GENSCANs is much higher (6,769, 14.13%). When we plot the distribution of the proteins detected in our experiments by size, the majority consists of 200-2000 amino acids, while short proteins (< 100 amino acids) constitute only 0.25% (Figure S3). We obtained experimental evidence for 124 annotated short proteins and confirmed 17 of the many predicted novel SEPs in both rat liver samples. All SEPs were supported by sORFs in the RNA-Seq data (Table S8B). These results show that mass spectrometry combined with transcriptomics provides an effective way to detect SEPs, but our data also suggest that caution is required when concluding that putative sORFs are actually translated. The current sORF annotation may severely overestimate the number of SEPs.

Detection of non-synonymous protein variants

Next, we explored to what extent the addition of strain-specific variants affected protein detectability. 4.3% of uniquely assigned spectra did discriminate between allele-specific protein isoforms (Tables S7 and S9). By applying a 0% FDR cut-off, we reassuringly did not identify any BN-*Lx* variants in the SHR samples, and *vice versa*. The fact that only a portion of non-synonymous variants was confirmed by peptide-based evidence can be explained by our experimental design in which only genes expressed in the liver could be detected. Clearly, the inclusion of allele-specific variants has a measurable impact on protein discovery and results in more balanced peptide count per strain. The latter is most notable for the SHR rat because its genome is more diverged from the reference BN strain (used for construction of the original database).

Peptide-based evidence for RNA-editing

To identify functional RNA-editing events, we mapped our peptide spectra to the set of potential RNA-editing events. In total, 54 out of the 196 non-synonymous editing events could be confirmed by unique pep-

tide-based evidence. (Table S9 and S10). Since unique peptide evidence needs to overlap with the predicted editing site, many of the remaining 142 edits are likely missed because of incomplete coverage or redundancy in peptide data. Whereas limitations in the MS technology obviously result in an under-representation of identified RNA edits, MS still provides the best means to confirm the presence of such post-transcriptional modifications in the expressed proteins. On the other hand, we cannot rule out a possibility that the relatively low percentage of confirmed events is a true representation of the actual level of post-transcriptional modifications that make it to mature proteins. This may be due to negative selection of modified mRNA molecules. The high level of RNA sequencing coverage and the strict calling settings used to define editing events make it unlikely that an overestimation of editing events is introduced during the RNA sequencing procedure and analysis.

It is worth noting that our comparison of *de novo* assembled and the annotated transcriptome may not only reveal genetic differences, transcript isoforms and common edited sites. Sequence and annotation imperfections within the current assembly and gene build can also be detected since the proteogenomics approach used in this study accounts for differences between observed and annotated transcriptome that originate from both biological and technical sources. Also, we emphasize that the *de novo* transcriptome assembly approach should be supplemented by regular transcriptome profiling if one aims to discover transcript variants that correspond to low abundance transcripts and low-frequency events. To this end we performed direct alignment of RNA-Seq data to the rat transcriptome (known proteins and GENSCAN predictions) and predicted additional modifications of annotated transcripts (Table S5).

Predicting the effects of germline variants on protein stability

To assess whether mutations affect protein stability we evaluated the effects of strain- and transcriptome-specific non-synonymous changes as predicted by SIFT (Ng and Henikoff, 2001) and Polyphen2 (Adzhubei et al., 2010) algorithms. We asked whether genes that harbor non-synonymous variants that are predicted to affect protein function could more frequently be found among differentially expressed genes, relative to those where changes in coding capacity are considered to be benign. Neither of

the algorithms revealed any significant difference at transcriptome level (Table S11). However, overrepresentation of potentially damaging mutations in differentially expressed proteins was clear from Polyphen2 results (chi-square P<0.002) where 51% of differentially expressed genes harbor deleterious mutations, while same is true for only 37% for genes where expression is not significantly altered between strains. A similar, but not significant trend observed for SIFT predictions (p=0.065). Our analysis shows that non-conservative and structural missense variants, may have limited influence on abundance of transcript, yet can show a pronounced effect on protein stability and, thus, expression level.

Relation between transcriptome and proteome levels

Next, we studied quantitative aspects by investigating the abundance of mRNA and protein levels. Although being derived from two different strains of rats, we observed a very high correlation of liver mRNA between BN-*Lx* and SHR (r = 0.98). Similarly, the correlation coefficient for protein expression between BN-*Lx* and SHR is also remarkably high r = 0.94 (Figure 2A).

Next we sought to define a correlation or relationship between mRNA and protein expression levels in our data. Making a direct correlation between mRNA and protein levels is hampered by the fact that in peptide-based proteomics many proteins contain similar peptide sequences. It is therefore hard to assign any of the shared peptides unambiguously to a protein, the so-called protein-inference problem (Grobei et al., 2009; Nesvizhskii and Aebersold, 2005; Serang et al., 2012). Consequently, it is hard to integrate the quantitative measurements, which are necessarily restricted to peptides, to a protein measurement. Still from numerous studies it has been concluded that the global correlation between mRNA and protein is certainly not linear and often an r of 0.4-0.5 is reported (Ning et al., 2012; de Sousa Abreu et al., 2009; Vogel and Marcotte, 2012). Such findings are corroborated by results that show that indeed only part of the variation in the protein levels can be explained by mRNA levels (Schwanhäusser et al., 2011). Here we use for quantification of protein levels a spectra-count based mass spectrometry method, and use data derived from using five different proteolytic enzymes, sufficient to exclude a proteolytic digest-specific bias (Peng et al., 2012). Although we did identify peptides that are not shared

between proteins (Table S12), we chose to take the total number of PSMs for every peptide matching a protein as a measurement of its abundance to increase the quantitative resolution per protein.
Subsequently, we determined the proteome-transcriptome correlation for BN-*Lx* and SHR, to be, r = 0.43 and 0.44, respectively, which is thus weak, albeit in line with the previous studies in other systems (Figure 2). Since various sequence features of transcripts have been documented to impact protein abundance (Vogel et al., 2010), we proceeded to determine which part of the transcript may be the most informative for estimating protein levels. To do so, we selected all coding transcripts with annotated 5'-UTRs and 3'-UTRs, exhibiting positive expression values as well as non-zero protein abundance (n=1,576). We observe that transcript abundance that is calculated from the 3'-UTR sequences alone improves the observed mRNA protein correlation substantially, when compared to measures ob-

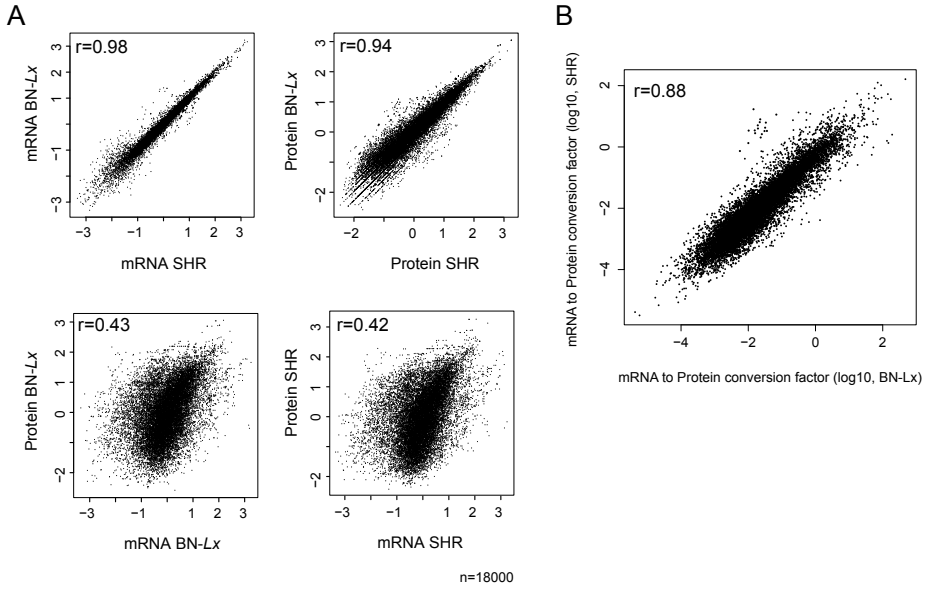


Figure 2. Global correlation plots displaying the complexity of mRNA and protein abundance. **(A)** Correlations between BN-Lx and SHR (top panels) for both mRNA and protein levels are shown as calculated using log10 normalized spectral counts (Log10SAF) and normalized RNA seq counts (Log10RPKM). The bottom two panels show the correlations between mRNA and protein abundance for BN-Lx (r=0.43) and SHR (r=0.42) respectively. **(B)** Scatter plot depicting the correlation between experimentally determined gene specific conversion factors (r=0.88) as calculated for BN-Lx and SHR.

tained from the 5'-UTR or the actual coding part of a transcript (Figure S5). This suggests that reads that correspond to the 3'-UTR part of the gene, thus reflecting both the expression level and degree of mRNA degradation may be a better proxy to approximate protein abundance.
Although globally the mRNA and protein data do show a rather weak direct correlation, our massive data allowed us to calculate an "mRNA to protein conversion factor" for each individual protein. These conversion factors span six orders of magnitude (Figure 2B). Interestingly, these factors correlate well between the two strains (r = 0.88). This conversion factor implies a convenient number with which possible biological parameters may be correlated. Also, it may serve as a more precise gene-specific proxy to determine the actual protein level that corresponds to the determined mRNA abundance. Evidently, we do not know yet whether these mRNA to protein conversion factors may be time and space dependent, and thus different in different organs or species.

Genetic control of quantitative proteome characteristics

To determine the effects of genetic variation on quantitative transcriptome and proteome characteristics, we compared the difference of mRNA and protein expression between the two rat strains. Only quantifiable genes (with evidence of their expression at both the protein and transcript level) were retained, which made it possible to compare 6,743 values (Figure 3 and Table S13). One hundred and thirteen of them showed differential expression when 4 BN-*Lx* transcriptomes and 4 SHR transcriptomes were compared (with expression change of 2 or above). The largest proportion of genes with differentially expressed transcripts (59) does not show comparable changes at proteomics level. These proteins potentially acquire stable expression through regulation of at the level of translation or through proteostasis. A small proportion of genes (13) showed discordant behavior with opposite expression patterns for transcripts and proteins. Both groups do not show any over-representation in gene ontology or pathways. These data clearly show the high global genome and proteome similarity between the two inbred rat strains. However, they also illustrate that inter-individual differences may be in the details, such as represented by changes in post-translational protein modifications and protein networks (Altelaar et al., 2013; Bensimon et al., 2012). Finally, 41 out of the 113 dif-

ferential genes show strain-specific expression changes that are consistent between transcriptome and proteome (Figure 3; Table S13). The products of these 41 genes relate to catalytic activity (28 genes, GO-Term enrichment p-value 1.4e-5) and metabolic pathways (13 genes, p=2.6e-4).

A germline promoter variant deregulates Cyp17a1 expression in spontanuously hypertensive rats

This set of 41 genes likely underlies some of the phenotypic differences known to exist between BN-*Lx* and SHR rats, like spontaneous hypertension (Okamoto and Aoki, 1963) and metabolic syndrome (Aitman et al., 1997, 1999). We therefore next investigated which genes were previously reported to be associated with hypertension in human or rat. First, 3 out of the 41 genes that are differential at both the mRNA and protein level were found to be associated with hypertension in the rat. Those 3 genes, *Hao2* (Lee et al., 2003), *Serpina3m* and *Cyp8b1* (Kinoshita et al., 2011), came out as top-hits while studying SHR (-related) strains or a panel of congenic rat strains to define candidates for hypertension. All 3 genes also overlap known blood pressure QTLs in the rat (RGD). A fourth gene, *Cyp17a1*, was identified by human genome-wide association studies as a top hit in relation to blood pressure and hypertension in European, Japanese and Chinese individuals (Li et al., 2013; Liu et al., 2011; Newton-Cheh et al., 2009; Takeuchi et al., 2010) (Table S14). This gene also overlaps a blood pressure QTL in rat and shows the most extreme down-regulation in SHR compared to BN-*Lx* in our analysis (Figure 3). Like *Cyp8b1, Cyp17a1* is a member of the cytochrome P450 (CYP450) superfamily (Danielson, 2002) of catalytic enzymes that mediate monooxygenase reactions and regulate drug metabolism. Interestingly, mutations in human *CYP17A1* are known to lead to congenital adrenal hyperplasia due to 17 alpha-hydroxylase deficiency, which results in hypogonadism, pseudohermaphroditism and severe hypertension (Biglieri, 1997; Biglieri et al., 1966; Geller et al., 1997; Goldsmith et al., 1967). To determine the genetic basis of the *Cyp17a1* expression differences between BN-*Lx* and SHR, we sought for germline variants in the annotated exons and flanking regulatory sequences, but none were present. Exploration of eQTL data based on the BXH/HXB recombinant inbred panel (which is derived from the BN-Lx and SHR strains) (Heinig et al., 2010), however, revealed a cis-effect, indicating that

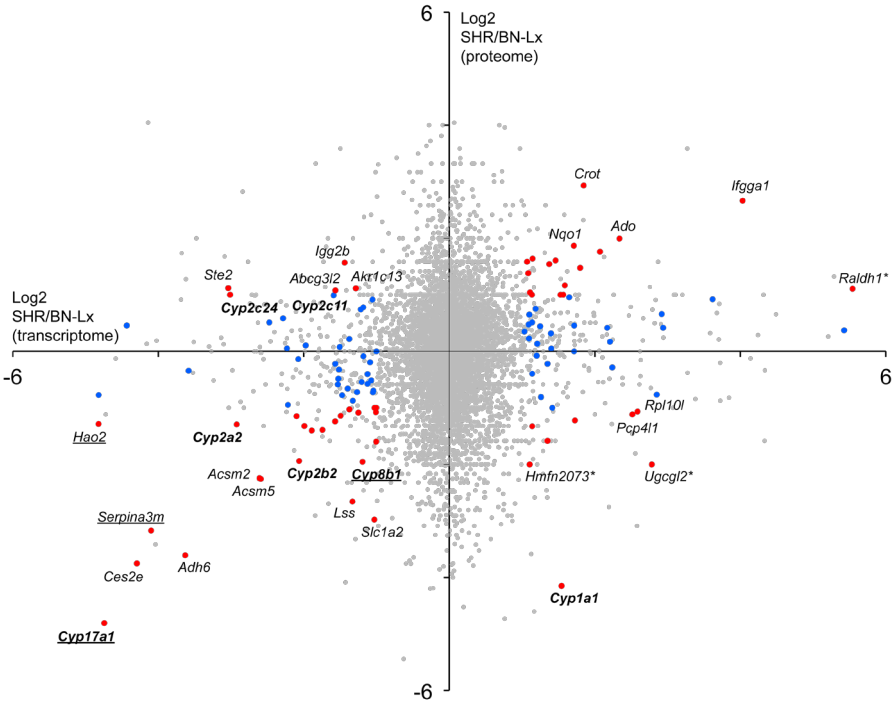


Figure 3. Gene-centric strain-to-strain comparison of significantly differentially expressed genes.Genes in BN-Lx and SHR with significantly deviating mRNA levels (blue dots; n=59) or mRNA and protein levels (red dots; n=54) are highlighted in this scatter plot. Gene names marked by an asterisk are given based on GENSCAN blast predictions derived from the closest predicted homology to human and mouse genes. Genes belonging to the CYP450 superfamily of catalytic enzymes are in bold and genes associated with hypertension in human or rat literature (Hao2, Serpina3m, Cyp8b1 and Cyp17a1) are underscored.

the measured expression difference is likely due to genetic variants in the gene itself or in neighboring regulatory elements. Upon closer inspection of the RNA sequencing data we found that the transcriptional start site (TSS) of the *Cyp17a1* gene was incorrectly annotated and resides approximately 2kb upstream of the currently annotated most 5' exon (Figure 4A). Interestingly, this promoter does harbor a germline variant in SHR that disrupts the core part of an evolutionary conserved forkhead-box DNA binding domain (Figure 4B, C) (Sandelin et al., 2004), specifically deregulating transcription in SHR (Figure 4A). Since this expression trait is regu-

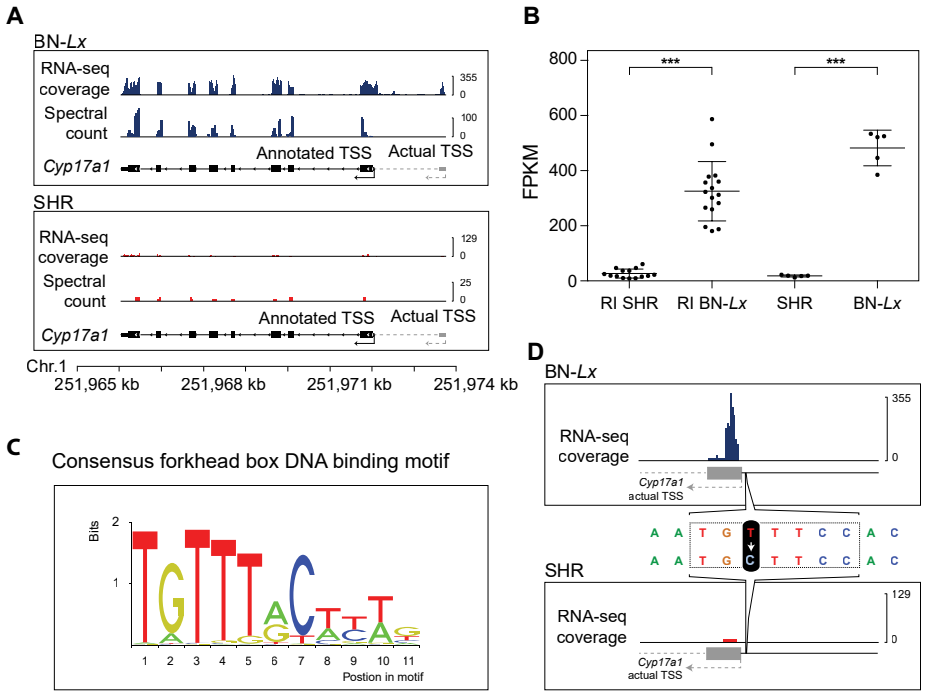


Figure 4. A germline promoter variant deregulates Cyp17a1 expression in spontanuously hypertensive rats. **(A)** RNA sequencing data and the proteomics spectral count are plotted along the gene body of Cyp17a1 for BN-Lx (blue) and SHR (red). The transcript is positioned on the reverse strand. Both the annotated TSS (black arrow) and the actual TSS (grey arrow) are shown. **(B)** Zoomed-in view of the actual TSS, with the position of the germline T/C SNV shown. The dashed box (grey) shows the core part of the forkhead box DNA binding motif. **(C)** Consensus forkhead box DNA binding motif, obtained from the JASPAR database FOXA1 motif (Sandelin et al., 2004).

lated in *cis* and this SNV is the only germline variant in the vicinity of the gene, our integrated genomics, transcriptomics and proteomics approach has most likely identified the source of expression variation. The overlap with the RGD blood pressure QTL, top GWAS loci in humans and known link to hypertension as a result of renal hyperplasia in patients carrying *CYP17A1* mutations are good indications that this promoter mutation in the SHR *Cyp17a1* gene contributes to the observed hypertensive phenotype of SHR rats.

Conclusions

The synergistic use of genomic, transcriptomic and proteomic technologies, presented here, demonstrates that one can achieve significant gains in informative data obtainable by proteomics and, complementarily, one can also refine and confirm predicted DNA or RNA variants. We show that when ultra-deep MS-based proteomics data is matched to a custom, personalized database built from the genome and transcriptome of the actual sample being studied, thousands of individual-specific peptides can be detected that would otherwise escape identification. Using the liver proteome analysis as an example, we show that, although proteomics has been making steady progress in the last 10 years (Table S15), integrating with NGS takes proteome analysis to a higher level. Ideally, and likely the future approach, the protein database used for MS-based proteomics would be generated by genome and transcriptome sequencing from the tissue or cell line under study. Technological advances in both the proteomics and the sequencing community now provide the ability to discriminate genetic and post-transcriptional polymorphisms at the proteome level, as well as to improve quantitation of gene expression because, as we and others have shown, transcriptome data alone is imprecise to predict protein level changes and study disease phenotypes. This implies that future efforts on both platforms benefit largely from the here as proof-of-concept presented combined approach.

Acknowledgments

This work was supported by the Netherlands Proteomics Centre, which is part of the Netherlands Genomics Initiative and a TOP grant from NWO-CW (N° 700.58.303) to EC. This work was financially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° HEALTH-F4-2010-241504 (EURATRANS) to EC and the PRIME-XS project grant agreement number 262067 to AJRH.

Author contributions

T.Y.L designed, performed and analyzed the proteomics experiments;

S.v.H. designed, performed and analyzed the RNA-seq experiments. H.vd.T. performed qualitative, quantitative and bioinformatics analysis on both transcriptomics and proteomics data. P.G. and A.C. performed mass spectrometry and data analysis. B.v.B. and S.M. provide consultation and support for bioinformatics and mass spectrometry. S.v.H, P.T. and V.G. performed and analyzed RNA-sequencing validation experiments. S.M., V.G., E.C. and A.J.R.H. contributed to conceptual design and scientific discussions. V.G. performed bioinformatics analysis on genomics, transcriptomics and proteomics data and is responsible for generating the new protein database. T.Y.L., S.v.H., H.vd.T., B.V.B., S.M., A.J.R.H., E.C. and V.G. wrote the manuscript.

Data Availability

MS data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Vizcaíno et al., 2013) with the dataset identifier PXD000131. DNA data were previously deposited in Sequence Read Archive (SRA) and are accessible via the following identifiers: BN-Lx genome: ERP001355; SHR genome: ERP001371; BN reference genome: ERP000510. RNA sequencing data were stored in ArrayExpress: E-MTAB-1666.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Ahrens, C.H., Brunner, E., Qeli, E., Basler, K., and Aebersold, R. (2010). Generating and navigating proteome maps using mass spectrometry. Nat. Rev. Mol. Cell Biol. *11*, 789–801.

Aitman, T.J., Gotoda, T., Evans, A.L., Imrie, H., Heath, K.E., Trembling, P.M., Truman, H., Wallace, C.A., Rahman, A., Doré, C., et al. (1997). Quantitative trait loci for cellular defects in glucose and fatty acid metabolism in hypertensive rats. Nat. Genet. *16*, 197–201.

Aitman, T.J., Glazier, A.M., Wallace, C.A., Cooper, L.D., Norsworthy, P.J., Wahid, F.N., Al-Majali, K.M., Trembling, P.M., Mann, C.J., Shoulders, C.C., et al. (1999). Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. Nat. Genet. *21*, 76–83.

Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. Nat. Rev. Genet. *14*, 35–48.

Atanur, S.S., Birol, I., Guryev, V., Hirst, M., Hummel, O., Morrissey, C., Behmoaras, J., Fernandez-Suarez, X.M., Johnson, M.D., McLaren, W.M., et al. (2010). The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. *20*, 791–803.

Bensimon, A., Heck, A.J.R., and Aebersold, R. (2012). Mass Spectrometry-Based Proteomics and Network Biology. Annu. Rev. Biochem. Vol 81 *81*, 379–405.

Biglieri, E.G. (1997). 17 alpha-Hydroxylase deficiency: 1963-1966. J. Clin. Endocrinol. Metab. *82*, 48–50.

Biglieri, E.G., Herron, M.A., and Brust, N. (1966). 17-hydroxylation deficiency in man. J. Clin. Invest. *45*, 1946–1954.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Cox, J., and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. Annu. Rev. Biochem. *80*, 273–299.

Danielson, P.B. (2002). The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr. Drug Metab. *3*, 561–597.

Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput. Biol. *4*, e1000176.

Dormeyer, W., Mohammed, S., Breukelen, B. van, Krijgsveld, J., and Heck, A.J.R. (2007). Targeted analysis of protein termini. J. Proteome Res. *6*, 4634–4645.

Farajollahi, S., and Maas, S. (2010). Molecular diversity through RNA editing: a balancing act. Trends Genet. *26*, 221–230.

Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T.L., and Grimmond, S.M. (2006). The abundance of short proteins in the mammalian proteome. PLoS Genet. *2*, e52.

Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A., and Couso, J.P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. *5*, e106.

Geller, D.H., Auchus, R.J., Mendonça, B.B., and Miller, W.L. (1997). The genetic and functional basis of isolated 17,20-lyase deficiency. Nat. Genet. *17*, 201–205.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493–521.

Goldsmith, O., Solomon, D.H., and Horton, R. (1967). Hypogonadism and mineralocorticoid excess. The 17-hydroxylase deficiency syndrome. N. Engl. J. Med. *277*, 673–677.

Grobei, M.A., Qeli, E., Brunner, E., Rehrauer, H., Zhang, R., Roschitzki, B., Basler, K., Ahrens, C.H., and Grossniklaus, U. (2009). Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Res. *19*, 1786–1800.

Hashimoto, Y., Niikura, T., Tajima, H., Yasukawa, T., Sudo, H., Ito, Y., Kita, Y., Kawasumi, M., Kouyama, K., Doyu, M., et al. (2001). A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. Proc. Natl. Acad. Sci. U. S. A. *98*, 6336–6341.

He, F. (2005). Human liver proteome project: plan, progress, and perspectives. Mol. Cell. Proteomics *4*, 1841–1848.

Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S.R., Bauerfeind, A., et al. (2010). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature *467*, 460–464.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet. *37*, 243–253.

Hurst, J.M., McMillan, L.E.M., Porter, C.T., Allen, J., Fakorede, A., and Martin, A.C.R. (2009). The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum. Mutat. *30*, 616–624.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Jaffe, J.D., Berg, H.C., and Church, G.M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics *4*, 59.

Jansen, R.C., and Nap, J.P. (2001). Genetical genomics: the added value from segregation. Trends Genet. *17*, 388–391.

Jensen, O.N. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. *8*, 33–41.

Johnson, M.D., He, L., Herman, D., Wakimoto, H., Wallace, C.A., Zidek, V., Mlejnek, P., Musilova, A., Simakova, M., Vorlicek, J., et al. (2009). Dissection of chromosome 18 blood pressure and salt-sensitivity quantitative trait loci in the spontaneously hypertensive rat. Hypertension *54*, 639–645.

Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.-C., Yang, H., Carter, C.D., Wheeler, D., Davis, R.W., Boeke, J.D., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res. *16*, 365–373.

Kinoshita, K., Ashenagar, M.S., Tabuchi, M., and Higashino, H. (2011). Whole rat DNA array survey for candidate genes related to hypertension in kidneys from three spontaneously hypertensive rat substrains at two stages of age and with hypotensive induction caused by hydralazine hydrochloride. Exp. Ther. Med. *2*, 201–212.

Kleinman, C.L., and Majewski, J. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302; author reply 1302.

Lee, S.J., Liu, J., Qi, N., Guarnera, R.A., Lee, S.Y., and Cicila, G.T. (2003). Use of a panel of congenic strains to evaluate differentially expressed genes as candidate genes for blood pressure quantitative trait loci. Hypertens. Res. *26*, 75–87.

Li, X., Ling, Y., Lu, D., Lu, Z., Liu, Y., Chen, H., and Gao, X. (2013). Common polymorphism rs11191548 near the CYP17A1 gene is associated with hypertension and systolic blood pressure in the Han Chinese population. Am. J. Hypertens. *26*, 465–472.

Lin, W., Piskol, R., Tan, M.H., and Li, J.B. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302; author reply 1302.

Liu, C., Li, H., Qi, Q., Lu, L., Gan, W., Loos, R.J., and Lin, X. (2011). Common variants in or near FGF5, CYP17A1 and MTHFR genes are associated with blood pressure and hypertension in Chinese Hans. J. Hypertens. *29*, 70–75.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *17*, 2337–2342.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P.G., Lenhard, B., Aturaliya, R.N., Batalov, S., Beisel, K.W., et al. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet. *2*, e62.

Merrihew, G.E., Davis, C., Ewing, B., Williams, G., Käll, L., Frewen, B.E., Noble, W.S., Green, P., Thomas, J.H., and MacCoss, M.J. (2008). Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res. *18*, 1660.

Mohammed, S., Lorenzen, K., Kerkhoven, R., van Breukelen, B., Vannini, A., Cramer, P., and Heck, A.J.R. (2008). Multiplexed proteomics mapping of yeast RNA polymerase II and III allows near-complete sequence coverage and reveals several novel phosphorylation sites. Anal. Chem. *80*, 3584–3592.

Munoz, J., Low, T.Y., Kok, Y.J., Chin, A., Frese, C.K., Ding, V., Choo, A., and Heck, A.J.R. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol. Syst. Biol. *7*, 550.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics *4*, 1419–1440.

Nesvizhskii, A.I., Roos, F.F., Grossmann, J., Vogelzang, M., Eddes, J.S., Gruissem, W., Baginsky, S., and Aebersold, R. (2006). 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 MCP *5*, 652–670.

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nat. Genet. *41*, 666–676.

Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res. *11*, 863–874.

Ning, K., Fermin, D., and Nesvizhskii, A.I. (2012). Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. J. Proteome Res. *11*, 2261–2271.

Oelkers, K., Goffard, N., Weiller, G.F., Gresshoff, P.M., Mathesius, U., and Frickey, T. (2008). Bioinformatic analysis of the CLE signaling peptide family. BMC Plant Biol. *8*, 1.

Okamoto, K., and Aoki, K. (1963). Development of a strain of spontaneously hypertensive rats. Jpn. Circ. J. *27*, 282–293.

Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T., and Sugano, S. (2007). Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol. Cell. Proteomics *6*, 1000–1006.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Paik, Y.-K., Jeong, S.-K., Omenn, G.S., Uhlen, M., Hanash, S., Cho, S.Y., Lee, H.-J., Na, K., Choi, E.-Y., Yan, F., et al. (2012). The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat. Biotechnol. *30*, 221–223.

Payne, S.H., Huang, S.-T., and Pieper, R. (2010). A proteogenomic update to Yersinia: enhancing genome annotation. BMC Genomics *11*.

Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A.J.R. (2012). Protease bias in absolute protein quantitation. Nat. Methods *9*, 524–525.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551–3567.

Pickrell, J.K., Gilad, Y., and Pritchard, J.K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302; author reply 1302.

Pravenec, M., and Kurtz, T.W. (2010). Recent advances in genetics of the spontaneously hypertensive rat. Curr. Hypertens. Rep. *12*, 5–9.

Pravenec, M., Zídek, V., Landa, V., Simáková, M., Mlejnek, P., Kazdová, L., Bílá, V., Krenová, D., and Kren, V. (2004). Genetic analysis of "metabolic syndrome" in the spontaneously hypertensive rat. Physiol. Res. *53 Suppl 1*, S15-22.

Printz, M.P., Jirout, M., Jaworski, R., Alemayehu, A., and Kren, V. (2003). Genetic Models in Applied Physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. J. Appl. Physiol. *94*, 2510–2522.

Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. Proteomics *11*, 620–630.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32*, D91-4.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Serang, O., Moruz, L., Hoopmann, M.R., and Käll, L. (2012). Recognizing Uncertainty Increases Robustness and Reproducibility of Mass Spectrometry-based Protein Inferences. J. Proteome Res.

Simonis, M., Atanur, S.S., Linsen, S., Guryev, V., Ruzius, F.-P., Game, L., Lansu, N., de Bruijn, E., van Heesch, S., Jones, S.J.M., et al. (2012a). Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biol. *13*, r31.

Simonis, M., Atanur, S.S., Linsen, S., Guryev, V., Ruzius, F.-P., Game, L., Lansu, N., de Bruijn, E., van Heesch, S., Jones, S.J., et al. (2012b). Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biol. *13*.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64.

Soon, W.W., Hariharan, M., and Snyder, M.P. (2013). High-throughput sequencing for biology and medicine. Mol. Syst. Biol. *9*, 640.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Mol. Biosyst. *5*, 1512–1526.

Starheim, K.K., Gevaert, K., and Arnesen, T. (2012). Protein N-terminal acetyltransferases: when the start matters. Trends Biochem. Sci. *37*, 152–161.

Su, Z.-D., Sun, L., Yu, D.-X., Li, R.-X., Li, H.-X., Yu, Z.-J., Sheng, Q.-H., Lin, X., Zeng, R., and Wu, J.-R. (2011). Quantitative detection of single amino acid polymorphisms by targeted proteomics. J. Mol. Cell Biol. *3*, 309–315.

Sun, A., Jiang, Y., Wang, X., Liu, Q., Zhong, F., He, Q., Guan, W., Li, H., Sun, Y., Shi, L., et al. (2010). Liverbase: a comprehensive view of human liver biology. J. Proteome Res. *9*, 50–58.

Swaney, D.L., Wenger, C.D., and Coon, J.J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J. Proteome Res. *9*, 1323–1329.

Takeuchi, F., Isono, M., Katsuya, T., Yamamoto, K., Yokota, M., Sugiyama, T., Nabika, T., Fujioka, A., Ohnaka, K., Asano, H., et al. (2010). Blood pressure and hypertension are associated with 7 loci in the Japanese population. Circulation *121*, 2302–2309.

Uhlen, M., and Ponten, F. (2005). Antibody-based proteomics for human tissue profiling. Mol. Cell. Proteomics *4*, 384–393.

Valentine, S.J., Sevugarajan, S., Kurulugama, R.T., Koeniger, S.L., Merenbloom, S.I., Bohrer, B.C., and Clemmer, D.E. (2006). Split-field drift tube/mass spectrometry and isotopic labeling techniques for determination of single amino acid polymorphisms. J. Proteome Res. *5*, 1879–1887.

Venter, E., Smith, R.D., and Payne, S.H. (2011). Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One *6*, e27587.

Vizcaíno, J.A., Côté, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J., et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. *41*, D1063-9.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. *6*, 400.

Volkening, J.D., Bailey, D.J., Rose, C.M., Grimsrud, P.A., Howes-Podoll, M., Venkateshwaran, M., Westphall, M.S., Ané, J.-M., Coon, J.J., and Sussman, M.R. (2012). A proteogenomic survey of the Medicago truncatula genome. Mol. Cell. Proteomics *11*, 933–944.

Yang, X., Tschaplinski, T.J., Hurst, G.B., Jawdy, S., Abraham, P.E., Lankford, P.K., Adams, R.M., Shah, M.B., Hettich, R.L., Lindquist, E., et al. (2011). Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res. *21*, 634–641.

Yates, J.R., Eng, J.K., McCormack, A.L., and Schieltz, D. (1995). Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database. Anal. Chem. *67*, 1426–1436.

Supplementary Materials and Methods

Identification of nonsynonymous genomic variants for BN-Lx and SHR

Single nucleotide variants and indels were obtained from previous genome sequencing efforts(Atanur et al., 2010; Simonis et al., 2012). Non-synonymous (amino acid changing) mutations were obtained using the Variant Effect Predictor tool (V2.1)(McLaren et al., 2010) For each genomic variant that can result in a polymorphic protein, strain-specific alleles were reconstructed and added to extended peptide reference database.

Transcriptome sequencing and assembly

Total RNA was isolated from snap-frozen and powdered liver tissue samples of 6-week old inbred BN-*Lx*/Cub and SHR/OlaIpcv males and females (2 per gender, per strain). Total RNA was purified prior to RNA sequencing library preparation using the RiboMinus™ Eukaryote Kit for RNA-Seq (Invitrogen). Libraries were prepared exactly according to manufacturer's instructions (SOLiD™ V4 guide for library preparation, Life Technologies) using the SOLiD™ total RNA-seq kit. Eight libraries were sequenced simultaneously using multiplexed paired-end technology (50 + 35 bp) on a single slide of SOLiD™ V4 system. For *de novo* transcriptome assembly, we used CLCBio assembly cell version 4 (CLC Bio, Aarhus, Denmark) to assemble transcriptomes of rat liver samples. All *de novo* assembly and scaffolding procedures were done for each of the eight samples separately. In addition we constructed transcriptomes by merging datasets coming from the same genetic background (BN-*Lx* or SHR). Transcriptomes were mapped against the reference genome assembly using BLAT software(Kent, 2002).

Detection of RNA editing and splicing in RNA sequencing data

To define accurately where RNA editing took place, we selected only the best read-alignments, with the second best hit having an at least 10% lower BLAT score. Also, gapped alignments produced in the previous step were scored for base-level inconsistencies between the genome and tran-

scriptome sequences. We required that each RNA editing candidate was flanked by at least 50 base pairs of uninterrupted sequence that match the genome in both the 5' and 3' direction. Transcriptome contigs that showed a discrepancy with the genome sequence (taking previously annotated strain-specific genomic variants into account) were compared with their corresponding protein sequences using NCBI BLAST (blastx). For all predicted nonsynonymous RNA editing variants, an individual entry was created in the extended rat protein database. For each entry we included which of the 8 rat liver samples showed this specific variant.
Similar to the detection of editing events, splice events were detected using alignments between the assembled transcriptome and genome, and compared to their corresponding proteins. Best, but structurally imperfect homologies between the annotated protein and translated transcript (due to insertions / deletions) comprised the source for alternative transcripts. Together with the genomic variants and RNA editing events, these alternative splice events were included into the extended protein search database.

GENSCAN gene predictions and support by RNA sequencing data

When best matches of transcriptome contigs corresponded to a GENSCAN prediction rather than to an annotated gene, we included it as possible translated sequence. About 10% of all GENSCAN predictions were supported by RNA-Seq data (~3,000), but as a control for the detection limit of the RNA sequencing data, all 47,450 GENSCAN predictions were included in our protein search database.

Sequence database compilation

We downloaded the annotated ENSEMBL(Birney et al., 2004; Curwen et al., 2004; Hubbard, 2002) rat protein FASTA (build 3.4.63) derived from the genome assembly of the Brown Norway (BN) strain as our foundation. Subsequently, to tailor-make an in-house rat protein database with enhanced comprehensiveness and precision, we modified and appended the original database with information derived from DNA re-sequencing and RNA-sequencing (RNA-seq) of the BN-Lx and SHR strain used in this study.

Quantification of transcriptome data

Relative expression for each gene was calculated from the alignment of RNA-Seq reads and genome annotation. RNA abundance for each gene was calculated as reads per kilobase per million sequences reads (RPKM) to normalize for transcript length.

Liver tissue sample preparation for proteomics

Snap-frozen liver tissues were re-suspended in 8 M urea in 50 mM ammonium bicarbonate (pH 8.0), supplemented with protease inhibitors (Complete protease inhibitor cocktail tablets, Roche Diagnostics). Lysates were then sonicated and cleared by centrifugation at 13,000×g. Protein lysates (300μg) were reduced with 1 mM dithiothreitol and alkylated with 5.5 mM iodoacetamide. For tryptic digestion, proteins were digested with endoproteinase Lys-C (Wako Chemicals) and sequencing grade modified trypsin (Promega) after 4-fold dilution in water. Protease digestion was stopped by addition of trifluoroacetic acid and precipitates were removed after centrifugation. Peptides were desalted using reversed-phase Sep-Pak C18 cartridges (Waters). The same digestion conditions were applied for GluC, AspN and Chymotrypsin.

Strong Cation Exchange Chromatography (SCX)

Desalted peptide samples were fractionated using strong cation exchange (SCX) system consisting of an Agilent 1100 HPLC system (Agilent Technologies) coupled to a Zorbax BioSCX-Series II column (0.8-mm inner diameter × 50-mm length, 3.5 μm). Solvent A consisted of 0.05% formic acid in 20% acetonitrile while solvent B was 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. The SCX salt gradient is as follows: 0-0.01 min (0-2% B); 0.01-8.01 min (2-3% B); 8.01-14.01 min (3-8% B); 14.01-28 min (8-20% B); 28-38 min (20-40% B); 38-48 min (40-90% B); 48-54 min (90% B); 54-60 min (0% B). A total of 50 SCX fractions (1 min each, i.e. 50-μl elution volume) were collected and dried in a vacuum centrifuge.

MS Analysis

The early fractions from SCX were analyzed with an Agilent 1290 Infinity

(Agilent Technologies, Waldbronn, DE) system coupled to a TripleTOF 5600 system (AB Sciex, Concord, ON). The UPLC was equipped with a double frit trapping column (ReproSil-Pur C18-AQ, 3µm, Dr. Maisch GmbH, Ammerbuch, Germany: 2 cm x 100 mID, packed in-house) and an analytical column (Agilent Zorbax SB-C18, 1.8 µm, 40 cm x 50 µm, packed in-house) for online trapping, desalting, and analytical separations. The solvents used were: buffer A 0.1% formic acid in water and buffer B 0.1% formic acid in 80% acetonitrile. Trapping and desalting were carried out at 5 µL/min for 10 min with 100% buffer A. For elution, the flow rate was passively split to 100 nL/min and the analytical separation was established in 2h gradient by the following conditions: immediately after loading the percentage of buffer B was increased to 15% in 0.1min, then buffer B was increased up to 21% and 35% respectively in 47.4 min and 47.5 min. Following the gradient was increased to 100% B for 2 min and maintained for 1 min. Initial chromatographic conditions were restored in 1 min and maintained for 10 min. Data acquisition was performed with a TripleTOF5600 System fitted with a Nanospray III source (AB SCIEX, Concord, ON) and a coated tip as the emitter (New Objectives, Woburn, MA). Data was acquired using an ion spray voltage of 2.7 kV, curtain gas of 10 PSI, nebulizer gas of 10 PSI, and an interface heater temperature of 100°C. The mass spectrometer was operated in information-dependent acquisition mode (IDA) and MS spectra were acquired across the mass range of 350–1250 m/z in high-resolution mode (> 30,000) using 250 ms accumulation time per spectrum. The 20 most abundant precursors ions per cycle at a threshold of 50 counts per second and peptides carrying from 2 up to 5 positive charges were chosen for fragmentation from each MS spectrum with 50 ms minimum accumulation time for each precursor. Dynamic exclusion was set 15 s, and then the precursor was refreshed off the exclusion list. Tandem mass spectra were recorded in high sensitivity mode (resolution > 15 000) with rolling collision energy on and with a collision energy spared of 15 V.

The late SCX fractions were analyzed with Nano-UPLC-MS/MS on a Proxeon EASY-nLC 1000 (Thermo Scientific, Odense, Denmark) connected to an LTQ-Orbitrap Velos (Thermo Fisher Scientific, Bremen, DE). The injected sample was first trapped with a double-fritted trapping column (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) before being separated in an

analytical column (Agilent Zorbax SB-C18, 1.8 µm, 35 cm x 50 µm). Solvent A consists of 0.1 M acetic acid while solvent B is 0.1 M acetic acid in 80% acetonitrile. Measurement time for each sample took 120 min. Samples are first loaded at a maximum pressure of 980 bar with 100% solvent A. Subsequently, peptides are chromatographically separated by a 91min gradient consisting of 15% to 40% solvent B at an un-split flow of 100 nL/min; then ramped to 100% B in 3 min and held in 100% B for another 2 min. This is finally followed by a 13-min equilibration with 100% A. For MS analysis, 1.7 kV was applied to the Nanospray needle. The survey scan was from 350 to 1500 m/z at a resolution of 30000 and for the MS2 the resolution was set to 7500. The 10 most intense precursors were selected for subsequent fragmentation using a direct dependent acquisition. A decision tree method previously described was used(Frese et al., 2011), only selecting the peptides with charge state higher than 3. Briefly, it chooses between ETD (with orbitrap or ion trap readout) and HCD as the fragmentation technique depending on the charge and m/z value of the precursor ion.

MS peak list generation

For Wiff files generated from TripleTOF 5600, tandem MS spectra were de-isotoped, charge- deconvoluted and peak lists converted to Mascot generic format (MGF) files using AB Sciex Data Converter (version 1.1). For data generated from the LTQ-Orbitrap Velos, Raw files were converted to MGF files using Proteome Discoverer (version 1.3). The non-fragment filter was used to simplify ETD spectra and the Top N filter (10 highest peaks admitted per 100 Da) for the HCD spectra. Three MGF files were generated (one for HCD, one for ETD IT and one for ETD FT). The files with an orbitrap readout were deisotoped and charge deconvoluted with the H-Score script, described elsewhere(Savitski et al., 2010).

Protein database searching

All MGF files were queried with Mascot search engine (version 2.3) via Proteome Discoverer version 1.3 (PD 1.3, Thermo Fisher) for submission. The spectra were searched against the in-house database named RAT_COMBINED. Each of the five different enzymes used (Trypsin/P, LysC/P,

Chymotrypsin, GluC-DE and AspN_ambic) were selected per file and up to 9 missed cleavages were allowed. Cysteine carbamidomethylation was set as fixed modification, and oxidation of methionine and acetylation of the N-terminal as variable modifications. Peptide tolerance was initially set to 50 ppm and the MS/MS tolerance was set to 0.1 Da (for TOF readout), 0.02 Da (Orbitrap readout) and 0.5 Da (ion trap readout).
All peptide-spectrum matches (PSMs) were evaluated with Percolator(Käll, Canterbury, Weston, Noble, & MacCoss, 2007) for validation. We classified each PSM based on their q value. We set a high stringency filter of q = 0 (0% FDR). The expected protein level FDR therefore is also 0%. Furthermore, only PSMs with a first rank in the search engine and a minimum length of 6 amino acids were kept. Unmatched spectra were exported for further analysis using PEAKS Studio (version 6.0). Peak lists were first filtered with a quality value of 0.65 as suggested by the manufacturer, followed by a tag database search. In this step, both peptide tolerance and MS/MS tolerance were set according to the Mascot search. To increase the options for matching the unassigned spectra, we additionally set deamidation of asparagine and glutamine, and pyro-glu from glutamic acid and glutamine as variable modifications, on top of the modifications used in the Proteome Discoverer search. The maximum allowed variable PTM per peptide was set to 3. Finally *de novo* interpreted PSMs were submitted to PEAKS DB database matching, this time allowing semi-enzymatic specificity and a maximum cleavages per peptide of 2. Again, the sequence database used was the newly constructed RAT_COMBINED database. The FDR was estimated using a concatenated decoy database and data was filtered according to an FDR threshold of 0.0%.

Quantitative comparison of proteome and transcriptome data

To combine quantitative data from all analysis methods, we developed a relational database schema (Supplementary Figure S5) to store all data necessary for analysis. The database schema was converted to Java (Java SE 7, Oracle California USA) entities, using the Java Persistence API (JPA version 2) implemented in EclipseLink version 2.3.2 (http://www.eclipse.org/eclipselink), using the tools provided in Netbeans IDE 7.3 (http://www.netbeans.org). The database used was MySQL version 5.5 (Oracle, California USA).

Briefly, the following steps were used to obtain spectral counts for our data:

- Import ensemble transcript to protein conversion table. Transcript identifiers were converted to Ensembl Protein identifiers using a list exported from Ensembl using the Biomart67 tool.
- Import RAT_COMBINED sequences, disambiguate protein sequences, create mapping table between non-redundant sequences and the original identifiers, create table to map the redundant identifiers to standard identifiers.
- Import Unimod: Entity classes were generated from the unimod schema using JAXB-2 Maven Plugin (version 0.8.2), and the unimod (http://www.19 Jul 2012) xml file was imported into the database.
- Import RNA-seq quantification, import PEAKS data, import Proteome Discoverer PTMs, group all PSM sequences into peptides, map all peptide sequences to the protein sequences, count all PSMs for the non-redundant protein entries in the database, count the PSMs only for the "unique" peptide sequences, calculate SAF values based on protein lengths.

RNA-seq quantifications are read from tabular text files containing mRNA abundance normalized to RPKM values. The proteomics data was read as PTMs from the PD .msf files using thermo_msf_reader (Colaert *et al*, 2011) version 2.0.3, only high quality matches (set to 0% FDR in PD) and first ranking search engine hits were accepted. Peaks results, exported as tab delimited peptide files were read into the database and filtered by FDR cutoff threshold scores determined in the Peaks software. In the database, peptides were grouped based on sequence and matched to the known protein sequences from the RAT_COMBINED database. For every protein the spectral counts were calculated based on the number of PSMs present in the database per strain. A semi-quantitative measure for expression SAF (spectral abundance factor) was calculated by dividing the spectral counts by the amino acid length of the protein. RNA-seq quantification values were averaged across all individuals (male and female), after confirmation that the quantification values were highly correlated (not shown). Tabular data was exported from the database and imported in R for further analyses.Significance analysis of different strains and the $\log_2$ ratios follow a

normal distribution with so-called "heavy" tails. If we assume most gene products do not change under different conditions, we propose that the genes present in these heavy tails are affected by an experimentally introduced bias. For this reason, we fitted a standard normal distribution using robust measures for standard deviation using the median absolute deviation (MAD), multiplying the outcome by 1.4826 to obtain the standard error and the median for the average. Q-q plots of the fitted values show that values follow a Gaussian distribution with heavy tails, indicating that the robust estimator can be used to infer the null distribution (Supplementary figure S4 B). From this normal distribution, p-values were calculated for every protein and corrected for multiple testing (1% FDR, two-sided) using the Benjamini-Hochberg step-up procedure.

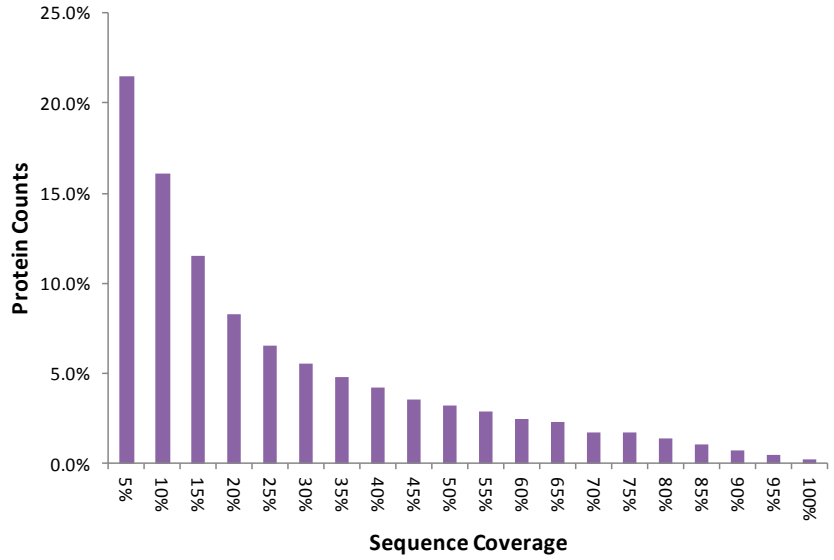
Supplementary tables

The supplementary tables are not depicted here for lack of space, please refer to the original publication.

Supplementary figures

Figure S1. Distribution of annotated genes, RNA-seq identified transcripts and proteomics-identified proteins according to rat chromosomes. Identified proteins are evenly distributed over the rat chromosomes, concordant to the distribution of genes and transcripts. The right part of the figure displays the physical location of the differential expression of the proteins on the chromosomes on a color scale from red (SHR highest) to white (BN-Lx highest). The end caps of the chromosomes are for graphical display only and do not represent physical dimensions of telomeres.

Rattus norvegicus chromosomes
12
20
19
18
16
17
11
10
15
14
13
9
8
7
X
6
5
3
4
2
1
Chromosome
0
1000
2000
3000
BNLX PSM
BNLX RPKM
SHR PSM
SHR RPKM
Genes



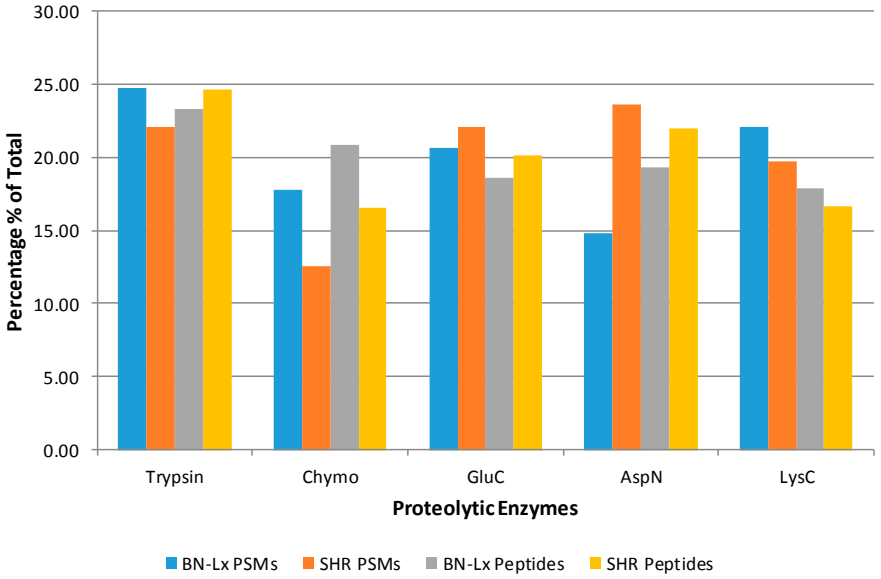


Figure S2. Comprehensiveness of proteomics data. A. The distribution of sequence coverage of proteins at a false discovery rate (FDR) of 0%. About 18% of proteins harbor at least 50% coverage. B. On average, each of the five proteases used contributed to at least 20% of peptide-spectral matches (PSMs) and non-redundant peptide sequences with trypsin being the best overall, about 25%.

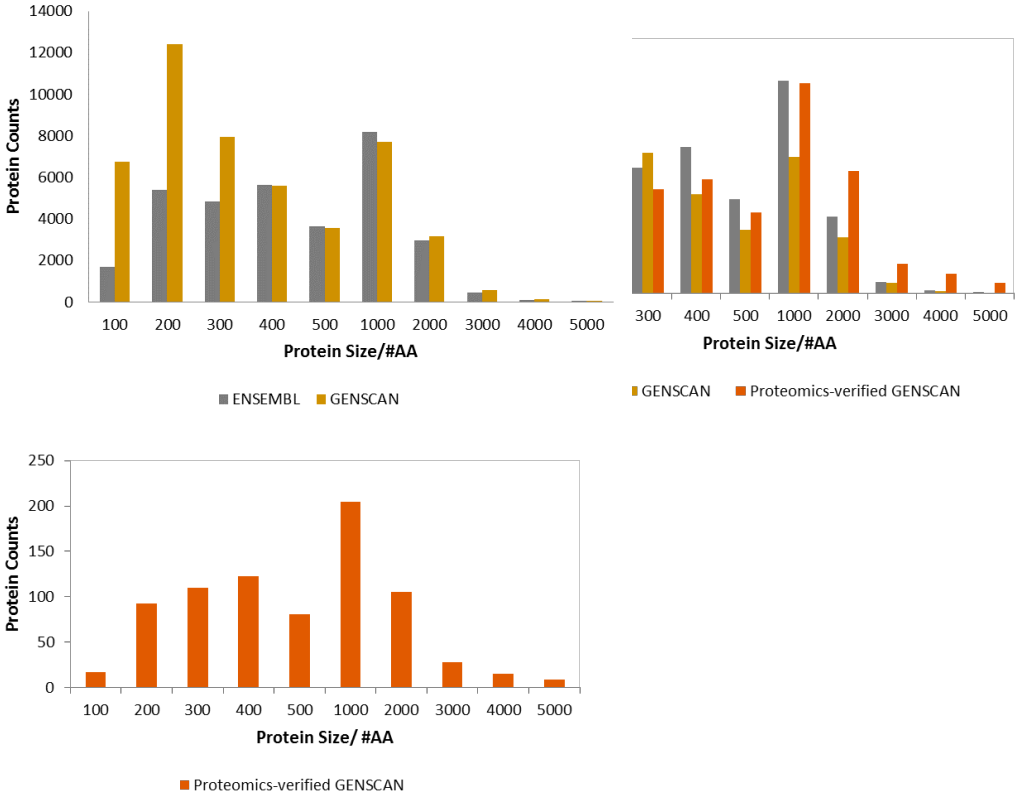


Figure S3. The distribution of size for all GENSCAN predicted proteins validated with proteomics. **(A.)** The distribution of sizes for all protein entries retrieved from ENSEMBL and GENSCAN databases shows that while GENSCAN predictions agree with ENSEMBL annotation for proteins larger than 400 amino acids; for proteins comprising < 300 amino acids, GENSCAN predicts a lot more proteins than those that are validated by ENSEMBL. **(B.)** At an FDR of 0% and using at least one unique peptide as supporting evidence to verify predicted gene models, we find that proteins < 100 amino acids are indeed present at comparatively lower levels in rat liver tissues. **(C.)** When data from **(A.)** and **(B.)** are normalized and plotted together, it is clear that both ENSEMBL-annotated proteins and MS-verified gene models agree in size distribution.

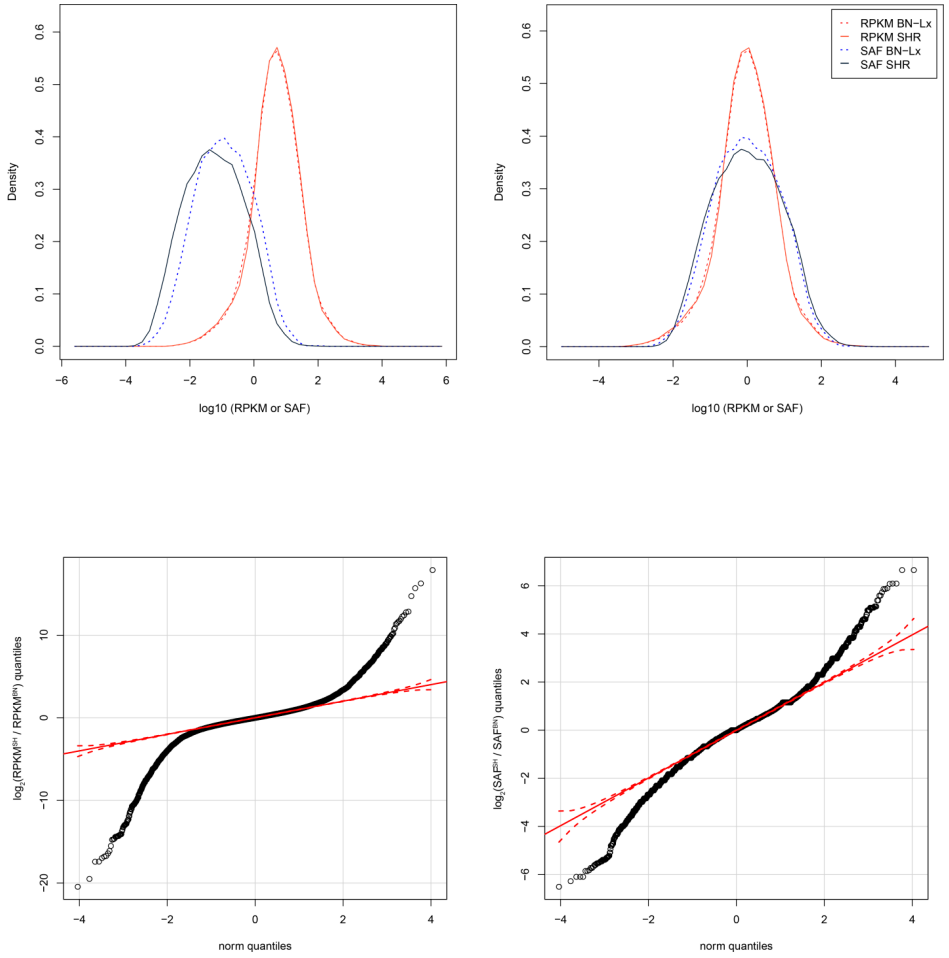


Figure S4.(A.) Normalization performed on the RPKM and SAF values by subtracting the median values of the measurements. Left panel: density plot of the distributions of RPKM and SAF values of both strains before normalization, Right panel: density plot of the same distributions after normalization. **(B.)** Q-Q plots of the distribution of the log2 ratios between SHR and BN-Lx measured as RPKM (left panel) and SAF (right panel). The plots show good concordance between the predicted distribution and the measured values in the central part, while the tails diverge from the expected line, indicative of the differentially expressed part of the measured distribution.

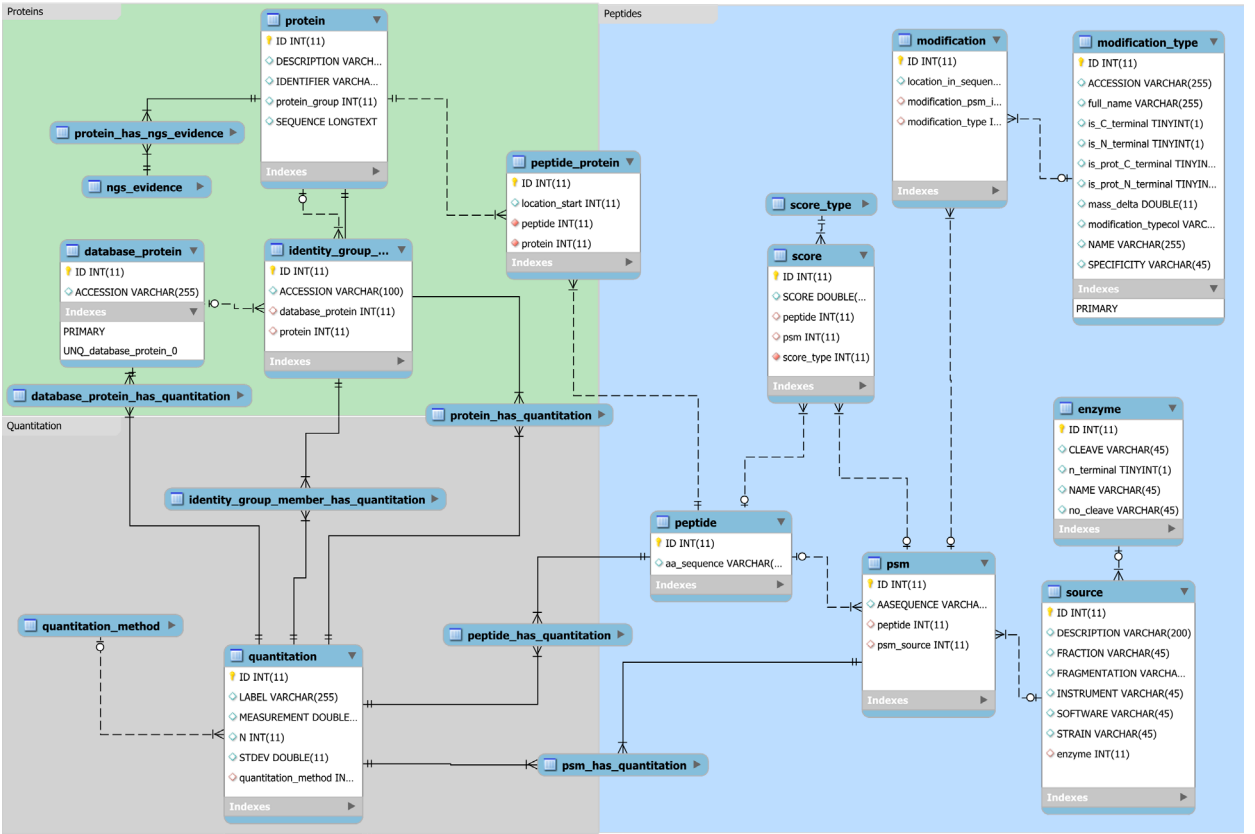


Figure S5. The schema of the relational database developed to store all transcriptomics and proteomics data. The database schema was converted to Java (Java SE 7, Oracle California USA) entities, using the Java Persistence API (JPA version 2) implemented in EclipseLink version 2.3.2 (http://www.eclipse.org/eclipselink), using the tools provided in Netbeans IDE 7.3 (http://www.netbeans.org). The database used was MySQL version 5.5 (Oracle, California USA).

Supplementary References

Atanur, S. S., Birol, I., Guryev, V., Hirst, M., Hummel, O., Morrissey, C., Behmoaras, J., et al. (2010). The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. *Genome Research*, *20*(6), 791–803. doi:10.1101/gr.103499.109

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., et al. (2004). An overview of Ensembl. *Genome Research*, *14*(5), 925–928. doi:10.1101/gr.1860604

Colaert, N., Barsnes, H., Vaudel, M., Helsens, K., Timmerman, E., Sickmann, A., Gevaert, K., et al. (2011). Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. *Journal of proteome research*, *10*(8), 3840–3. doi:10.1021/pr2005154

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., & Clamp, M. (2004). The Ensembl automatic gene annotation system. *Genome research*, *14*(5), 942–50. doi:10.1101/gr.1858004

First insight into the human liver proteome from PROTEOME(SKY)-LIVER(Hu) 1.0, a publicly available database. (2010).*Journal of Proteome Research*, *9*(1), 79–94. doi:10.1021/pr900532r

Foster, L. J., De Hoog, C. L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V. K., & Mann, M. (2006). A mammalian organelle map by protein correlation profiling. *Cell*, *125*(1), 187–199. doi:10.1016/j.cell.2006.03.022

Frese, C. K., Altelaar, A. F. M., Hennrich, M. L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A. J. R., et al. (2011). Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of Proteome Research*, *10*(5), 2377–2388. doi:10.1021/pr1011729

Hubbard, T. (2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*(1), 38–41. doi:10.1093/nar/30.1.38

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*(11), 923–925. doi:10.1038/nmeth1113

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation*, *2011*(0), bar049. doi:10.1093/database/bar049

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome research*, *12*(4), 656–64. doi:10.1101/gr.229202. Article published online before March 2002

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, *26*(16), 2069–70. doi:10.1093/bioinformatics/btq330

Pan, C., Kumar, C., Bohl, S., Klingmueller, U., & Mann, M. (2009). Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Molecular & Cellular Proteomics: MCP*, *8*(3), 443–450. doi:10.1074/mcp.M800258-MCP200

Savitski, M. M., Fischer, F., Mathieson, T., Sweetman, G., Lang, M., & Bantscheff, M. (2010). Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *Journal of the American Society for Mass Spectrometry*, *21*(10), 1668–1679. doi:10.1016/j.jasms.2010.01.012

Shi, R., Kumar, C., Zougman, A., Zhang, Y., Podtelejnikov, A., Cox, J., Wiśniewski, J. R., et al. (2007). Analysis of the mouse liver proteome using advanced mass spectrometry. *Journal of Proteome Research*, *6*(8), 2963–2972. doi:10.1021/pr0605668

Simonis, M., Atanur, S. S., Linsen, S., Guryev, V., Ruzius, F.-P., Game, L., Lansu, N., et al. (2012). Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. *Genome biology*, *13*(4). doi:10.1186/gb-2012-13-4-r31

Sun, A., Jiang, Y., Wang, X., Liu, Q., Zhong, F., He, Q., Guan, W., et al. (2010). Liverbase: a comprehensive view of human liver biology. *Journal of Proteome Research*, *9*(1), 50–58. doi:10.1021/pr900191p

Yan, W., Lee, H., Deutsch, E. W., Lazaro, C. A., Tang, W., Chen, E., Fausto, N., et al. (2004). A dataset of human liver proteins identified by protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry. *Molecular & Cellular Proteomics: MCP*, *3*(10), 1039–1041. doi:10.1074/mcp.D400001-MCP200

Ying, W., Jiang, Y., Guo, L., Hao, Y., Zhang, Y., Wu, S., Zhong, F., et al. (2006). A dataset of human fetal liver proteome identified by subcellular fractionation and multiple protein separation and identification technology. *Molecular & Cellular Proteomics: MCP*, *5*(9), 1703–1707. doi:10.1074/mcp.M500344-MCP200

8. Summary, samenvatting, future outlook, curriculum vitae, publications, acknowledgements

Summary

In this thesis, I describe methods and applications of computer techniques and bioinformatics in the field of mass spectrometry based proteomics. **Chapter 1** contains a basic introduction into the field of proteomics, how computers are an essential part of this research area, and some background on the biological questions addressed in this thesis.

There were several specific issues in the analysis of mass spectrometry based proteomics data that we attempted to address by developing our own software. We describe these issues in chapters 2 and 3. The first issue was that the number of sequencing events in the mass spectrometer reached such size that the search engine output became cumbersome. We therefore developed a software suite called RockerBox, which we describe in detail in **chapter 2**. This software reduces the size of the "dat" file output of Mascot, which is an intermediary file containing all the search results in an unformatted way. The size reduction consisted of filtering out low-quality matches. There are three ways of filtering: Using simple score thresholds, by calculating false discovery rate (FDR) values based on target and decoy database matches and introducing a score threshold for a FDR threshold, or by calculating the FDR with the Percolator algorithm. The filtered file retains the correct structure, making it possible to visualize the results in the Mascot output window, but also to use the file in other software that rely on this file format. Aside from filtering, extensive visualization methods are present to assess the quality of each search result.

A second issue was the lack of an easy-to-use statistical analysis package for quantitative data. We therefore developed a graphical desktop package called Statquant, described in **chapter 3**. It performs outlier detection and manual curation of peptide quantification with real-time adjustment of the significance values. It also provides methods to compensate for chemical artifacts like the well-known arginine to proline conversion that can occur in SILAC labeling.

An important part of proteomics research is the optimization of the methods. In **chapter 4**, we therefore investigated the difference between two different fragmentation methods, electron transfer dissociation (ETD), and the conventional collision induced dissociation (CID). The major conclusion is that ETD and CID are complementary fragmentation techniques that are optimal for different populations of peptides, based on charge, and

therefore length. Based on these findings we suggest the use of decision tree analysis of different peptides during the mass spectrometry. **Chapter 5** describes another optimization for the detection of peptides is combining the results of different peptidases on the same samples. As a proof of principle, we created a human phosphopeptide atlas, based on peptides acquired with five different proteolytic enzymes, followed by Ti4+ enrichment. We show that the different enzymes yield different and orthogonal peptide species. This lead to the detection of 37,771 unique phosphopeptides, containing 18,430 high quality unique phosphosites. We published the results with an interactive web site that contains different views of the data that allows the retrieval of the spectra of interest.

The widespread functional occurrence of tyrosine phosphorylation may have been introduced in life in parallel with the origin of multicellularity. To investigate this, we focused on the kinome of one of the most primitive animals, Trichoplax adhaerens, as described in **chapter 6**. We could identify and quantify several kinases, among which several tyrosine kinases. Notably, we also found an increased number of tyrosine phosphorylation events compared to other species, supporting the theory that the evolution of multicellularity co-occurred with the origination of tyrosine kinases. This research would not have been possible without the sequencing of the Trichoplax genome, but it did require annotation of the kinases. Since innovations of DNA and RNA sequencing allows sequencing of not only individual organisms, but also single organs, we were able to compare and analyze the liver of two using proteogenomics analysis. The experiments are described in chapter 7, where we compare two different rat strains, the standard laboratory strain Bn-Lx and a related strain called spontaneously hypertensive rat (SHR). We were able to link a germ line mutation in the promoter region of Cyp17a1, to the reduction of expression at both RNA and protein level in the SHR background. An established link between Cyp17a1 and hypertension in human existed before, but here we show it appears to exist for rat as well. With these experiments, it is clear that the combination of proteomics and next generation sequencing techniques adds mutual benefit, in that the novel and organism-specific sequencing data is used for database searching. Conversely, the proteomics analysis can verify the expression of certain specific variants, while adding the possibility to analyze post-translational modifications.

Samenvatting

In dit proefschrift worden computer en bioinformatica methoden beschreven, die ondersteunend zijn voor de bestudering van eiwitten met behulp van massaspectrometrie (*proteomics*). Hoofdstuk 1 bevat een introductie over proteomics, en hoe computers een essentieel deel vormen van dit onderzoeksgebied. Verder geef ik enige achtergrond over de biologische vragen die in dit proefschrift worden behandeld.
Massaspectrometrie gebaseerde proteomics kende een aantal technische problemen waarvoor wij een oplossing creëerden door zelf programmatuur te ontwikkelen. Deze problemen en onze oplossingen worden beschreven in **hoofdstukken 2 en 3**. Het eerste probleem dat, zeker in het verleden, veel problemen opleverde was dat het aantal sequentie-gebeurtenissen in de massaspectrometer voor een enkel monster zo groot werd, dat het lastig werd de resultaten van de analyse in een computer te bekijken of te bewerken. Om dat te adresseren ontwikkelden we programmatuur met de naam 'RockerBox', dat we beschrijven in **hoofdstuk 2**. Het programma reduceert de grootte van de analyseresultaten van de Mascot software, een zoek-algoritme dat eiwitsequenties toewijst aan massaspectra. Bestanden van het Mascot-bestandstype '.dat' kunnen worden verkleind door het deel van de resultaten met een lage kwaliteit volledig uit het bestand te verwijderen. Er zijn hiervoor drie methoden van filteren aanwezig: het gebruik van eenvoudige criteria zoals score, een meer geavanceerde criterium gebaseerd op de schatting van het aantal vals positieven (*False Discovery Rate, FDR*), door de zoekresultaten van een database met echte eiwitsequenties ('*target*') te vergelijken met een die gehusselde sequenties bevat ('*decoy*'). Een derde methode probeert deze FDR te berekenen met behulp van het 'Percolator' algoritme, dat gelijktijdig gebruik maakt van veel verschillende eigenschappen van de zoekresultaten. Het gefilterde .dat bestand behoudt de structuur, waardoor Mascot zelf, of andere analysesoftware de hoge kwaliteit data kan openen. Buiten filteren kan 'RockerBox' uitgebreide plots creëren voor de inschatting van de kwaliteit van de zoekresultaten. In **hoofdstuk 3** wordt de 'StatQuant' programmatuur beschreven, een grafisch computerprogramma waarmee op eenvoudige wijze statistische analyse van kwantitatieve gegevens kan worden uitgevoerd. Visuele inspectie van de gekwantificeerde gegevens, en het manueel aanpassen is mogelijk, alsook het automatisch detecteren

en verwijderen van uitbijters, terwijl de p-waarden van de significantietest onmiddellijk wordt aangepast. Als een specialistisch voorbeeld kan het ook compenseren voor chemische artefacten zoals de bekende arginine naar proline conversie wanneer men SILAC-labels gebruikt.
Een belangrijk deel van proteomics is de optimalisering van methodes. In **hoofdstuk 4** wordt daarom gekeken naar het verschil tussen twee verschillende fragmentatiemethoden, te weten elektronoverdracht dissociatie (*electron transfer dissociation*, ETD) en de meer conventionele botsing geïnduceerde dissociatie (*collision induced dissociation*, CID). Hier is de belangrijkste conclusie, dat ETD en CID complementaire technieken zijn, die optimaal werken op verschillende populaties van peptiden, in het bijzonder met verschillende lading, dus ook op lengte. Gebaseerd op dit onderzoek suggereren wij het gebruik van een beslis-schema voor het selecteren van de dissociatiemethode tijdens de massaspectrometrie.
In **hoofdstuk 5** beschrijven we een andere methode van optimalisatie, namelijk die van het gebruik van verschillende proteasen op hetzelfde monster. Als een *proof of principle* hebben we een fosfopeptide-atlas gebouwd, gebaseerd op peptiden van eiwitten geknipt met verschillende proteolytische enzymen. De fosfopeptiden werden verrijkt met behulp van een Ti^{4+}-kolom. We rapporteren 377.771 unieke fosfopeptide, met in totaal 18.430 unieke gefosforyleerde locaties op het eiwit. Om deze resultaten toegankelijk te maken, hebben we daarvoor een uitgebreide interactieve webpagina ontwikkeld, waarmee op verschillende manieren in de gegevens kan worden gezocht en gebladerd, en kan worden gedownload, tot op het niveau en het van het individuele fragmentatiespectrum.
Het brede voorkomen van functionele fosforylering van tyrosine zou gedurende de evolutie parallel kunnen zijn ontstaan met meercelligheid. Om dit verder te onderzoeken hebben we een van de meest primitieve dieren onderzocht, te weten het plakdiertje *Trichoplax adhaerens*. Dit onderzoek is beschreven in **hoofdstuk 6**. We konden verschillende kinasen, waaronder tyrosine kinasen identificeren en kwantificeren. Interessant genoeg vonden we ook een verhoogde fractie van tryosine fosforylatie vergeleken bij andere diersoorten, wat de theorie van gelijktijdigheid van het ontstaan van tyrosine fosforylatie en meercelligheid onderschrijft. Dit onderzoek zou niet mogelijk zijn geweest zonder dat de DNA-sequentie van het *Trichoplax* genoom aanwezig was, het zelf annoteren van de

kinasen was echter nog wel nodig.
Innovaties op het gebied van het nucleotide sequentiebepaling, bekend als *next gen[eration] sequencing*, maken het steeds beter mogelijk DNA- en RNA-sequenties binnen steeds kortere tijd te verkrijgen. Daardoor wordt het ook steeds beter mogelijk om te focussen kleinere deelgebieden, zoals de verschillende organen van een enkel organisme. In **hoofdstuk 7** combineerden we deze techniek met *proteomics* om de lever van twee verschillende stammen van ratten te vergelijken: de standaard laboratoriumrat Bn-Lx en een afgeleide stam die spontane hypertensie vertoont, genaamd SHR (*spontaneous hypertensive rat*). We maten een verhoogde hoeveelheid van zowel het RNA als het eiwit van het Cyp17a1 gen. Voor dit gen bestaat al wel een verband met hypertensie in de mens, maar was nog niet bekend bij ratten. Buiten de directe resultaten van deze experimenten, is het duidelijk dat het combineren van *proteomics* en *next gen sequencing* een voordeel bieden dat beide kanten op werkt. Nieuwe DNA en sequenties maken het mogelijk om in de massaspectrometrie zoekmachine meer, en individu-specifieke eiwitten te vinden. Van de andere kant maakt de proteomics analyse duidelijk welke eiwitten daadwerkelijk tot expressie komen, ook als het gaat over specifieke varianten. Daarnaast kan essentiële informatie worden gevonden over de post-translationele modificaties van de eiwitten. De combinatie van beide technieken is daarom zeer waardevol.

Future Outlook

Software for analysis of proteomics data

In parallel with the rapid advancement of proteomics technologies there have been many improvements in the software for the analysis of mass spectrometry data. Developing good user-friendly software always takes time, it is always somewhat behind compared to developments in instrumentation. Therefore, we needed to develop in-house tools that would also provide ideas for future developments. The functionality in the RockerBox program can now be found in many other software tools, notably Proteome Discoverer (Thermo Fisher Scientific) or MaxQuant (Cox and Mann, 2008; Cox et al., 2011) and Perseus (Tyanova et al., 2016) with their built-in visualization of results, such as freely assignable plotting commands. Luckily, one of the major bottlenecks in the analysis that were addresed by the RockerBox software: the lack of memory and processing power, are largely overcome with the growth of possibilities with modern day computers.

Improvements in proteomics mass spectrometry

The comparison between Electron Transfer Dissociation (ETD) and Collision Induced Dissociation (CID) revealed that these techniques have different efficiencies for different physical properties of peptides, making them excellent complementary techniques for the analysis of complex samples. Since the publication of these findings, there have been many new strategies added to the toolkit for analyses of samples. A direct consequence of the realization of the differences between ETD and CID is the decision tree method: Mass spectrometers selectively use ETD or CID for dissociation based on the properties of the selected peptide (Swaney et al., 2008). A variant of CID method called HCD (Higher-energy Collisional Dissociation) is available especially for Orbitrap mass spectrometers, with no lower mass cutoff and slightly different fragmentation bias, again having a complementary role next to CID and ETD. Interestingly, instrumentation now allows for the simultaneous use of both ETD and CID or HCD fragmentation on the same precursor. The combination of ETD and HCD

leads to very high fragmentation coverage of peptides with a wide range of physicochemical properties (Frese et al., 2012, 2013; Mommen et al., 2014). Even though current software is able to analyze this kind of fragmentation spectra, there is still room for improvement in terms of weighting of fragments for scoring. Another important aspect of identification is the type of peptide that is presented to the mass spectrometer. Trypsin is the traditional choice for most mass spectrometry experiments, since it cleaves at the basic residues lysine (K) and arginine (R), so that a charge can be present both on the N-terminus and at the K or R amine group. Therefore, we also test different proteolytic enzymes that create peptides with different chemicophysical properties. An example of these enzymes is Lys-N that cleaves the proteins N-terminally, before a lysine (Gauci et al., 2009). Not only does this create longer peptides than trypsin (that cleaves on both lysine and arginine) but it also concentrates the charge on the N-terminal side of the peptide, yielding a single ion series that can be sequences following the subsequent fragment peaks, in a de-novo sequencing method (Gauci et al., 2009; Hennrich et al., 2009; Taouatas et al., 2008). But not only for ETD enzymes with a different specificity will produce orthogonal results, since sometimes a tryptic enzyme does not contain an optimal charge distribution for fragmentation, especially if the peptide contains a charged modification like phosphorylation. Combining different enzymes therefore will give a great increase of detection at the cost of analysis time. We have successfully used the well-known enzymes in our phosphoproteome atlas, but it would probably be advantageous to investigate additional novel proteases that would give peptides with even more distinctive properties (Tsiatsiani and Heck, 2015).

Combining multiple –omics techniques

In the multidisciplinary analysis of the rat liver, we found clear advantages and disadvantages of the combination of DNA sequencing, RNA sequencing and proteomics. Notably, there is a difference in sampling efficiency, since an amplification step, as provided by PCR for DNA, does not exist for proteins. Having DNA and RNA sequences of specific individual samples is a major advantage for proteomics, as the generally available common genome database does not cover all genetic variation for each individual species. Conversely, it is not always clear if RNA that contains a point mu-

tation will actually be active as a protein. In our study, we found that the Cyp17a1 gene promoter mutation lead to decreased levels of CYP17A1 protein in the spontaneous hypertensive rat. Intriguingly, the moment of measurement is before the onset of hypertension, so future research may elucidate whether the level of Cyp17a1 is useful as an early indicator for hypertension. Although not the focus in this study, proteomics is also the only technique that can elucidate post-translational modifications. This is information vital to understand function as PTMs play a crucial role in many biological processes.

There are still a lot of a challenges in combining RNA-sequencing and proteomics quantification because of the typically low correlation in quantification between these data. Therefore, combining the expression levels to reinforce the quantification between both techniques does not lead to a more powerful analysis. Interestingly, the ratios between two conditions do correlate well between the protein and RNA measurements, meaning that if the level of a particular mRNA doubles, the level of the translated protein is likely to double as well. Therefore, optimistically, there may be a fixed and gene-specific conversion factor between mRNA and protein levels that by the calculation of ratios falls out of the equation. The fact that this prediction was made independently in several studies (Edfors et al., 2016; Wilhelm et al., 2014) from different angles reinforces this idea. Ideally, it would be possible to estimate the conversion factor from intrinsic properties of the RNA coding or non-coding sequences.

Phosphorylation in evolution

Investigating the Trichoplax kinome and phosphoproteome reinforced the ideas put forward earlier that the emergence of the tyrosine kinase system (Lim and Pawson, 2010) ran parallel with the formation of multicellular organisms and inter-cellular signaling. To further prove this hypothesis, it would be beneficial to investigate more animals at the base of the evolutionary tree, such as colony-forming single cells to a state that involves differentiating cells. Unfortunately, there is no possibility to roll back time and observe what happened, but with more circumstantial evidence, we should be able to make a good guess, and understand part of how life as we know it now evolved.

References

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J. V, and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. 10, 1794–1805.

Edfors, F., Danielsson, F., Hallström, B.M., Käll, L., Lundberg, E., Ponten, F., Forsström, B., and Uhlen, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Syst. Biol. 12, 883.

Frese, C.K., Altelaar, A.F.M., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A.J.R., and Mohammed, S. (2012). Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Anal. Chem. 84, 9668–9673.

Frese, C.K., Zhou, H., Taus, T., Altelaar, A.F.M., Mechtler, K., Heck, A.J.R., and Mohammed, S. (2013). Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). J. Proteome Res. 12, 1520–1525.

Gauci, S., Helbig, A.O., Slijper, M., Krijgsveld, J., Heck, A.J.R., and Mohammed, S. (2009). Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. Anal. Chem. 81, 4493–4501.

Hennrich, M.L., Boersema, P.J., van den Toorn, H., Mischerikow, N., Heck, A.J.R., and Mohammed, S. (2009). Effect of chemical modifications on peptide fragmentation behavior upon electron transfer induced dissociation. Anal Chem 81, 7814–7822.

Lim, W.A., and Pawson, T. (2010). Phosphotyrosine signaling: evolving a new cellular communication system. Cell 142, 661–667.

Mommen, G.P.M., Frese, C.K., Meiring, H.D., van Gaans-van den Brink, J., de Jong, A.P.J.M., van Els, C.A.C.M., and Heck, A.J.R. (2014). Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). Proc. Natl. Acad. Sci. U. S. A. 111, 4507–4512.Swaney, D.L., McAlister, G.C., and Coon, J.J. (2008). Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat. Methods 5, 959–964.

Taouatas, N., Drugan, M.M., Heck, A.J.R., and Mohammed, S. (2008). Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase. Nat. Methods 5, 405–407.

Tsiatsiani, L., and Heck, A.J.R. (2015). Proteomics beyond trypsin. FEBS J. 282, 2612–2626.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods 13, 731–740.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587.

Curriculum vitae

2007-now	**Bioinformatician at Biomolecular Mass Spectrometry and Proteomics Group (Prof. Dr. Albert J.R. Heck), Utrecht University** Scientific research using mass spectrometry and proteomics, Statistical analysis of proteomics experiments, Installation and maintainance of storage systems.
2002-2007	**Bioinformatician at Molecular Genetics groep (Prof. Dr. Ben Scheres), Utrecht University** Creating the Design of a web-based (lims) systeem for storage and analysis of microarray-data Statistical analysis of microarray data, Maintenance of microarray analysis software, analysis en storage of expressed-sequence tag (EST) sequence data of plant-pathogen downy mildew. Education of a student in programming and database usage, who created a web-based search system
1998– 2002	**Researcher at Molecular Genetics group (Prof. Dr. Ben Scheres), Utrecht University** Studying Plant developmental biology, Fluorescence microscopy, confocal microscopy, Molecular techniques
1990– 1998	**Study of Biology, Utrecht University** Majored in molecular genetics and evolutionary ethology

Publications

2008

Toorn, H.W.P. van den, Mohammed, S., Gouw, J.W., Breukelen, B. van, and Heck, A.J.R. (2008). Targeted SCX Based Peptide Fractionation for Optimal Sequencing by Collision Induced, and Electron Transfer Dissociation. J. Proteomics Bioinform. 1, 379.

2009

Aye, T.T., Mohammed, S., van den Toorn, H.W.P., Van Veen, T.A.B., van der Heyden, M.A.G., Scholten, A., and Heck, A.J.R. (2009). Selectivity in enrichment of cAMP-dependent protein kinase regulatory subunits type I and type II and their interactors using modified cAMP affinity resins. Mol Cell Proteomics 8, 1016–1028.

van Breukelen, B., van den Toorn, H.W.P., Drugan, M.M., and Heck, A.J.R. (2009). StatQuant: a post-quantification analysis toolbox for improving quantitative mass spectrometry. Bioinformatics 25, 1472–1473.

Hennrich, M.L., Boersema, P.J., van den Toorn, H., Mischerikow, N., Heck, A.J.R., and Mohammed, S. (2009). Effect of chemical modifications on peptide fragmentation behavior upon electron transfer induced dissociation. Anal Chem 81, 7814–7822.

2010

van Gestel, R.A. van, van Solinge, W.W., van den Toorn, H.W.P., Rijksen, G., Heck, A.J., van Wijk, R., and Slijper, M. (2010). Quantitative erythrocyte membrane proteome analysis with Blue-native/SDS PAGE. J Proteomics, pp. 456–465.

2011

van den Toorn, H.W.P., Muñoz, J., Mohammed, S., Raijmakers, R., and Heck, A.J.R. (2011). RockerBox: analysis and filtering of massive proteomics search results. J. Proteome Res. 10, 1420–1424.

Zhou, H., Low, T.Y., Hennrich, M.L., van der Toorn, H., Schwend, T., Zou, H., Mohammed, S., and Heck, A.J.R. (2011). Enhancing the identification of phosphopeptides from putative basophilic kinase substrates using Ti (IV) based IMAC enrichment. Mol. Cell. Proteomics 10, M110.006452.

2012

Bleijerveld, O.B., Wijten, P., Cappadona, S., McClellan, E.A., Polat, A.N., Raijmakers, R., Sels, J.-W.W., Colle, L., Grasso, S., van den Toorn, H.W.P., et al. (2012). Deep proteome profiling of circulating granulocytes reveals bactericidal/permeability-increasing protein as a biomarker for severe atherosclerotic coronary stenosis. J Proteome Res 11, 5235–5244.

Cote, R.G., Griss, J., Dianes, J.A., Wang, R., Wright, J.C., van den Toorn, H.W.P., van Breukelen, B., Heck, A.J.R., Hulstaert, N., Martens, L., et al. (2012). The PRoteomics IDEntification (PRIDE) Converter 2 Framework: An Improved Suite of Tools to Facilitate Data Submission to the PRIDE Database and the ProteomeXchange Consortium. Mol. Cell. Proteomics 11, 1682–1689.

Frese, C.K., Altelaar, A.F.M., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A.J.R., and Mohammed, S. (2012). Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Anal. Chem. 84, 9668–9673.

Hennrich, M.L., van den Toorn, H.W.P., Groenewold, V., Heck, A.J.R., and Mohammed, S. (2012). Ultra acidic strong cation exchange enabling the efficient enrichment of basic phosphopeptides. Anal Chem 84, 1804–1808.

2013

Halim, V.A., Alvarez-Fernandez, M., Xu, Y.J., Aprelia, M., van den Toorn, H.W.P., Heck, A.J.R., Mohammed, S., and Medema, R.H. (2013). Comparative phosphoproteomic analysis of checkpoint recovery identifies new regulators of the DNA damage response. Sci. Signal. 6, rs9-rs9.

Low, T.Y., VanHeesch, S., VandenToorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., VanBreukelen, B., Mohammed, S., et al. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 5, 1469–1478.

Pollo-Oliveira, L., Post, H., Acencio, M.L., Lemke, N., van den Toorn, H.W.P., Tragante, V., Heck, A.J.R., Altelaar, A.F.M., and Yatsuda, A.P. (2013). Unravelling the Neospora caninum secretome through the secreted fraction (ESA) and quantification of the discharged tachyzoite using high-resolution mass spectrometry-based proteomics. Parasit. Vectors 6, 335.

Ringrose, J.H.J., van den Toorn, H.W.P., Eitel, M., Post, H., Neerincx, P., Schierwater, B., Altelaar, A.F.M., and Heck, A.J.R. (2013). Deep proteome profiling of Trichoplax adhaerens reveals remarkable features at the origin of metazoan multicellularity. Nat. Commun. 4, 1408.

2014

Matheron, L., van den Toorn, H., Heck, A.J.R., and Mohammed, S. (2014). Characterization of Biases in Phosphopeptide Enrichment by Ti^{4+} -Immobilized Metal Affinity Chromatography and TiO_2 Using a Massive Synthetic Library and Human Cell Digests. Anal. Chem. 86, 8312–8320.

Walzer, M., Pernas, L.E., Nasso, S., Bittremieux, W., Nahnsen, S., Kelchtermans, P., Pichler, P., van den Toorn, H.W.P., Staes, A.A., Vandenbussche, J., et al. (2014). qcML: an exchange format for quality control metrics from mass spectrometry experiments. Mol. Cell. Proteomics 13, 1905–1913.

2015

Giansanti, P., Aye, T.T., van den Toorn, H., Peng, M., van Breukelen, B., and Heck, A.J.R. (2015). An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. Cell Rep. 11, 1834–1843.

Selvan, N., Mariappa, D., Van Den Toorn, H.W.P., Heck, A.J.R., Ferenbach, A.T., and Van Aalten, D.M.F. (2015). The early metazoan Trichoplax adhaerens possesses a functional O-GlcNAc system. J. Biol. Chem. 290, 11969–11982.

2017

Cristobal, A., Marino, F., Post, H., van den Toorn, H.W.P., Mohammed, S., and Heck, A.J.R. (2017a). Toward an Optimized Workflow for Middle-Down Proteomics. Anal. Chem. 89, 3318–3325.

Cristobal, A., van den Toorn, H.W.P., van de Wetering, M., Clevers, H., Heck, A.J.R., and Mohammed, S. (2017b). Personalized Proteome Profiles of Healthy and Tumor Human Colon Organoids Reveal Both Individual Diversity and Basic Features of Colorectal Cancer. Cell Rep. 18, 263–274.

Tsiatsiani, L., Giansanti, P., Scheltema, R.A., van den Toorn, H., Overall, C.M., Altelaar, A.F.M., and Heck, A.J.R. (2017). Opposite Electron-Transfer Dissociation and Higher-Energy Collisional Dissociation Fragmentation Characteristics of Proteolytic K/R(X)n and (X)nK/R Peptides Provide Benefits for Peptide Sequencing in Proteomics and Phosphoproteomics. J. Proteome Res. 16, 852–861.

Acknowledgements

I would like to start by thanking Albert Heck for allowing me to work in his great lab, and involving me in many different projects with a great variety of subjects. It was always a pleasure to discussing them, and more than once, you have been the person to give direction on how to move forward if we were struggling with the subject. I also want to thank you very much for giving me the opportunity to write this thesis, even after I had expressed my doubts about doing it. You let me decide for myself, but when I did, you helped me in all the ways needed to get it done.

Bas, I will always remember that you asked me to apply in the group many years ago. I was very happy to learn so much about this exciting new technology, and it has been a great experience ever since. Because of all of your diverse interests, I have done so many more things than just bioinformatics, with notable moments like filming the inaugural speech by Alexander Makarov, but also aiding in setting up the IT infrastructure in our group, leading to eating many hamburgers with company representatives. I think we have become a great team over the years and for me it is always easy to enter your office to discuss science, web pages, and education or even Apple devices.

Thank you, the rest of the 'staff'. Shabaz Mohammed, for the many discussions and for helping me out many times. Maarten thanks for the nice chats and great collaborations, Thanks for guiding us through Madrid. Richard, for always bringing many new ideas, and for showing me C#. Wei, for your nice discussions, great to see you becoming a PI. Monique, for the discussions about education and acknowledgements. Simone, for all the nice collaborations and 'fun' with peer reviews. Celia, for widening our view beyond proteins. Corine, without you, I actually think the lab will stop working. Thank you for being such a great secretary, but also a wonderful person whose common sense often resolves perceived problems before they arise.
Thanks to the lab, past and present. Adja, Alex, Ana, Andrea, Andreas,

Andrey, Aneika, Anita, Anja, Anna H., Anna R., Arjan, Arjen, Armel, Ayse, Barbara, Bas, Basak, Bohui, Celia, Celine, Charlotte, Charlotte U., Chiara, Clement, Dave, David, Dominique, Donna, Eleni, Esther, Esther van D., Fan, Fiona, Geert H., Geert M., Gianluca, Glen, Guanbo, Harm, Hongtao, Jamila, Javier, Jeffrey, Jessica, Jin, Jing, Joost S., Juan, Kees V., Kelly D., Kelly S., Kristina, Kyle, Leticia, Liana, Lucrece, Maarten, Marco H., Martje, Matina, Miao-Hsia, Michiel, Mirjam, Monique, Nadine, Nadine B., Natalie, Nicolas, Niels, Nikolai, Oleg, Onno, Pepijn, Philip, Pierre, Pieter, Qingyang, Rebecca, Reinout, Renske P., Renske van G., Riccardo, Richard, Rozalia, Saar, Sander, Sem, Serena, Shabaz, Sibel, Sietske, Simone, Soenita, Suzy, Theo, Thierry, Thomas S., Tobias, Tomislav, Vincent, Violettte, Vita, Vojtech, Wei, Wim, Yang, Yorrick and Yu-Hsien, and everyone I forgot to mention. I really enjoy working with you all. I want to mention some people of this group, with whom I spent some more time. The following is a bit of a stream-of-consciousness, therefore no particular order is intended. Thanks Teck for passing by my office so many times, just to talk to me and amaze me with your knowledge about the most diverse scientific trivia. The rat project was daunting but the collaborations were great. Alba, we've spent so much time behind the computer it was almost strange to think that the articles were finished and you're a doctor now. Thank you for defending a bit of my work during your defense. Fabio, thanks for being a great roommate, I really enjoyed having you there, also as a great travel companion. Which brings me to the 'coffee club' of which the first rule is: bring coffee and talk about coffee, or something else. Thank you for keeping me awake after lunch. Jeffrey, for the great stories and for impressing with your eating habits, and for thinking of the great research project I was allowed to help you with (published herein).
I'd like to especially thank all the technicians. Harm, for all of the 'real' work you do in the lab, so I have something to work with, Mirjam and Soenieta. Also Arjan, for his humor and being a great colleague.

Thanks also to Geert, who keeps all of the computers running, and I try to help every now and then, we're a great team, from setting up hardware to keeping the servers running.

My fellow bioinformaticians in the lab: Bas van Breukelen, Robert

Kerkhoven, Danny Navarro, Wim Spee, Alex Georgiou, Madalina Drugan, Pieter Neerincx, Mao Peng, Martin Fitzpatrick, Linsey Raaimakers, Salvatore 'Salvo' Cappadona, Richard Scheltema and Bohui Li. It is always nice to have people there that understand what you're talking about. Mao, for being a wonderful (albeit sometimes sleepy) roommate. Alex, it was really nice to work with you, I've always valued your programming skills and have learnt a lot from you. Salvo, I'll never forget our visit to the Mormon temple, I've not been the same since (not really). Pieter I enjoyed the movie nights. Martin, cheers for a lot of good fun, lunch breaks and beers, not to mention PaDua. Wim, for all the work you did. Linsey for the cheerful moments, and the possibility to talk about the practicalities of promotions. Bohui, my current roommate/bioinformatician, I really enjoy our conversations and the analysis of the Dutch language. Armel, great having you in the room (sometimes), discussing about computer science stuff that I half understand -- it's interesting anyway.

My gratitude goes out to all of the collaborators in different projects, including Edwin Cuppen, Sebastiaan van Heesch and Joep de Ligt, Hans Clevers and Marc van de Wetering, Bernd Schierwater, Michael Eitel, Lennard Martens, Niels Hulstaert, Kenny Helsens, Davy Maddelein, Florian Reisinger, Richard Côté, Juan Antonio Vizcaíno, Yasset Perez-Riverol, Attila Csordas, Oliver Kohlbacher, Lukas Käll, Harald Barsnes and Henning Hermjakob.

Next to work, I really want to thank people from my personal surroundings. I want to thank my friends and family, but first, I would like to thank my love Katrien, who is always there for me, not just as my lover, but also my friend. I would not know what to do without you. I am also very happy you want to be my paranimf.

I would especially like to thank my mother Annemarie, my brother Albert, my sister Annette, my brother-in-law Jasper, my nephew Simon, for being such a great family that I can always rely on. I am very happy Albert wants to be my paranimf, and that Annette was kind enough to design the cover for this book. Also my 'extended' family Mirjam, Daan, Jan-Robert, Carlein, Frans, Eel, Wil, and all the children.

Last but not least, thank you Birkit, Muzaffer, Willem, Constance, Mine, Emma, Remco, Marion, Pim, Pascale, Ana, Daniel, Marjolein, Lidwin, Harry, Don, Truus and Paul, for all your interest and encouragements.

Of course I would like to thank everyone I forgot to mention here.