From DEPARTMENT OF ONCOLOGY AND PATHOLOGY
Karolinska Institutet, Stockholm, Sweden

# MASS SPECTROMETRY BASED PROTEOMICS: DATA ANALYSIS AND APPLICATIONS

Yafeng Zhu

Stockholm 2018

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by E-print AB 2018

ISBN 978-91-7831-044-9

# Mass spectrometry based proteomics: data analysis and applications
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

## Yafeng Zhu

*Principal Supervisor:*
Professor Janne Lehtiö, Ph.D.
Karolinska Institutet
Department of Oncology-Pathology

*Co-supervisor(s):*
Jenny Forshed, Ph.D., M.D
Karolinska Institutet
Department of Oncology-Pathology

*Opponent:*
Professor David Fenyö
NYU School of Medicine
Department of Biochemistry and Molecular Pharmacology

*Examination Board:*
Johan Malmström
Lund University
Division of Infection Medicine

Sara Lind
Uppsala University
Department of Chemistry

Lisa Klasson
Uppsala University
Department of Cell and Molecular Biology

The public defense of this thesis will take place

9:00 am, Friday, 25th May, 2018

CMB föreläsningssalen
Berzelius väg 21
Karolinska Institutet, Solna

做学问的三重境界：

“昨夜西风凋碧树，
独上高楼，
望尽天涯路”

此第一境也。

“衣带渐宽终不悔，
为伊消得人憔悴”

此第二境也。

“纵里寻他千百度，
蓦然回首，
那人却在灯火阑珊处”

此第三境也。

古之学问大成者，罔不历此三重境界。

——王国维 《人间词话》

古人之境界，欣然往之，
虽徐徐不能近，以此勉励。

# ABSTRACT

Mass spectrometry (MS) based proteomics has become a widely used high throughput method to investigate protein expression and functional regulation. From being able to study only dozens of proteins, state-of-art MS proteomic techniques are now able to identify and quantify ten thousand proteins. Nevertheless, MS proteomics are facing problems investigating protein variants derived from alternative splicing, detecting peptides from novel coding sequences, identifying peptide variants from genetic changes and statistical analysis of quantitative proteome. The work present in this thesis start from these problems and contribute solutions to them.

In standard shotgun proteomics studies, protein identifications are inferred from a list of identified peptides using Occam Razor's rule, which outputs a minimum list of proteins sufficient to explain peptide evidences. The protein inference process creates a potential problem in protein level quantification, resulting mixture of quantitative signals from different splice variants if the inferred proteins do not correctly represent the peptide populations. **Paper I** present a tool to investigate splice variants using MS proteomics data. By clustering the quantitative pattern of peptides and showing their transcript positions, it is able to reveal splice variants specific peptides with different quantitative signal. The tool was applied to a comprehensive proteomics data of A431 cells treated with Gefitinib (EGFR inhibitor). For certain genes, we observed splice-variant-centric quantification differs from traditional protein-centric or gene-centric quantification, suggesting differentially regulated splice variants after Gefitinib treatment.

Previously, MS proteomics has been used to refine genome annotation. However, the applications were limited to validate and confirm predicted gene models. In **Paper II**, we demonstrate an integrative genome annotation workflow that combines MS proteomics data and RNA-sequencing to perform evidence-based whole genome annotation of a newly sequenced commensal yeast. The workflow showed higher accuracy of protein coding gene annotation compared to conventional way of using only RNA-sequencing data. The study exemplifies that proteomics data used in combination with RNA-seq data is able to produce a more accurate and complete whole genome annotation.

**Paper III** shows an integrative proteogenomics analysis workflow. Compared to standard proteomics which analyzes known proteins in reference database, proteogenomics aims to discover peptides from novel coding sequences and disease relevant mutations. To identify novel coding sequences in well annotated genomes, such as human, it is particular challenging due to several reasons. First, protein-coding sequences in the human genome consists of only 2%-3% of the total sequences. There are approximately one million peptides from known coding genes, and the novel peptides from undiscovered coding loci constitutes a minor part of the total peptide population. That means the vast majority of experimental spectra are produced from known peptides. Identification of peptides with MS proteomics technique relies on correct matching between experimental spectra to *in silico* generated spectra of the peptides in search space. Detecting of novel peptides requires correct spectra matching for both known and novel peptides, and the process is doomed to produce false positives. Previously, conservative criteria and manual curation has been applied to ensure the quality of findings. **Paper III** presents a workflow which improves the reliability of proteogenomics findings by automated extensive data curation and evidence searching in orthogonal data. In analysis of the proteomics data of a cancer cell line and five normal human tissues, the workflow successfully detected novel peptides from unknown coding regions and peptide variants from non-synonymous single

nucleotide polymorphisms (nsSNPs) and mutations, with multiple sources of evidence provided. Moreover, our quantitative MS data indicated that certain pseudogenes and lncRNAs were expressed and translated in tissue-specific manner.

**Paper IV** addresses the statistical analysis of quantitative proteomics. Currently, there is no consensus in the usage of statistical methods to analyze labelled and label-free proteomics data. One of the main reasons is the lack of statistical tool with high performance, ease to use, and broad applicability to various proteomics datasets. The presented statistical method, DEqMS, is a robust and universal tool to assess differential protein expression for quantitative MS proteomics. DEqMS takes into account the variance dependence on the number of peptides/PSMs used for protein quantification in statistical significance test. Compared to existing methods in several benchmarking datasets, DEqMS was demonstrated with both high statistical accuracy and general applicability.

In summary, the work included in this thesis contributes with improved data interpretation and applications of MS proteomics data in analysis of splice variants, genome annotation, proteogenomics studies and statistical analysis of protein expression changes. Development of these methods facilitate a wide range of applications of MS proteomics data in the systems biology research.

# LIST OF SCIENTIFIC PAPERS

I. **Yafeng Zhu,** Lina Hultin-Rosenberg, Jenny Forshed, Rui M. M. Branca, Lukas M. Orre, and Janne Lehtiö.
SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Molecular & Cellular Proteomics 2014 Jun;13(6):1552-62.*

II. **Yafeng Zhu**§, Pär G Engström§, Christian Tellgren-Roth, Charles Baudo, John C Kennell, Sheng Sun, R. Blake Billmyre, Markus S. Schröder, Anna Andersson, Tina Holm, Benjamin Sigurgeirsson, Guangxi Wu, Sundar Ram Sankaranarayanan, Rahul Siddharthan, Kaustuv Sanyal, Joakim Lundeberg, Björn Nystedt, Teun Boekhout, Thomas L Dawson Jr., Joseph Heitman, Annika Scheynius*, Janne Lehtiö*.
Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of Malassezia sympodialis. *Nucleic Acids Research, 2017 Mar 17;45(5):2629-2643.*
§ These authors contributed equally to this work as the first authors.

III. **Yafeng Zhu**, Lukas M. Orre, Henrik J. Johansson, Mikael Huss, Jorrit Boekel, Mattias Vesterlund, Alejandro Fernandez-Woodbridge, Rui M. M. Branca* & Janne Lehtiö*.
Discovery of coding regions in human genome using an integrated proteogenomics analysis workflow. *Nature Communications. 2018 Mar 2;9(1):903.*

IV. **Yafeng Zhu**, Lukas M. Orre, Georgios Mermelekas, Henrik J. Johansson, Alina Malyutina, Simon Anders, Janne Lehtiö.
DEqMS: a robust and universal statistical method for quantitative mass spectrometry proteomics. *Manuscript.*

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| MS | Mass spectrometry |
| TOF | Time-of-flight |
| MALDI | Matrix-assisted laser desorption ionization |
| ESI | Electrospray ionization |
| LC | Liquid chromatogram |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| FASP | Filter-aided sample preparation |
| SP3 | Single-Pot Solid-Phase-enhanced Sample Preparation |
| SCX | Strong cation exchange |
| IEF | Isoelectric focusing |
| IPG | Immobilized pH gradient |
| HiRIEF | High resolution isoelectric focusing |
| hp-RP | High pH reverse phase |
| lp-RP | Low pH reverse phase |
| PTM | Post-translational modification |
| PSM | Peptide spectrum match |
| TDA | Target decoy approach |
| FDR | False discovery rate |
| AUC | Area under curve |
| SILAC | Stable isotope labelling by amino acids in cell culture |
| iTRAQ | Isobaric tags for relative and absolute quantitation |
| TMT | Tandem mass tag |
| PCA | Principal component analysis |
| ANOVA | Analysis of variance |
| LIMMA | Linear models for microarray data |
| MNAR | Missing not at random |

| | |
|---|---|
| MAR | Missing at random |
| TCGA | The cancer genome atlas |
| ORF | Open reading frame |
| COSMIC | Catalogue of somatic mutations in cancer |
| SNP | Single-nucleotide polymorphism |

# INTRODUCTION

## From genomics to proteomics

The word *proteome*, representing the whole set of cellular proteins, was first coined by Marc Wilkins in 1996 when he carried out the analysis of 20 *E.coli* proteins using two-dimensional (2-D) electrophoresis mass spectrometry (MS)[1]. As an analogy to genomics (study of the human genome), the term "proteomics" was later created in 1997 to describe the global study of complete protein components of cells[2].

With the development of next generation sequencing techniques in the last decade[3], genomics studies have advanced our understanding of genetic mechanisms underlying various type of cancers. Proteomics serves a complementary tool for researchers who investigate protein level regulation globally in a high throughput manner.

There are two alternative approaches for proteomics, antibody-based proteomics[4-8](also referred as affinity proteomics) and MS-based proteomics[9-12]. Antibody-based proteomics relies on specific binding affinity between antibodies and their targets to detect and quantify the target proteins. This technique requires a list of predefined proteins to be measured and the availability of corresponding protein antibodies. MS-based proteomics is a hypothesis-free discovery approach to characterize proteins without dependence on antibodies. This thesis is mainly focused on MS-based proteomics.

## Mass spectrometry

Mass Spectrometry (MS) is an analytical technique to measure the mass-to-charge ratio and intensity of charged molecules or particles. The direct readouts from a mass spectrometer are mass-to-charge ratio (m/z) values and their signal intensities, which are often visualized as a *mass spectrum* with m/z values on the x-axis and intensities on the y-axis (Figure 1).

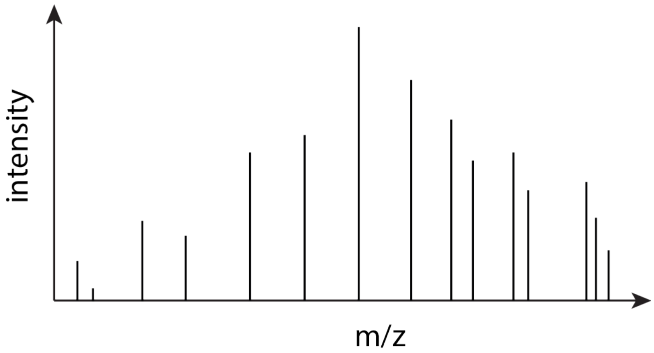

**Figure 1. Illustration of a mass spectrum.**

A mass spectrometer consists of three parts: ion source, mass analyzer and detector. Each of them is designated for a specific purpose. The ion source is to ionize the analytes so that they can be manipulated by magnetic or electric fields. Depending on the amount of energy that imparts on the analytes, ionization processes are divided into hard ionization and soft ionization. Hard ionization imparts high residual energy on the molecules causing more fragmentation (brakeage) within the molecules. In contrast, soft ionization imparts less residual energy and results in little fragmentation within the molecules. The function of a mass analyzer is to sort charged molecules according to their m/z values. Analytes with different mass move in different trajectories in electromagnetic filed according to Lorentz force law and Newton's second law.

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{\nu} \times \boldsymbol{B}) \qquad \text{Equation 1}$$

$$\boldsymbol{F} = m\boldsymbol{a} \qquad \text{Equation 2}$$

$$\boldsymbol{a} = \frac{\boldsymbol{E} + \boldsymbol{\nu} \times \boldsymbol{B}}{m/q} \qquad \text{Equation 3}$$

The above formulas explain how ions are sorted in a mass analyzer. Equation 1 is Lorentz force law where $\boldsymbol{F}$ is referred to Lorentz force. $q$ is the ion charge, $\boldsymbol{E}$ is the electric field, $\boldsymbol{\nu}$ is the ion velocity and $\boldsymbol{B}$ is the magnetic field. Equation 2 is Newton's second law where $m$ is the mass of the ion, $\boldsymbol{a}$ is the acceleration calculated based on Equation 3. For ions entering the electromagnetic field with equal velocity, the acceleration depends only on mass-to-charge ratio.

The same principles are implemented in different types of mass analyzer with different designs: time-of-flight (TOF) analyzer, quadrupole mass analyzers, and orbitrap. There are several characteristic parameters for mass analyzers: mass resolving power (resolution), mass accuracy and mass range. The mass resolution refers to the ability to distinguish two mass peaks with slightly different m/z values. Resolving power is calculated as $M/_{\Delta}m$ (molecular mass divided by peak width at half height). For the same instrument, peak width increases as the molecular mass of the analytes increase. The mass accuracy indicates the measurement error of m/z values compared to the true m/z. The mass range is the range of m/z that a mass analyzer is able to measure. The purpose of a detector is to record the ion current (signal intensity) produced when an ion hits its surface in TOF analyzer. In orbitrap analyzer, ions oscillate in high vacuum space around a central electrode (orbital motion) and horizontally in between two outer electrodes. The oscillation frequencies of different ions detected by the outer electrodes are Fourier transformed into a mass spectrum. **Orbitrap was the mass analyzer used in all four Papers (I-IV).**

The first attempt of using MS to identify amino acids (building bricks for proteins) was made by K. Biemann et al in 1959 in which they tried to identify structure formation of ethyl esters, derivatives of amino acids[13, 14]. At that time, the ionization techniques were not able to study large molecules using MS due to unintended fragmentation in the process. The problem was overcome by the development of fast atom bombardment ionization technique in 1980. In this "soft" ionization approach, macromolecules are hit by a beam of high-energy atom to generate intact ions[15-17]. This technique served as theoretical basis of two widely used ion sources, matrix-assisted laser desorption ionization (MALDI)[18] and electrospray ionization (ESI)[19], which made mass spectrometer applicable to analyze large biomolecules such as peptides and proteins.

### Soft ionization methods for peptide and protein

As previously mentioned, matrix-assisted laser desorption ionization (MALDI)[18] and electrospray ionization (ESI)[19] are two widely used ion sources. In MALDI process, the matrix is prepared as a mixture of water and organic solvent (acidic) which serves as the proton donor to ionize the analytes. After the matrix is mixed with analytes, they are co-crystallized on a metal plate. A pulsed laser is then added to irradiate the sample to trigger ionization process. The analytes carry single charge in MALDI process.

Electrospray ionization is able to ionize large biomolecules in solution, and it is commonly coupled with liquid chromatogram (LC) to analyze complex samples. Peptide mixtures are loaded into LC and separated according to hydrophobicity. Eluting peptides are dispersed by electrospray into small liquid droplets. As the droplets evaporate progressively and decrease in size, charges are split to smaller peptide droplets and finally into gas phase. In ESI process, peptides get multiple charges. **ESI was used as ion source in all four Papers (I-IV).**

## MS-based proteomics

MS-based proteomics is also known as shotgun proteomics. Similar to shotgun genomic where long double-strand DNA are broken to fragments prior to sequencing, shotgun proteomics starts by cleaving proteins into peptides, which then are analyzed by MS.

There are two commonly used sample preparation protocols for shotgun proteomics: filter-aided sample preparation (FASP)[20] and Single-Pot Solid-Phase-enhanced Sample Preparation (SP3)[21]. The workflow starts by extracting proteins from cell lysates. Proteins are then digested into peptides by proteolytic enzymes, such as trypsin. In many cases, peptide mixtures are subjected to a pre-fractionation step to reduce sample complexity before they are injected into LC-MS system. Peptides eluting out from LC are ionized in the ion source and then intact peptides are analyzed by MS, so called MS1 analysis. The m/z values of all peptide ions and their intensities are measured and depicted as MS1 spectrum. In data dependent acquisition (DDA) mode, a preset number of top intense peptides are selected for fragmentation, and an exclusion time period is set to avoid the instrument sampling the high abundant peptides redundantly. The fragmented product ions, MS2 ions, are analyzed in the second MS to acquire MS/MS (MS2) mass spectrum (Figure 2). MS2 spectrum is compared against computed spectra based on the peptide sequences in the search database. According to scoring function within the peptide search engine, the best matched peptide sequence will be assigned as the identity of the corresponding MS2 spectrum. Ideally, a pure MS2 spectrum that contains only product ions from one peptide is more likely to find a correct match. However, in analysis of complex samples, MS2 spectra usually suffer from precursor interference, resulted from impure selection of peptides at the fragmentation step[22]. **DDA mode was used in all four Papers (I-IV) to acquire the experimental data.**

### Comprehensive peptide pre-fractionation increases proteome coverage

MS is usually coupled with LC to separate peptide mixtures based on hydrophobicity. The purpose is to avoid peptide competition in ionization process and reduce precursor interference. However, the separation capacity of LC can be saturated when analyzing complex peptide mixtures. Protein digestion of whole cell lysates generates complex peptide mixtures which span a concentration range of seven orders of magnitude[23, 24]. In plasma samples, the estimated dynamic range is even bigger, spanning ten orders of magnitude[25-27]. Peptides of some highly abundant proteins occupy most of the instrument time and create a noisy background to identify peptides from low abundant proteins. Therefore, more efficient orthogonal methods for peptide fractionation are needed to increase sensitivity.

Strong cation exchange (SCX) is able to separate peptides by charge states in addition to hydrophobicity in the LC. Since this method can be easily integrated online with LC system, peptide mixtures can be directly injected to start the analysis. This approach was demonstrated to identify 1484 proteins from the proteome of Saccharomyces cerevisiae[28, 29]. The downside of SCX method is that sample desalting is required to avoid the negative effects of salts on the ionization process. Another method to pre-fractionate peptide mixtures is two-dimension LC (2D-LC) applying high pH reverse phase (hp-RP) in the first dimension and low pH reverse phase (lp-RP) in the second dimension[30]. Charge distribution of peptides at different pH conditions is changed and therefore it creates a different separation profile. This method was proved to provide separation orthogonality comparable to that of SCX-LC and allows concatenation of fractions in the first dimension to save analysis time with minimum reduction of protein identifications[30-32].

In GeLC-MS analysis, SDS-PAGE is used to separate protein lysates according to their molecular masses. Different protein bands are cut into fractions, followed by in gel digestion and LC-MS/MS analysis.

Gel-based isoelectric focusing (IEF) pre-fractionates peptide mixtures based on their isoelectric points using a polyacrylamide gel containing an immobilized pH gradient (IPG)[33-35]. Peptide IEF shows much higher resolution and reproducibility, whereas SCX method is more sensitive in the case of peptide and protein identifications at low sample amount (10 microgram)[36].

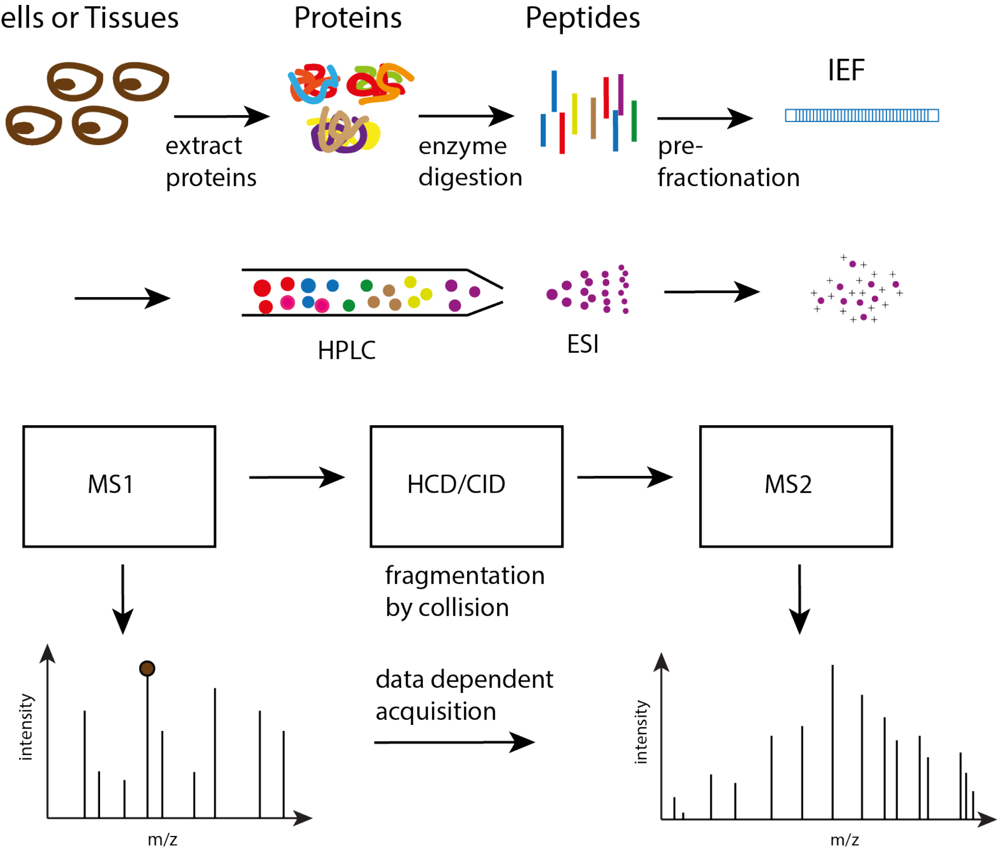

**Figure 2. Illustration of shotgun proteomics workflow.**

High resolution isoelectric focusing (HiRIEF) enables extensive separation of peptides on an immobilized pH gradient (IPG) strip which are further divided into 72 fractions[37], illustrated in Figure 3. HiRIEF coupled LC-MS was able to identify 13078 and 10637 proteins in human and mouse proteome[37]. However, this method requires long MS run time and affects protein sequence coverage after two-thirds of peptides are excluded. The recent HiRIEF method has been extended to cover peptides in pH range 3-10, which identifies more peptides and proteins when used in combination with pH range 3.7-4.9[38]. **The HiRIEF method was used in all four studies (I - IV) to generate experimental MS data.**

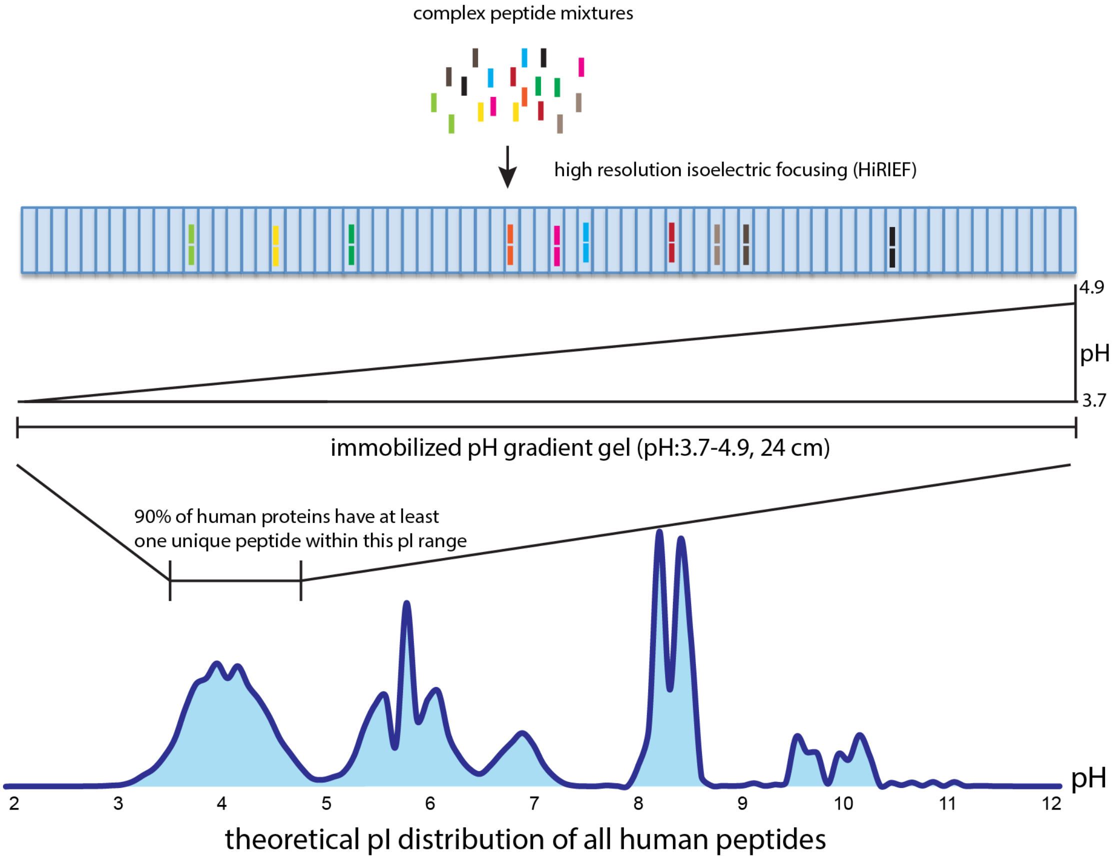
complex peptide mixtures
high resolution isoelectric focusing (HiRIEF)
4.9
pH
3.7
immobilized pH gradient gel (pH:3.7-4.9, 24 cm)
90% of human proteins have at least
one unique peptide within this pI range
pH
2
3
4
5
6
7
8
9
10
11
12
theoretical pI distribution of all human peptides

**Figure 3. Illustration of HiRIEF method for separating complex peptide mixtures with minimum loss of proteome coverage.**

To evaluate the analytical depth and dynamic range of proteome analysis, a simulation tool has been developed to assess the two global quality measurements and to guide better experimental design to enhance proteome analysis[39].

### Identify peptide sequences from mass spectra

Unlike genome sequencing or RNA-sequencing by which the nucleotide sequences are direct readouts from the instruments, mass spectrometer generates mass spectra which contains mass over charge (m/z) values of each product ions fragmented from a certain peptide, called peptides' mass "finger print". In early 1990s, it was still a difficult task to solve the amino acid sequences based on peptide mass spectra. Only experienced mass spectrometry experts were able to perform this analysis and it was very time-consuming. In 1994, Jimmy K. et al developed an algorithm, named SEQUEST, which correlates tandem mass spectra to amino acid sequences in a given reference protein database[40]. Since then, many software have been developed to automate this process, such as Mascot[41], TANDEM[42], RADARS[43], Probity[44], OMSSA[45], InSpecT[46], Andromeda[47], Morpheus[48], and MS-GF+[49-51]. Combination of multiple algorithms is likely to reduce false positive matches[52]. All of these peptide identification algorithms share a similar concept. A score is calculated for each peptide-spectrum match (PSM) to indicate the matching quality between experimental spectra and theoretical spectra of peptides in the search database. A mass spectrum is assigned with amino acid sequences of the peptide whose theoretical mass spectrum matches the best to the experimental mass spectrum. The peptide sequences are then used to infer which proteins are present in the samples. For mass spectra generated from co-fragmented peptides, it is difficult to revolve the amino acid sequences. To cope this problem, some tools have been developed to de-convolute the mass spectra of co-fragmented peptides and are possible to identify multiple peptides from one spectrum[53]. **SEQUEST was used in Paper I and Paper II, and MS-GF+ was used in Paper III and Paper IV for peptide identification.**

### Error rate in peptide and protein identifications

Scores of peptide-spectrum matches (PSMs) from database search only indicate how well the match between the observed and theoretical spectrum is, not telling if they are correct or incorrect matches. Through manual examination of PSMs with different scores, one can arbitrarily determine a score threshold to classify PSMs with equal or higher score than the threshold into correct matches and the remaining ones into incorrect matches. However, it is not trivial to find a threshold balancing well between sensitivity and specificity since search engines produce different scores and there is often an overlap of score distributions between correct and incorrect PSMs. Moreover, it poses a problem to compare results from different datasets that use different scoring functions[54]. In high throughput analysis of millions of spectra, it is important to accurately estimate error rate of peptide identifications from database search results.

Previously, a model of random spectra matching has been used to assess if a peptide-spectrum match scores higher than the scores expected from random matches[55]. Target decoy approach (TDA) provides an effective way to estimates false discovery rates in peptide identifications[56-58]. In this approach, a decoy database is created by reversing or shuffling protein sequences in the target database. The number of spectra matches to the decoy database is used to estimate the expected number of false positive matches. The false discovery rate is then calculated as the number of decoy hits divided by the number of target hits above a certain score threshold. The disadvantage of TDA is that error probabilities of individual PSMs are not known. A semi-supervised learning algorithm, Percolator, was developed to classify incorrect and correct PSMs that could be better distinguished by multiple scores and features instead of one single score threshold[59]. Percolator automatically generates a training set using top scored PSMs from the target database and PSMs from the decoy database to represent correct and incorrect matches respectively. A linear combination of several scores and features reported from peptide search engines is used to train a support vector machine classifier[59]. **Percolator was used in all four Papers (I-IV).**

### Protein inference

The final goal of proteomics study is to identify and quantify proteins. Due to the lost connection between proteins and peptides after digestion, it becomes difficult to correctly derive a list of proteins based on detected peptides, especially in the presence of "degenerate" peptides, which are shared by multiple proteins. This process is referred as protein inference[60-62]. One of the commonly used rules is the minimum set cover (Occam's Razor) which infers a minimum list of proteins that explains all peptide evidences. **Occam's Razor rule was used in all four Papers (I-IV) to infer the protein list.**

The proteins inferred from peptides or PSMs with 1% FDR do not inherit the same error rate[63]. Because false matches spread evenly among different decoy proteins while correct peptides or PSMs accumulate over proteins truly present. This leads to an inflated protein FDR compared to that at PSM or peptide level[64]. Protein level false discovery rate is usually calculated as the expected number of false protein identifications divided by the total number of protein identifications from the target database. Different methods have been developed for protein inference and estimation of the error rates in protein identifications, such as ProteinProphet[65], MAYU[64], IDPicker[66] and Fido[67]. ProteinProphet computes a protein's probability of being present in the samples on the basis of peptides' probability assuming unique peptides are independently identified[65]. The statistical model takes into account two factors associated with error rates: proteins identified with multiple peptides are more likely to be true than single-hit proteins; a minimum list of proteins is reported to cover all observed peptide evidence. MAYU[64] and Picked FDR[68] are target decoy based methods to calculate protein level FDR, which are scalable to large datasets with better accuracy than the classical target decoy approach. **Picked FDR approach was used in Paper III to calculate protein level FDR.**

## Quantitative MS proteomics

### Label free protein quantification

Spectra counting quantifies a protein according to the number of identified peptide-spectrum matches (PSMs) of this protein because more abundant peptides generate more fragment spectra and the number of PSMs is positively proportional to the protein amount[69]. Spectra counting was modified to integrate a normalization factor accounting for length of proteins, the number of observable peptides and expectation of observing peptides in an experiment[70-72]. Issues with spectra counting are statistical accuracy of comparing proteins with low spectra count and possible saturation effects of highly abundant proteins. An alternative method to spectra counting is to use peptide chromatographic peak area, referred as area under peak (AUC), which has been demonstrated to be linearly proportional to the concentration of measured peptides ($r^2$>0.99 in range of 10 fmol-100 pmol)[73]. Considerations of using AUC are the technical variations such as retention time, peak intensity and noise level when samples are analyzed in different MS runs. Another issue using label free method is the poor overlap of protein identifications when analyzing many samples in different MS runs. Some tools have been developed to tackle this issue, such as MaxLFQ[74], which aligns peptide MS1 features between runs and therefore creates better overlap of protein identifications when multiple samples are analyzed in different MS runs.

### Protein quantification by labelling

One of appealing advantages using labelling methods is that multiple samples can be analyzed in one MS experiment. Isotope-coded affinity tags (ICATs) was first used to label cysteine containing peptides[75]. Heavy ICAT reagent contains eight deuterium while the light form has no deuterium, creating a paired peptide peak with mass difference of 8 Da. Relative peptide abundance in two cell populations is calculated as the ratio of intensities of the paired peptides.

### Stable isotope labelling by amino acids in cell culture (SILAC)

SILAC utilizes the normal metabolic process by culturing mammalian cells in the medium containing essential amino acids labelled by stable isotopes[76]. Two cell populations are cultured in two separate mediums, light medium containing normal amino acids and heavy medium containing isotope labelled amino acids with $^{2}$H instead of H, $^{13}$C instead of $^{12}$C, or $^{15}$N instead of $^{14}$N, which create a known mass shift to the light peptides. SILAC is shown to have no effect on cell morphology or growth rates[77, 78]. The ratio of peak intensities of paired heavy and light peptides indicates the relative protein abundance in two cell populations. SILAC method is limited to analyze living cells. Spike in SILAC samples as a standard can be applied to analyze tissue samples[79]. When analyzing complex tumor samples, one SILAC labelled cell line may not be representative. Super SILAC method uses several SILAC labelled cell lines together to generate more accurate protein quantification[80].

### Isobaric labelling by iTRAQ and TMT

Stable isotope labelling methods were limited to analyze only two samples at the beginning. Its multiplexing capacity was increased later[81-83]. With SILAC methods, deuterium labelled peptides are slightly separated in chromatography compared to normal peptides. Moreover, the use of isotope to create paired peptide peaks with certain mass differences increases the complexity of MS1 spectra.

In contrast, iTRAQ uses isobaric reagents that always add a constant mass to peptides[84, 85]. The isobaric tag consists of three parts: a peptide reactive group linked to peptide's N-terminal amine or amine at lysine, a mass balance group and a reporter group. The total mass of reporter and balance group is constant while having different isotopes enrichment of $^{13}$C, $^{15}$N, and $^{18}$O atoms. Since different isobaric tags add the same mass shift, peptides from different samples labelled with iTRAQ are indistinguishable in MS1. When peptides are fragmented in MS2, reporter ions of different isobaric tags dissociate from the peptides and appear in low mass region ranging from m/z 114.1 to 117.1[84]. Relative peptide abundance in different samples is calculated as the ratio of intensities of the reporter ions. iTRAQ was originally designed as 4-plex, now it has been extended to 8-plex[85]. **In Paper I, iTRAQ 8-plex was used for quantitative analysis.**

Same as iTRAQ labelling, tandem mass tag (TMT) labelled peptides are also indistinguishable in MS1. When it was first presented, only comparison of two samples were allowed[86]. However, the multiplexing capacity was extended to 6-plex with newly designed reporter and balance group[87]. The TMT reagents now allow multiplexing 11 samples together[88]. **TMT 10-plex was used in paper III and IV for quantitative analysis.**

## Statistical analysis of quantitative MS proteomics

The distinctive structure and properties of MS proteomics data has differed its statistical analysis from other high throughput experiments such as RNA-seq and microarray, which are alternative methods for gene expression profiling. MS proteomics data has a hierarchical structure in which PSMs are repeated measurement of peptides and multiple peptides are detected for one protein. Moreover, peptides from one protein differs in signal intensity depending on ionization properties, digestion efficiency and other stochastic events.

Statistical methods in quantitative MS proteomics has been mostly solving two major problems: protein abundance summarization and significance test to assess differential protein expression.

### Methods for protein abundance summarization

Although the quantification can be done at peptide level[89, 90], it has several drawbacks. The biological functions are carried out by proteins and not by tryptic peptides. Peptide level quantification is less accurate than protein level quantification which combines multiple peptides. Moreover, peptide level statistical test is prone to generate more false positives[91].

There are many different methods to combine peptide intensities into protein abundance estimate, including averaging, median sweeping[92, 93], weighted mean[94], total least squares[95], sum of intensities[96, 97], principal component analysis (PCA)[98], factor analysis[99, 100], linear model and mixed-effects model[101, 102]. However, protein inference process creates a potential problem in protein level quantification, resulting a mix of quantitative signals of peptides derived from different proteoforms or post-translational modifications. Methods like PQPQ[103] and Diffacto[100] can be used to curate outlier peptides with incoherent quantitative signals.

These methods have been tested and compared in different datasets[92, 96, 104]. It is not trivial to choose among these methods as they are demonstrated in different types of data. Understanding the underlying assumptions of different methods and examining the data distribution can indicate which methods to apply.

### Statistical methods to assess differential protein expression

Traditional statistical methods such as t-test and ANOVA can be used to assess differential expression when protein abundance is estimated. Although t-test has several limitations such as applicable to only two classes comparison and lack of statistical power for high dimension data, it is still being used in recent proteomics studies[105-107], mainly because it is easy to implement. A moderated t-test, previously developed in Limma R package[108] to analyze microarray data has been recently suggested to analyze proteomics data[93, 109, 110]. Limma applies Bayesian approach to increase statistical power for experiments with small sample size by shrinking variance toward a pooled estimate. According to D'Angelo et al evaluation, Limma is preferred considering both sensitivity and specificity compared to general linear model (GLM) and mixed models in a TMT 10plex labelled proteomics dataset[93]. For label-free data sets which often contain many missing values, there are methods to impute the values for both missing not at random (MNAR) and missing at random (MAR)[111, 112]. A specific test has also been developed for label-free data to detect differentially expressed proteins without imputing values[113]. However, previous statistical methods have not taken into account the impact of the number of peptides on quantification accuracy which has been shown in independent studies[94, 96]. **Paper IV present a tool, DEqMS, which inherits the advantage of Limma and provides an additional feature to account for the impact of the number of peptides on quantification when statistically assessing differential protein expression.**

## Proteogenomics – bridging between genomics and proteomics

With the development of MS instruments together with quantification methods and bioinformatics, MS-based proteomics has become a more mature technique for high throughput quantitative proteome analysis. The published first draft map of human proteome detected protein products of about 17,000 (84%) human protein-coding genes[114, 115], marking a milestone in MS proteomics study. In the meanwhile, the next generation sequencing technology has become a widely accessible techniques to perform whole genome, exome and transcriptome analysis[116].

By integrating genomics and proteomics, proteogenomics has emerged as a powerful approach to study the interplay of genome and proteome, such as how chromosomal aberration and genetic mutations impact protein level regulation[37, 117-126]. In comparison to standard proteomics, which focuses only on canonical proteins in a reference protein database, the primary goal of proteogenomics is to detect *unusual* peptides from unannotated or non-coding loci, novel splicing, non-synonymous SNPs and mutations, chromosomal rearrangements and other aberrations in the genome. Such integrated proteogenomics analysis has been applied to characterize proteomic variation in colorectal cancer[127], ovarian cancer[128] and breast cancer[129, 130].

### Proteogenomics database

Proteogenomics workflow consists of three major steps: database construction, peptide search, and curation of the results. Database construction is the first and vital step which determines the types of *unusual* peptides to be detected. It is important to clearly define them before conducting the search.

To discover peptides from unannotated or non-coding regions in genome, hypothetical peptides translated from genome sequences can be included in the search space. Six-frame translation search is an unbiased approach to discover coding regions in the genome, particularly useful in species with compacted genome (high gene density). **This approach was used in Paper II to discover coding sequences in a newly sequenced commensal yeast genome.** However, it is challenging to apply it in large genomes with majority of sequences being non-coding. For example, protein coding regions in the human genome consist a minor fraction, 2%-3%, and inclusion of hypothetical peptides from non-coding regions increases the search space several hundred times. It is challenging for peptide identification algorithms to find correct peptide-spectrum matches in a large database[119, 131]. The effect of increased database size on the reliability of peptide identifications is evaluated[132, 133]. Database reduction for species with large genome size such as human can increase the sensitivity of six-frame translation search. Peptide isoelectric focusing coupled LC-MS is a feasible method to reduce the peptide search database[33, 37]. Peptide spectra can be searched in a database restricted by the corresponding pI ranges in which the peptides are collected from the IPG strip. Therefore, the sensitivity and specificity of peptide identifications is increased[33, 37]. **Peptide isoelectric point restricted database search was applied in Paper III to discover coding sequences in human genome.**

Proteogenomics has been applied to refine genome annotation in many organisms such as bacteria[134, 135], yeast[136], *Drosophila melanogaster*[137], *Arabidopsis thaliana*[138, 139], mouse[140] and human[37, 141, 142]. RNA sequencing data is another commonly used resource to construct databases for proteogenomics studies, moreover it provides a sample specific peptide database from transcribed genes including splice junction sequences and missense variants[143-145]. Several tools have been developed specifically to facilitate peptide database construction from RNA-sequencing data[146, 147].

To detect variant peptides from cancer specific mutations, the search database can be generated from known mutations database such as COSMIC or detected variants from sequencing data. However, identification of variant peptides is an error prone process and suffers high risk of false positives. **Paper III contributes a tool, SpectrumAI, to reduce false discoveries by automatically inspecting the experimental spectra of the variant peptides for supporting**

**peaks of the substituted amino acids.** Whole-genome or whole exome sequencing from tumors and matched normal tissues can be used to construct a data supported variant database to increase the reliability of detecting tumor specific mutations by MS proteomics techniques. Many different bioinformatics tools have recently been developed for proteogenomics[148-156].

### False discovery rate of proteogenomics findings

Estimation of false discovery rate (FDR) of proteogenomics findings is not trivial since the expected findings only consist of a minor fraction compared to peptide identifications from known proteins[157]. The false discovery rate needs to be estimated separately for *unusual* and known peptides, termed as novel only or class specific FDR[37, 119]. To calculate novel only FDR, target and decoy hits in known and novel search space are counted separately to have accurate estimate for the number of correct and incorrect matches of novel peptides. In practice, prior to FDR calculation, it is necessary to remove novel peptide hits resulted from the artefact changes of amino acids such as deamidation (asparagine to aspartic acid, glutamine to glutamic acid) and changes between isoleucine and leucine, to avoid underestimating the FDR by populating the novel peptides with spurious hits.

Class specific FDR should be used as first step to reduce false novel peptide identifications. However, it is not sufficient to control the error rate. Because target decoy FDR calculation is heavily weighted on match scores reported by peptide search engines, incorrect peptides with partial sequences matched to theoretical spectra can still receive high scores and thus can pass certain FDR cutoff. These incorrect matches could arise from unexpected modified known peptides that possess the same molecular mass to the novel peptides.

### Curation of proteogenomics findings

In the task of discovering novel peptides, the step after FDR calculation is to curate the candidates against a known protein database. Novel peptides can refer to peptides absent in reference protein databases, or peptides from supposedly non-coding gene or unannotated regions in the genome. Reference proteome databases such as Ensmebl and Uniprot contain predicted protein sequences from non-coding genes, of which no protein level evidence has been observed previously. A clear definition in the beginning will avoid incorrect tagging of known or novel peptides in subsequent filtering. In curation step, peptide variants derived from non-synonymous SNPs should not be considered as novel peptides. Additional evidence in orthogonal data, such as conservation, RNA-seq and ribosomal profiling data can be used to reinforce the confidence level of proteogenomics findings[158, 159].

## AIMS AND OBJETIVES OF THE PRESENT THESIS

MS based proteomics provides an opportunity to advance our understanding of cancer biology by studying protein expression and regulation. Specifically, this thesis focuses on the data analysis and applications of MS based proteomics to generate valid biological information about the cellular and tissue proteome by using existing tools and further developing novel bioinformatics methods.

**Specific objectives:**

**Paper I:**

Develop a bioinformatics tool to analyze splice variants at protein level using MS proteomics data.

**Paper II:**

Develop an integrative genome annotation pipeline combining proteomics and transcriptomics data and compare it with the conventional annotation approach.

**Paper III:**

Develop a proteogenomics pipeline to discover coding regions in the human genome and variant peptides derived from mutations and nsSNPs using MS proteomics data.

**Paper IV:**

Develop a robust statistical method to assess differential protein expression for both labelled and label-free proteomics data.

# RESULTS AND DISCUSSION

## Paper I - Splice variant analysis in MS proteomics

1. Based on in silico analysis of human peptidome, about 20 percent of splice variants are theoretically possible to be identified with unique peptides using trypsin or lysC digestion.
2. Among the 7000 genes identified in A431 cell lines, about 600 genes have known splice variants identified with unique peptides evidence.
3. Discrepancy was observed in comparison of gene centric, protein centric and splice variant centric quantitative analysis. Several genes were shown to have differential splice variants regulation in A431 cells after treated with Gefitinib.

In MS-based proteomics, peptides are aggregated into proteins through the protein inference process which applies Occam's Razor rule to report the minimum number of protein variants that are sufficient to explain all peptide evidences. Thereafter, the abundance of protein is estimated by averaging over the peptides (or by other weighted mean methods). In protein or gene centric analysis, there is a possibility that only one protein isoform is reported by the protein inference process, whereas in fact another shorter protein isoform is co-expressed. Depending on which protein isoform is more abundant, the summarized signal for protein and gene resembles the dominant protein isoform. When we developed and used our previous tool PQPQ[103] to curate quantitative proteomics data, we observed that a few high-quality peptides showed altered quantitative pattern compared to the bulk of peptides. And we hypothesized that these could be derived from splice variants, hence we explored this further by developing SpliceVista. With SpliceVista, it is possible to investigate if the group of peptides causing discrepant signal to gene centric analysis are derived from a specific splice variant undergoing differential regulation.

In this study, certain splice variants showed different quantification results to protein centric and gene centric analysis. An example is shown in Figure 4 where Protein NM_022170 was reported based on peptide evidences. However, three peptides in cluster 1 showed a different quantitative response after the treatment compared to the remaining peptides. Inspecting their positions in transcripts, these peptides map to the unique region from NM_022170 variant. The evidence indicated another protein isoform present in addition to NM_022170, and the other hidden protein isoform was more abundant and remained unchanged after the treatment.

The recent work by Wang et al corroborated that detection of protein isoforms by MS-based proteomics is limited by trypsin cleavage specificity[160]. In addition, their conservation analysis showed that the trypsin cleavage sites, lysine and arginine, are more frequently observed in exon-ending and exon-exon junctions due to the evolutionally conserved nucleotide usage at splicing sites. As a consequence, it limits the detection of splicing junction peptides using trypsin as digestion enzyme. Using other enzymes, such as chymotrypsin, it is possible to increase detection rate of splice junction peptides[160].

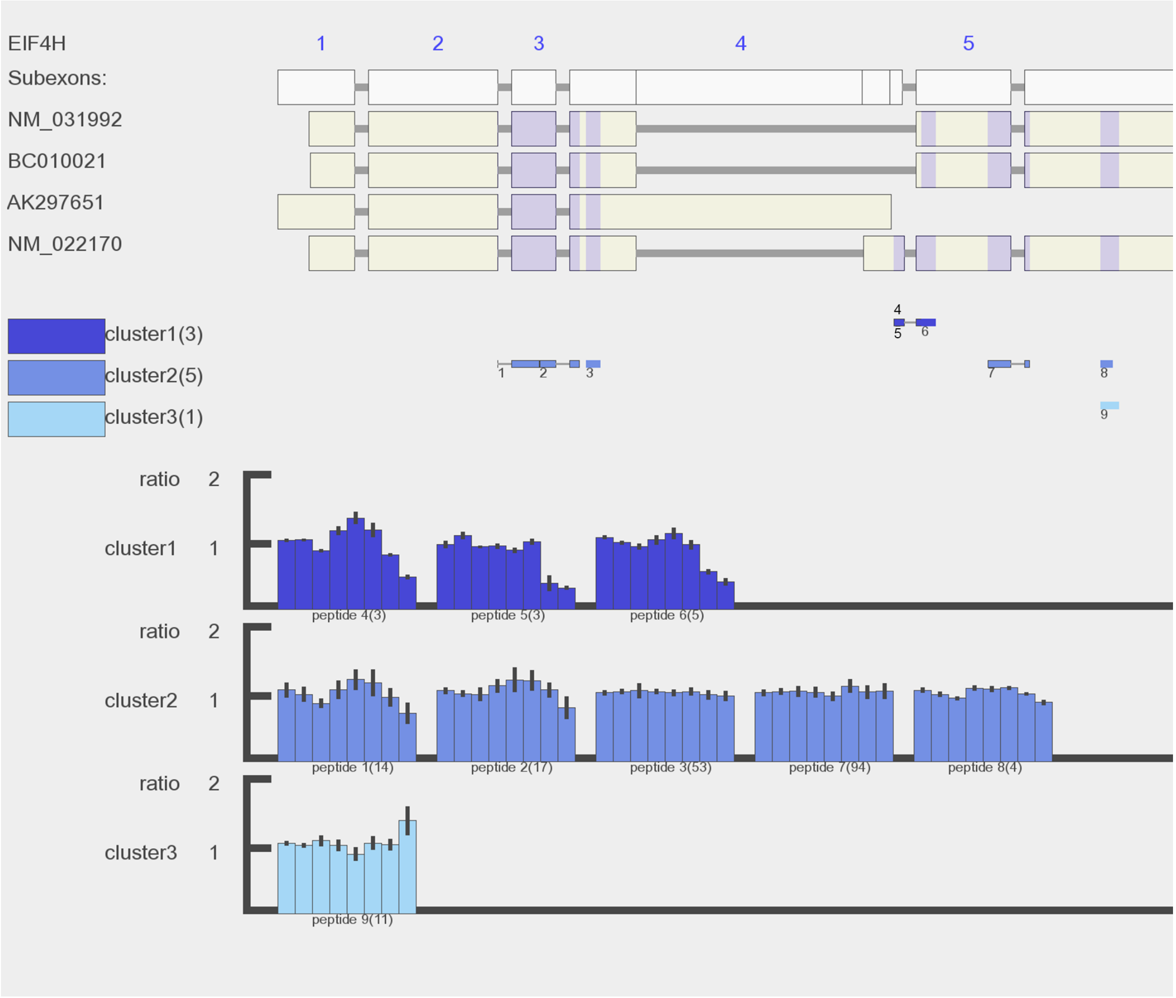

**Figure 4.** An example of detected gene EIF4H with incoherent peptide quantitative patterns in A431 cell line after Gefitinib treatment. Gene EIF4H has six exons and four known splice variants (exon 6 is cut out to enable better resolution). Nine unique peptides were identified for EIF4H and grouped in 3 clusters. The eight bars represent peptide relative abundance in the eight samples at control, 2h, 6h, and 24h after Gefitinib treatment, two biological replicates at each condition. Cluster 1 (dark blue) which includes peptide 4, 5 and 6 has distinct pattern showing down-regulation at 24h after drug treatment. Peptide 4 (DDFNSGFR) and peptide 5 (DDFNSGFRDDFLGGR) are uniquely mapped to splice variant NM_022170. Peptide 6 (DDFLGGR) is not unique to splice variant NM_022170, but it is very likely that this peptide was a digested product from NM_022170 based on its similar quantitative pattern to peptide 4 and 5. In the mid panel, the number in the bracket after each cluster is the number of unique peptides grouped in this cluster. In the bottom panel, the number in the bracket after each peptide is number of PSMs.

**Paper II – Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation.**

Current genome annotation strategies utilize mostly protein homologs from related species and transcriptomics data from RNA-sequencing. Recently, proteomic data have been primarily used to confirm gene models and discover missing genes after annotation obtained by RNA-seq or homology based means, and thus not as part of an integrative workflow combining proteomics and RNA-seq for genome-wide annotation (shown in Figure 5).

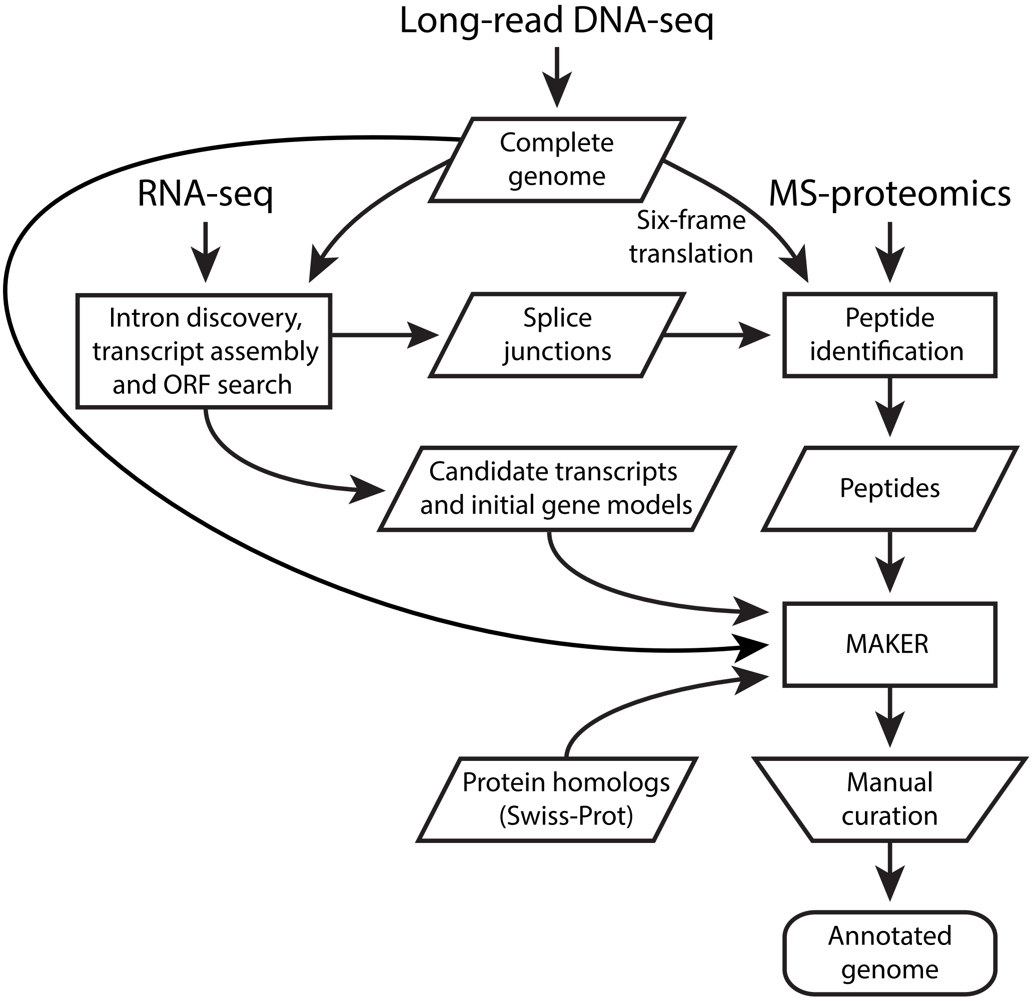

**Figure 5. Integrative genome annotation workflow**. Data from four different sources (long-read DNA sequencing, RNA-seq, MS-based proteomics and Swiss-Prot reviewed proteins) were integrated using an evidence-based genome annotation framework (MAKER)[161]. Transcripts were assembled from RNA-seq reads using Trinity[162] and PASA[163] was used to identify likely protein-coding regions to provide gene models for initial gene predictions. Three *ab initio* gene predictors (GeneMark-ES[164], Augustus[165] and SNAP[166]) were included in MAKER. Augustus and SNAP were iteratively trained based on MAKER-generated gene models. The computationally inferred gene structures were manually curated. Shapes are used according to workflow figure standards (rectangles show processes, data are in parallelograms, the trapezoid indicates a manual step and the rounded rectangle represents output).

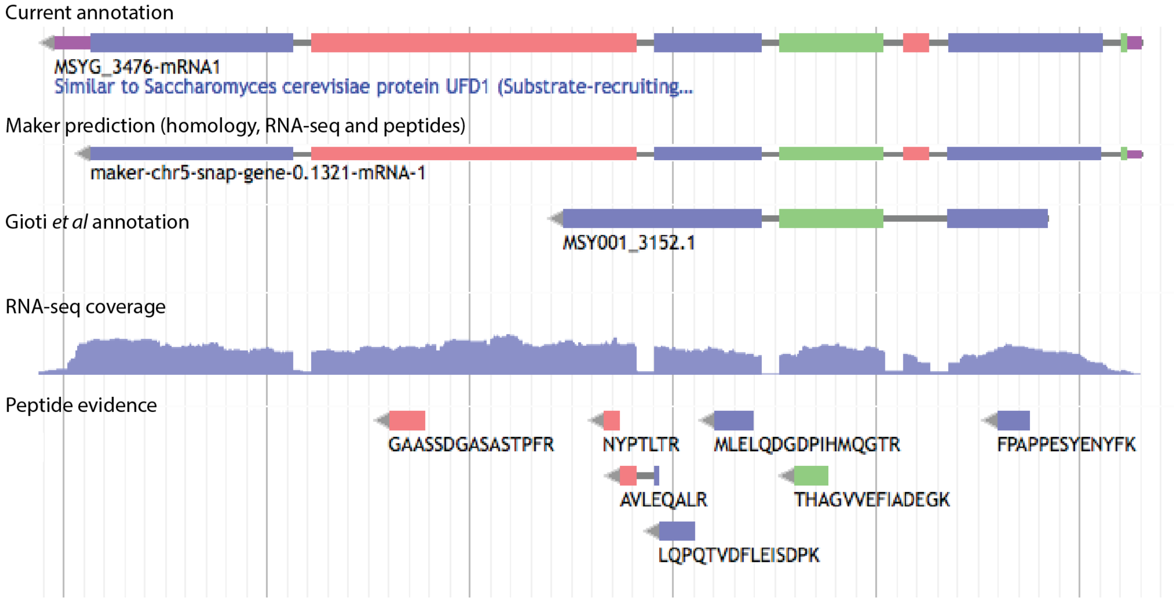

**Figure 6. Gene annotation facilitated by RNA-seq and peptide evidence.** Screenshot from the WebApollo[167] genome annotation editor showing a locus where RNA-seq and peptide evidence improved gene annotation compared to the previous annotation described by Gioti *et al*[168]. The 5´-UTR and protein-coding segments were identified by the MAKER-based pipeline integrating RNA-seq and peptide data. Manual curation added a 3´-UTR (uppermost track). The colors of exons and peptides indicate reading frame, such that exons and peptides with the same color are in the same reading frame. UTRs are indicated in purple and introns in gray. RNA-seq coverage is shown for the genomic minus strand (i.e. the strand of the annotated gene) and indicates the number of read pairs at each base.

The major conclusions from this study are:

1. Integration of proteomics data in gene annotation pipeline increases sensitivity of protein-coding sequences detection as compared to using homology and RNA-seq data alone. RNA-seq data is particularly useful for intron detection. Figure 6 shows an example of one annotated gene using the integrative workflow (top two tracks) and the traditional approach with RNA-seq data and homology (third track from top).
2. In addition, proteomics data increases accuracy of protein coding gene annotation, evaluated by conserved protein domain analysis.
3. Proteomics data provides additional level of confidence for annotated genes and is able to assess the completeness of gene annotation.

## Paper III – an integrative proteogenomics workflow to discover protein-coding region in human genome and variant peptides derived from mutations

Proteogenomics analysis enables the discovery of protein coding regions and sequence variants but their verification remains challenging. This study presents an integrated proteogenomics analysis workflow that combines peptide discovery, curation and validation, robustly identifying unknown coding regions and mutations using MS data.

1. Detected peptide evidences suggest translation of pseudogenes, lncRNAs, short ORFs, alternative ORFs, N-terminal extensions and intronic sequences in human genome.
2. Certain pseudogenes and lncRNAs are translated in tissue specific manner.
3. Class specific FDR is necessary but not sufficient to control error rates in proteogenomics findings.
4. SpectrumAI which verifies peptide-spectrum matches through MS2 spectra significantly reduces false identifications of variant peptides with single amino acid substitution.
5. External evidence in orthogonal data such as conservation, RNA-seq and ribosomal profiling data increases the confidence of detected protein coding regions.

Verification of detected protein-coding region is challenging in proteogenomics studies. Previous studies have been either too stringent, such as demanding minimum two unique peptides supporting a protein coding loci or requiring the novel peptides with at least two mismatches to known peptides. Although the criteria can remove some false discoveries, it probably misses many potential true findings at the same time. In our A431 cell line data where 374 protein coding loci were discovered, only 42 loci were supported with two and more peptides (Figure 7). Instead of discarding all protein coding loci with only one unique peptide support, we searched external evidence such as detection in independent datasets, conservation analysis, RNA-seq, and ribosomal profiling. These additional levels of evidence could be used to filter and prioritize the candidate novel peptides.

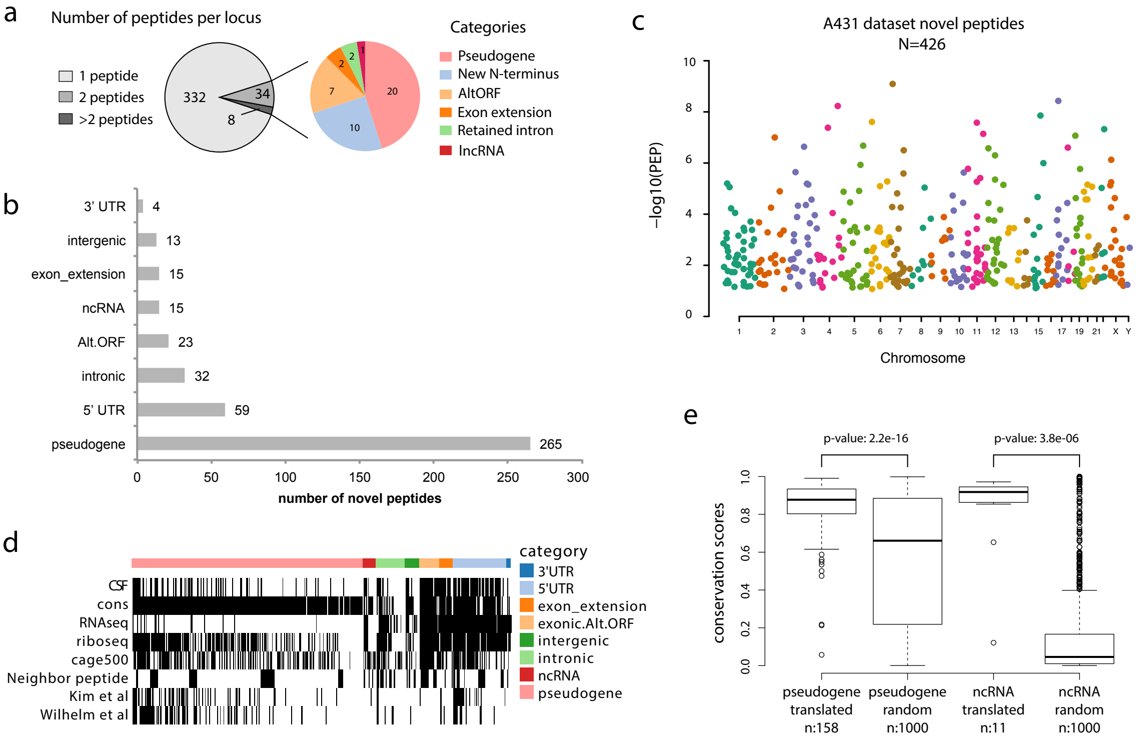

**Figure 7. Unannotated protein-coding loci found in the A431 cells dataset. a**) The left pie chart shows the number of unannotated protein-coding loci supported by one, two or more peptides (peptides within 10kb distance were grouped into one locus); the right pie chart shows the different types of unannotated coding events supported by multiple peptides. **b**) Automatic categorization of novel peptides by Annovar[169] using RefSeq gene annotation. **c**) Manhattan plot of novel peptides, where the y-axis represents the peptide's posterior error probability (PEP). **d**) Orthogonal data support for novel peptides, including PhyloCSF[170] coding potential, conservation analysis, A431 cell line RNA-seq reads evidence, ribosome profiling[171, 172], CAGE[173] (up to 500 bp upstream from peptide location), presence of neighboring peptides (within 10kb), and whether the peptide was identified in the draft proteome data of Kim *et al*[114] and Wilhelm *et al*[115]. Continuous variables were discretized to binary values 0 or 1 for visualization purposes. 10,000 random genomic loci were used to determine the threshold to call if Ribo-seq or CAGE data were supportive or not. **e**) The conservation score (PhastCons[174] score) distribution of pseudogenes and lncRNAs for which peptides were found was compared to that of 1000 randomly selected pseudogenes and lncRNAs. In the box plots, center line corresponds to median, box boundaries correspond to the first and third quartiles (Q1 and Q3), the upper whisker is min(max(x), Q3+1.5*IQR) and lower whisker is max(min(x), Q1-1.5*IQR).

Identification of single amino acid variant peptides is an error-prone process with high risk of false positives. Previously, manual curation has been used to ensure the validity of matches, but it is a laborious and time-consuming process. SpectrumAI automates the process of verifying the experimental peptide spectra for the presence of MS2 ions supporting the substituted amino acid. Being tested in cell line data with whole genome sequencing and RNA-seq data support, SpectrumAI is proved to be robust eliminating incorrect identification of single amino acid variants (Figure 8). It is worth to note that variant peptides passing SpectrumAI are not necessarily sequence variants before ruling out the artefact amino acid changes occurred in sample preparation.

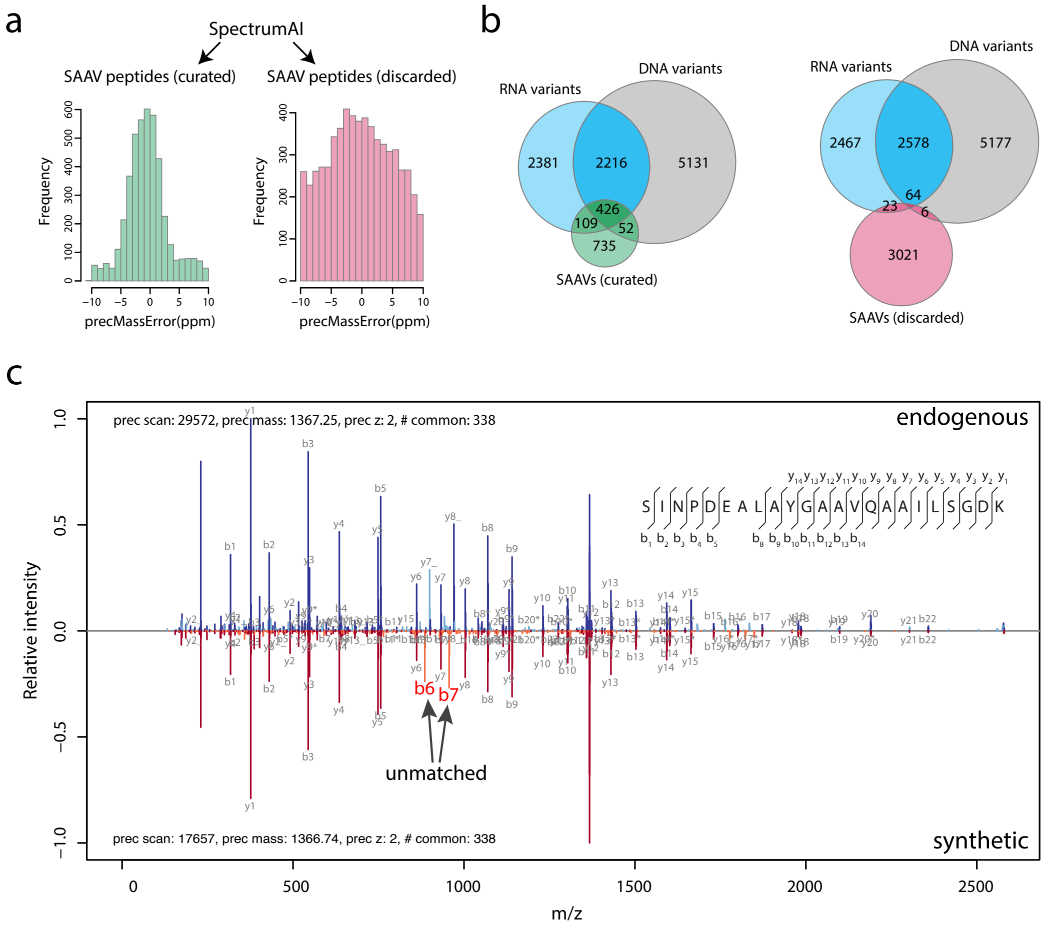

**Figure 8. SpectrumAI increases identification accuracy of peptides with single amino acid changes.** **a**) Precursor mass error distributions of peptides classified as curated and discarded by SpectrumAI. **b**) Curated SAAV peptides have more overlap with missense variants identified at DNA and RNA level. **c**) Mirror plot of an incorrectly identified peptide (that yet had passed discovery stage with class-specific FDR 1%) with a single residue substitution (V>L, at position 8) that was subsequently discarded by SpectrumAI. Annotated MS2 spectrum of the endogenous peptide is shown on top whereas that of the respective synthetic peptide is inverted and shown on bottom. This incorrect peptide identification detected by SpectrumAI shows mismatching b6 and b7 product ions (highlighted in the synthetic side and missing in the endogenous side) which ought to have flanked the substituted residue, indicating that the endogenous amino acid sequence is incorrect between its sixth and eighth residues.

Pseudogenes constitute the major categories which contained more than half of the detected novel peptides. Some of them, e.g. *TATDN2P1* and *UBE2L5P,* were shown tissue specific expression. In comparison to previous effort using RNA-seq data to detect pseudogene expression which claimed expression of thousands of pseudogenes at transcript level[175], our proteomics data suggest a relatively smaller scale of pseudogene expression at protein level. The reasons can be that some transcribed pseudogenes execute their functions as RNA transcripts, e.g. acting as decoys for miRNA or some have undergone translational control[176].

### Paper IV – DEqMS is a robust statistical tool to assess differential protein expression for both labelled and label-free proteomics data

Currently, various statistical methods are being used for quantitative proteome analysis. The lack of consensus is due to several reasons. Some methods are limited for specific experimental design with only two classes or limited to specific data type (labelled or label-free) and difficult to choose. In addition, previous methods neglect the impact of the number of peptides on quantification accuracy, shown in Figure 9. DEqMS is a robust and universal tool to statistically assess differential protein expression by taking into account the number of PSMs or peptides used for quantification.

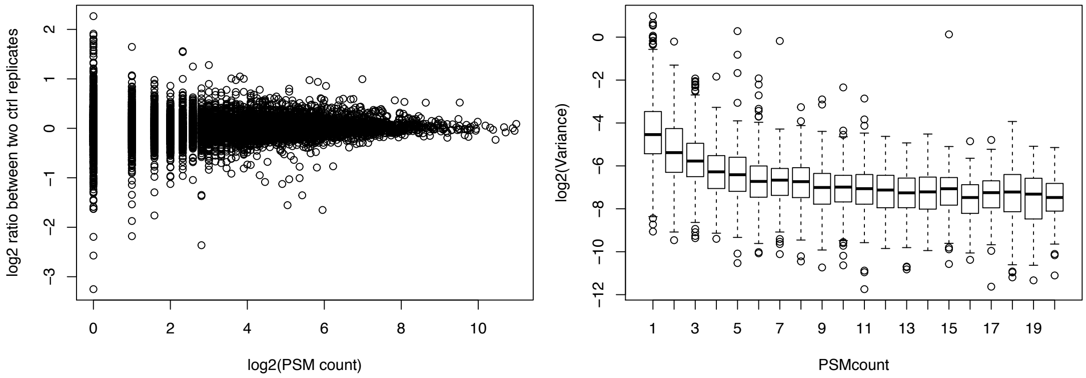

**Figure 9. Protein variance in relation to the number of PSMs in TMT labelled dataset.** A) each dot represents one protein. The log2 ratio is calculated using median sweeping method as previously described[93]. x-axis is the number of PSM count in log2 scale. B) y-axis is the pooled variance calculated using all 10 samples. The boxplot showed the median of protein variance with number of PSMs identified from 1 to 20.

DEqMS has several advantages over existing methods:

1. Unlike previous methods neglecting variance heterogeneity in proteomics data, DEqMS takes into account the number of peptides used for quantification in variance estimation.
2. DEqMS has higher accuracy of detecting differentially expressed proteins compared to existing methods, demonstrated in independent benchmark datasets (Figure 10).
3. DEqMS is a universal tool to perform differential protein expression analysis with MS proteomics data, applicable to both labelled and label-free datasets.
4. DEqMS is flexible to allow various experimental designs.

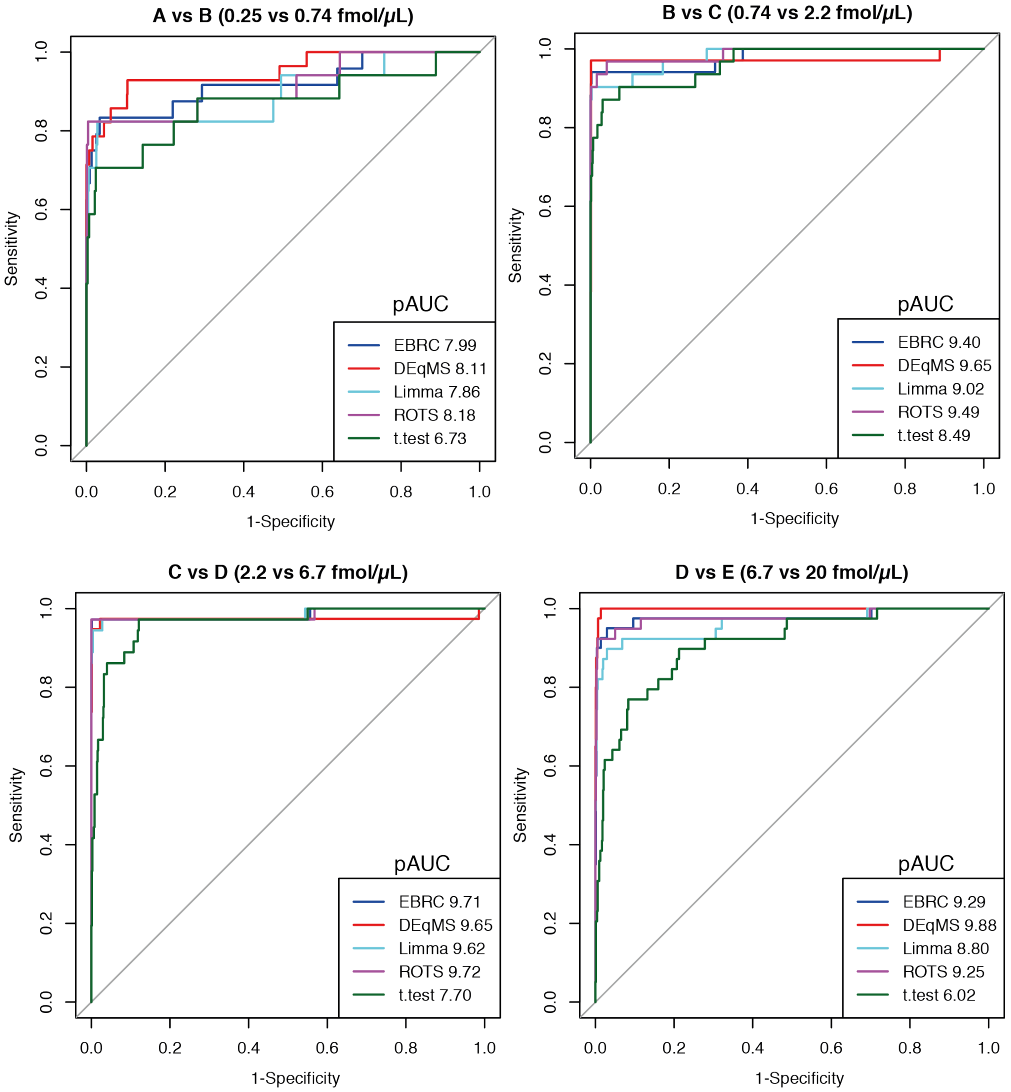

**Figure 10. Receiver operating characteristics (ROCs) on CPTAC study 6 data**[177]. In this dataset, 48 UPS1 proteins were spiked into yeast proteome at five different concentrations (subsequent dilution at the factor of 3): 0.25, 0.74, 2.2, 6.7 and 20 fmol/μL, labelled as A to E, with three replicates at each concentration. Partial area under the curves (pAUC) of each model between false positive rates 0 and 0.1 are shown in the bottom-right legend. pAUC values were multiplied 100 times, the perfect model will have a value 10.

# CONCLUDING REMARKS AND FUTURE PERSPECTIVES

## Concluding remarks

The presented thesis has contributed methods and tools for various analysis and applications of MS-based proteomics in biological research.

**Paper I** addressed an issue in protein quantification after the protein inference process. It described a tool to de-convolute potentially mixed quantitative signals of peptides derived from different splice variants that are incorrectly inferred. The present tool allows splice variant centric quantification and enables investigation of splice variant regulation using MS proteomics data. **Paper II** and **III** demonstrated applications of MS data in proteogenomics studies. In **Paper II**, MS data was used in combination with RNA-seq data to annotate a newly sequenced yeast genome. Such integrative approach demonstrated higher accuracy than the conventional method without using MS data. **Paper III** extended the application to a more challenging case, aiming for discovery of coding sequencing in the human genome. An automated pipeline has been developed to curate the findings and search their evidences in multiple sources of orthogonal data. Detection of differentially expressed genes with statistical significance is an important task in biomarker research. **Paper IV** contributed a tool for statistical analysis of quantitative MS proteomics. Tested in several different datasets, the tool demonstrated higher accuracy and broader applicability compared to existing methods.

The applications of MS-based proteomics and proteogenomics have great potential to discover new biomarkers and therapeutic targets in cancers [118, 124, 126, 130, 178-184]. The methodologies developed in this thesis have contributed bioinformatics tools to facilitate the future applications.

## Future perspectives

### The identities of a large fraction of MS2 spectra are still unknown

Recent proteomics studies have showed detection of ten thousand proteins in one MS experiment, but this only corresponds to 30%-40% of generated MS2 spectra. The identities of the remaining MS2 spectra are still unknown. Different computational methods have been applied to find matches for the unexplained spectra, using for example sequence tagging[185-187], *de novo* sequencing[188-194], a hybrid of *de novo* sequencing and database search[195], correlating or clustering spectra to known peptides[196, 197], and error-tolerant search approaches[185, 198]. It is postulated that the majority of the remaining unknown spectra are from peptides with unexpected post translational modifications (PTMs) and sequence variants. PTMs plays an important role in signal transduction pathways and regulating cellular functions. Currently, enrichment is required for selected PTMs analysis, such as phosphorylation, to be identified on proteome-wide scale[199-206]. It is still yet not possible to perform unbiased detection of unmodified peptides and PTMs simultaneously due to the lack of tools for proteome-wide analysis of multiple PTMs.

With the advances of computing capacity and machine learning algorithms, *de novo* sequencing is gaining interest since it has the possibility to identify new peptides and PTMs without the need for a database. A recent publication by Tran *et al* used deep learning for peptide *de novo* sequencing and demonstrated considerable improvement over previous *de novo* methods[194].

Multi-dimension peptide pre-fractionation methods are necessary to improve proteome coverage. However, to reach complete coverage of the proteome including PTMs, new separation and enrichment methods or instrumentation are needed to increase sensitivity for PTMs detection.

### Multiplexing capacity of MS-based proteomics technique is still limited

Current isobaric labelling reagents allow multiplexing up to 11 samples to be analyzed at the same time. In order to analyze large clinical cohort of hundreds or thousands of samples, samples need to be split into separated MS experiments, with the risk of getting fewer and fewer overlapping proteins when combining data from multiple MS experiments. Moreover, the time required for sample preparation and MS analysis to reach in-depth proteome analysis is still considerably long, ranging from a few days to a week for one MS experiment. It is therefore in need of new quantification method or new types of instruments with increased multiplexing capacity and throughput so that MS proteomics techniques are applied in large clinical studies.

### Standardize formats to facilitate proteomics data integration with other omics data

A standard and compatible format is the foundation to facilitate genome-wide integrative studies combining multi-omics data, such as genomics, transcriptomics and proteomics. In response to The Human Proteome Organization (HUPO) Proteomics Standards Initiative, two novel standard formats, proBAM and proBED, have been developed to store proteomics data in genomic context to facilitate proteogenomics studies[207, 208]. These two formats inherit features from two widely used genomics data format, BAM and BED, to store PSMs in genomic context. Many previously genomics tool such as SAMtools[209] and BEDTools[210], and new tools like proBAMconvert[211] are available for file indexing and conversion. In the future, proteogenomics studies will embrace the new data formats to increase data sharing and reanalysis.

# ACHKNOWLEDGEMENTS

Throughout the journey of my PhD study, I have so many people to thank. First of all, I would like to thank my main supervisor, Janne Lehtiö, for selecting me out of many applicants to work on these projects, for supporting me both with his professional advices and in finance, for inspiring and encouraging me in all projects, for trusting me in research, for being a humble person that I can discuss scientific ideas freely and fearlessly with. Secondly, I also want to thank my co-supervisor, Jenny Forshed, for supporting and guiding me in my PhD, for being like a friend and for giving me considerate advice both in career and life.

I would like to express my gratitude to Rui for teaching me tons of mass spectrometry knowledge with always elaborated answers, for bearing all my naïve questions with his humbleness and patience all the time; to Henrik, for mentoring and encouraging me while working together in the breast cancer project and being always gentle in all kinds of discussions; to Lukas for mentoring me in research, inspiring me with all his challenging and sharp ideas in scientific discussion. I also want to thank Helena, for not only being supportive in all administrative work, but also active discussions and exchange of ideas in all projects.

Many thanks to the other lab members who have jointly created an open and friendly work environment, to talk about research, to have lunch together, to FIKA together and speak jokes together. Special thanks to Hillevi and Ghazaleh (our guardian angles: P), for managing the lab and arrangement of group activities; to Yan for accompanying me in the PhD study and being a good friend to talk about everything, and Elena, for studying and hanging out together in and outside the lab. To Jorrit, for teaching me programing tricks and sharing your Dutch jokes. To Maria, for scientific discussion in all projects and also for organizing the climbing activities sharing fun with us. To Georgios, for teaching and sharing with me the fun of lab work, To Rozbeh and Jürgen, for exchanging ideas in CETSA and Click-chemistry and being humorous scientist around us. To Nate, for all the discussion we had and your considerate tips in choosing postdoc labs. To Taner and Ioannis, for joining me in the journey as bioinformatician, and special thanks to Ioannis for the cover image, well done! To Husen, thanks for reviewing my thesis text, and it was very helpful!

And thanks to our previous lab members Alex and Eric, for bringing “shiny” things to our group and my knowledge; to Lina, for helping me in the SpliceVista project; to Davide, for teaching me mass spec knowledge, for running and playing KI football together (salute to our *Speedvac* team in KI football cup). I also want to thank Pär Engström and Annika Scheynius, for mentoring me while we were working together for the Malassezia project.

In addition to all mentioned above, I want to tell everyone in our group that this PhD journey is one of my greatest joy in science and life.

Last, I want to thank my friends from ShuGuang Chinese Christian fellowship, for all the days we have spent together, studying bible and going worship together, eating together, go hiking and skiing, playing guitars and singing beautiful songs. To Jingwen, for all the interesting discussion we had both in science and religion, and for all the help while I was preparing for my thesis. All of these have become my memories of happiness in Stockholm.

In the end, I want to thank my parents for raising me with all their love, selflessly and unreservedly.

# REFERENCES

1. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I *et al*: **From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis**. *Bio/technology* 1996, **14**(1):61-65.
2. James P: **Protein identification in the post-genome era: the rapid rise of proteomics**. *Quarterly reviews of biophysics* 1997, **30**(4):279-331.
3. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies**. *Nature reviews Genetics* 2016, **17**(6):333-351.
4. Uhlen M, Ponten F: **Antibody-based proteomics for human tissue profiling**. *Molecular & cellular proteomics : MCP* 2005, **4**(4):384-393.
5. Brennan DJ, O'Connor DP, Rexhepaj E, Ponten F, Gallagher WM: **Antibody-based proteomics: fast-tracking molecular diagnostics in oncology**. *Nature reviews Cancer* 2010, **10**(9):605-617.
6. Stoevesandt O, Taussig MJ: **Affinity proteomics: the role of specific binding reagents in human proteome analysis**. *Expert review of proteomics* 2012, **9**(4):401-414.
7. Solier C, Langen H: **Antibody-based proteomics and biomarker research - current status and limitations**. *Proteomics* 2014, **14**(6):774-783.
8. Wingren C: **Antibody-Based Proteomics**. *Adv Exp Med Biol* 2016, **926**:163-179.
9. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.
10. Beranova-Giorgianni S: **Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations**. *TrAC Trends in Analytical Chemistry* 2003, **22**(5):273-281.
11. Domon B, Aebersold R: **Mass spectrometry and protein analysis**. *Science* 2006, **312**(5771):212-217.
12. Han X, Aslanian A, Yates JR, 3rd: **Mass spectrometry for proteomics**. *Current opinion in chemical biology* 2008, **12**(5):483-490.
13. Biemann K, Seibl J, Gapp F: **Mass spectrometric identification of amino acids**. *Biochemical and Biophysical Research Communications* 1959, **1**(6):307-311.
14. Biemann K, Vetter W: **Quantitative amino acid analysis by mass spectrometry**. *Biochemical and Biophysical Research Communications* 1960, **2**(2):93-96.
15. Morris HR, Panico M, Barber M, Bordoli RS, Sedgwick RD, Tyler A: **Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis**. *Biochem Biophys Res Commun* 1981, **101**(2):623-631.
16. Barber M, Bordoli RS, Sedgwick RD, Tyler AN: **Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry**. *Journal of the Chemical Society, Chemical Communications* 1981(7):325-327.
17. Barber M, Bordoli RS, Sedgwick RD, Tyler AN: **Fast atom bombardment of solids as an ion source in mass spectrometry**. *Nature* 1981, **293**(5830):270-275.
18. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**. *Analytical chemistry* 1988, **60**(20):2299-2301.
19. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules**. *Science* 1989, **246**(4926):64-71.
20. Wisniewski JR, Zougman A, Nagaraj N, Mann M: **Universal sample preparation method for proteome analysis**. *Nature methods* 2009, **6**(5):359-362.

21. Hughes CS, Foehr S, Garfield DA, Furlong EE, Steinmetz LM, Krijgsveld J: **Ultrasensitive proteome analysis using paramagnetic bead technology**. *Molecular systems biology* 2014, **10**:757.
22. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM: **Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies**. *Journal of proteome research* 2010, **9**(8):4152-4160.
23. Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC: **The dynamic range of protein expression: a challenge for proteomic research**. *Electrophoresis* 2000, **21**(6):1104-1115.
24. Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, Aebersold R: **The quantitative proteome of a human cell line**. *Molecular systems biology* 2011, **7**:549.
25. Anderson NL, Anderson NG: **The human plasma proteome: history, character, and diagnostic prospects**. *Molecular & cellular proteomics : MCP* 2002, **1**(11):845-867.
26. Mitchell P: **Proteomics retrenches**. *Nat Biotech* 2010, **28**(7):665-670.
27. Pernemalm M, Lehtio J: **Mass spectrometry-based plasma proteomics: state of the art and future outlook**. *Expert review of proteomics* 2014, **11**(4):431-448.
28. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR, 3rd: **Direct analysis of protein complexes using mass spectrometry**. *Nature biotechnology* 1999, **17**(7):676-682.
29. Washburn MP, Wolters D, Yates JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology**. *Nature biotechnology* 2001, **19**(3):242-247.
30. Gilar M, Olivova P, Daly AE, Gebler JC: **Orthogonality of separation in two-dimensional liquid chromatography**. *Analytical chemistry* 2005, **77**(19):6426-6434.
31. Dwivedi RC, Spicer V, Harder M, Antonovici M, Ens W, Standing KG, Wilkins JA, Krokhin OV: **Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics**. *Analytical chemistry* 2008, **80**(18):7036-7042.
32. Wang Y, Yang F, Gritsenko MA, Wang Y, Clauss T, Liu T, Shen Y, Monroe ME, Lopez-Ferrer D, Reno T *et al*: **Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells**. *Proteomics* 2011, **11**(10):2019-2026.
33. Cargile BJ, Bundy JL, Freeman TW, Stephenson JL, Jr.: **Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification**. *Journal of proteome research* 2004, **3**(1):112-119.
34. Cargile BJ, Sevinsky JR, Essader AS, Stephenson JL, Jr., Bundy JL: **Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics**. *Journal of biomolecular techniques : JBT* 2005, **16**(3):181-189.
35. Cargile BJ, Talley DL, Stephenson JL, Jr.: **Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides**. *Electrophoresis* 2004, **25**(6):936-945.
36. Slebos RJ, Brock JW, Winters NF, Stuart SR, Martinez MA, Li M, Chambers MC, Zimmerman LJ, Ham AJ, Tabb DL *et al*: **Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry**. *Journal of proteome research* 2008, **7**(12):5286-5294.
37. Branca RM, Orre LM, Johansson HJ, Granholm V, Huss M, Perez-Bercoff A, Forshed J, Kall L, Lehtio J: **HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics**. *Nature methods* 2014, **11**(1):59-62.

38. Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, Fernandez-Woodbridge A, Branca RMM, Lehtio J: **Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow**. *Nat Commun* 2018, **9**(1):903.
39. Eriksson J, Fenyo D: **Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs**. *Nature biotechnology* 2007, **25**(6):651-655.
40. Eng JK, McCormack AL, Yates Iii JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *Journal of the American Society for Mass Spectrometry* 1994, **5**(11):976-989.
41. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**(18):3551-3567.
42. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**(9):1466-1467.
43. Field HI, Fenyo D, Beavis RC: **RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database**. *Proteomics* 2002, **2**(1):36-47.
44. Eriksson J, Fenyo D: **Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results**. *Journal of proteome research* 2004, **3**(1):32-36.
45. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *Journal of proteome research* 2004, **3**(5):958-964.
46. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: **InsPecT: identification of posttranslationally modified peptides from tandem mass spectra**. *Analytical chemistry* 2005, **77**(14):4626-4639.
47. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: **Andromeda: a peptide search engine integrated into the MaxQuant environment**. *Journal of proteome research* 2011, **10**(4):1794-1805.
48. Wenger CD, Coon JJ: **A proteomics search algorithm specifically designed for high-resolution tandem mass spectra**. *Journal of proteome research* 2013, **12**(3):1377-1386.
49. Kim S, Gupta N, Pevzner PA: **Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases**. *Journal of proteome research* 2008, **7**(8):3354-3363.
50. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJ, Pevzner PA: **The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search**. *Molecular & cellular proteomics : MCP* 2010, **9**(12):2840-2852.
51. Kim S, Pevzner PA: **MS-GF+ makes progress towards a universal database search tool for proteomics**. *Nat Commun* 2014, **5**:5277.
52. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS *et al*: **An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis**. *Proteomics* 2005, **5**(13):3475-3490.
53. Zhang B, Pirmoradian M, Chernobrovkin A, Zubarev RA: **DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry**. *Molecular & cellular proteomics : MCP* 2014, **13**(11):3211-3223.

54. Fenyo D, Beavis RC: **A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes**. *Analytical chemistry* 2003, **75**(4):768-774.

55. Eriksson J, Fenyo D: **A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis**. *Proteomics* 2002, **2**(3):262-270.

56. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**. *Nature methods* 2007, **4**(3):207-214.

57. Kall L, Storey JD, MacCoss MJ, Noble WS: **Assigning significance to peptides identified by tandem mass spectrometry using decoy databases**. *Journal of proteome research* 2008, **7**(1):29-34.

58. Elias JE, Gygi SP: **Target-decoy search strategy for mass spectrometry-based proteomics**. *Methods in molecular biology (Clifton, NJ)* 2010, **604**:55-71.

59. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ: **Semi-supervised learning for peptide identification from shotgun proteomics datasets**. *Nature methods* 2007, **4**(11):923-925.

60. Nesvizhskii AI, Aebersold R: **Interpretation of shotgun proteomic data: the protein inference problem**. *Molecular & cellular proteomics : MCP* 2005, **4**(10):1419-1440.

61. Huang T, Wang J, Yu W, He Z: **Protein inference: a review**. *Briefings in bioinformatics* 2012, **13**(5):586-614.

62. Li YF, Radivojac P: **Computational approaches to protein inference in shotgun proteomics**. *BMC bioinformatics* 2012, **13 Suppl 16**:S4.

63. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM: **Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study**. *Nature biotechnology* 2006, **24**(3):333-338.

64. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R: **Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry**. *Molecular & cellular proteomics : MCP* 2009, **8**(11):2405-2417.

65. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: **A statistical model for identifying proteins by tandem mass spectrometry**. *Analytical chemistry* 2003, **75**(17):4646-4658.

66. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW *et al*: **IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering**. *Journal of proteome research* 2009, **8**(8):3872-3881.

67. Serang O, MacCoss MJ, Noble WS: **Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data**. *Journal of proteome research* 2010, **9**(10):5346-5357.

68. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M: **A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets**. *Molecular & cellular proteomics : MCP* 2015, **14**(9):2394-2404.

69. Liu H, Sadygov RG, Yates JR, 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics**. *Analytical chemistry* 2004, **76**(14):4193-4201.

70. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M: **Exponentially modified protein abundance index (emPAI) for estimation of absolute protein**

**amount in proteomics by the number of sequenced peptides per protein**. *Molecular & cellular proteomics : MCP* 2005, **4**(9):1265-1272.

71. Lu P, Vogel C, Wang R, Yao X, Marcotte EM: **Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation**. *Nature biotechnology* 2007, **25**(1):117-124.
72. Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, Lee A, van Sluyter SC, Haynes PA: **Less label, more free: approaches in label-free quantitative mass spectrometry**. *Proteomics* 2011, **11**(4):535-553.
73. Chelius D, Bondarenko PV: **Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry**. *Journal of proteome research* 2002, **1**(4):317-323.
74. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M: **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ**. *Molecular & cellular proteomics : MCP* 2014, **13**(9):2513-2526.
75. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nature biotechnology* 1999, **17**(10):994-999.
76. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Molecular & cellular proteomics : MCP* 2002, **1**(5):376-386.
77. Mann M: **Functional and quantitative proteomics using SILAC**. *Nat Rev Mol Cell Biol* 2006, **7**(12):952-958.
78. Ong SE, Mann M: **A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)**. *Nature protocols* 2006, **1**(6):2650-2660.
79. Geiger T, Wisniewski JR, Cox J, Zanivan S, Kruger M, Ishihama Y, Mann M: **Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics**. *Nat Protocols* 2011, **6**(2):147-157.
80. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M: **Super-SILAC mix for quantitative proteomics of human tumor tissue**. *Nat Meth* 2010, **7**(5):383-385.
81. Boersema PJ, Aye TT, van Veen TA, Heck AJ, Mohammed S: **Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates**. *Proteomics* 2008, **8**(22):4624-4632.
82. Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR: **Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics**. *Nat Protocols* 2009, **4**(4):484-494.
83. Wu Y, Wang F, Liu Z, Qin H, Song C, Huang J, Bian Y, Wei X, Dong J, Zou H: **Five-plex isotope dimethyl labeling for quantitative proteomics**. *Chemical communications* 2014, **50**(14):1708-1710.
84. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al*: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents**. *Molecular & cellular proteomics : MCP* 2004, **3**(12):1154-1169.
85. Choe L, D'Ascenzo M, Relkin NR, Pappin D, Ross P, Williamson B, Guertin S, Pribil P, Lee KH: **8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease**. *Proteomics* 2007, **7**(20):3651-3660.
86. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C: **Tandem mass tags: a novel quantification**

**strategy for comparative analysis of complex protein mixtures by MS/MS**. *Analytical chemistry* 2003, **75**(8):1895-1904.

87. Dayon L, Hainard A, Licker V, Turck N, Kuhn K, Hochstrasser DF, Burkhard PR, Sanchez JC: **Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags**. *Analytical chemistry* 2008, **80**(8):2921-2931.
88. Werner T, Sweetman G, Savitski MF, Mathieson T, Bantscheff M, Savitski MM: **Ion coalescence of neutron encoded TMT 10-plex reporter ions**. *Analytical chemistry* 2014, **86**(7):3594-3601.
89. Suomi T, Corthals GL, Nevalainen OS, Elo LL: **Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins**. *Journal of proteome research* 2015, **14**(11):4564-4570.
90. Ting YS, Egertson JD, Payne SH, Kim S, MacLean B, Kall L, Aebersold R, Smith RD, Noble WS, MacCoss MJ: **Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data**. *Molecular & cellular proteomics : MCP* 2015, **14**(9):2301-2307.
91. Clough T, Key M, Ott I, Ragg S, Schadow G, Vitek O: **Protein quantification in label-free LC-MS experiments**. *Journal of proteome research* 2009, **8**(11):5275-5284.
92. Herbrich SM, Cole RN, West KP, Jr., Schulze K, Yager JD, Groopman JD, Christian P, Wu L, O'Meally RN, May DH *et al*: **Statistical inference from multiple iTRAQ experiments without using common reference standards**. *Journal of proteome research* 2013, **12**(2):594-604.
93. D'Angelo G, Chaerkady R, Yu W, Hizal DB, Hess S, Zhao W, Lekstrom K, Guo X, White WI, Roskos L *et al*: **Statistical Models for the Analysis of Isobaric Tags Multiplexed Quantitative Proteomics**. *Journal of proteome research* 2017, **16**(9):3124-3136.
94. Hultin-Rosenberg L, Forshed J, Branca RM, Lehtio J, Johansson HJ: **Defining, comparing, and improving iTRAQ quantification in mass spectrometry proteomics data**. *Molecular & cellular proteomics : MCP* 2013, **12**(7):2021-2031.
95. de Groen PPN: **An Introduction to Total Least Squares**. *ARXIV* 1998, **eprint arXiv:math/9805076**.
96. Carrillo B, Yanofsky C, Laboissiere S, Nadon R, Kearney RE: **Methods for combining peptide intensities to estimate relative protein abundance**. *Bioinformatics* 2010, **26**(1):98-103.
97. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, Koziol JA, Schnitzer JE: **Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis**. *Nature biotechnology* 2010, **28**(1):83-89.
98. Pan C, Kora G, Tabb DL, Pelletier DA, McDonald WH, Hurst GB, Hettich RL, Samatova NF: **Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics**. *Analytical chemistry* 2006, **78**(20):7110-7120.
99. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics* 2006, **22**(8):943-949.
100. Zhang B, Pirmoradian M, Zubarev R, Kall L: **Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences**. *Molecular & cellular proteomics : MCP* 2017, **16**(5):936-948.
101. Daly DS, Anderson KK, Panisko EA, Purvine SO, Fang R, Monroe ME, Baker SE: **Mixed-effects statistical model for comparative LC-MS proteomics studies**. *Journal of proteome research* 2008, **7**(3):1209-1217.
102. Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, Cooper LT, Onuma OK, Spiro C, Therneau TM *et al*: **Statistical analysis of relative**

**labeled mass spectrometry data from complex samples using ANOVA**. *Journal of proteome research* 2008, **7**(1):225-233.

103. Forshed J, Johansson HJ, Pernemalm M, Branca RM, Sandberg A, Lehtio J: **Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ)**. *Molecular & cellular proteomics : MCP* 2011, **10**(10):M111 010264.

104. Hill EG, Schwacke JH, Comte-Walters S, Slate EH, Oberg AL, Eckel-Passow JE, Therneau TM, Schey KL: **A statistical model for iTRAQ data analysis**. *Journal of proteome research* 2008, **7**(8):3091-3101.

105. Han D, Jin J, Woo J, Min H, Kim Y: **Proteomic analysis of mouse astrocytes and their secretome by a combination of FASP and StageTip-based, high pH, reversed-phase fractionation**. *Proteomics* 2014, **14**(13-14):1604-1609.

106. Pan D, Gao J, Zeng X, Ma G, Li N, Huang X, Du X, Miao Q, Lian J, Xu L *et al*: **Quantitative proteomic Analysis Reveals up-regulation of caveolin-1 in FOXP3-overexpressed human gastric cancer cells**. *Scientific reports* 2017, **7**(1):14460.

107. Zhou Y, Frings O, Branca RM, Boekel J, le Sage C, Fredlund E, Agami R, Orre LM: **microRNAs with AAGUGC seed motif constitute an integral part of an oncogenic signaling network**. *Oncogene* 2017, **36**(6):731-745.

108. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.

109. Kammers K, Cole RN, Tiengwe C, Ruczinski I: **Detecting Significant Changes in Protein Abundance**. *EuPA Open Proteom* 2015, **7**:11-19.

110. Efstathiou G, Antonakis AN, Pavlopoulos GA, Theodosiou T, Divanach P, Trudgian DC, Thomas B, Papanikolaou N, Aivaliotis M, Acuto O *et al*: **ProteoSign: an end-user online differential proteomics statistical analysis platform**. *Nucleic acids research* 2017, **45**(W1):W300-W306.

111. Karpievitch YV, Dabney AR, Smith RD: **Normalization and missing value imputation for label-free LC-MS analysis**. *BMC bioinformatics* 2012, **13 Suppl 16**:S5.

112. Lazar C, Gatto L, Ferro M, Bruley C, Burger T: **Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies**. *Journal of proteome research* 2016, **15**(4):1116-1125.

113. Koopmans F, Cornelisse LN, Heskes T, Dijkstra TM: **Empirical Bayesian random censoring threshold model improves detection of differentially abundant proteins**. *Journal of proteome research* 2014, **13**(9):3871-3880.

114. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al*: **A draft map of the human proteome**. *Nature* 2014, **509**(7502):575-581.

115. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.

116. Schuster SC: **Next-generation sequencing transforms today's biology**. *Nat Meth* 2008, **5**(1):16-18.

117. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing**. *Genome Biology* 2010, **11**.

118. Alfaro JA, Sinha A, Kislinger T, Boutros PC: **Onco-proteogenomics: cancer proteomics joins forces with genomics**. *Nat Meth* 2014, **11**(11):1107-1113.

119. Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies**. *Nat Meth* 2014, **11**(11):1114-1125.

120. Pandey A, Pevzner PA: **Proteogenomics**. *Proteomics* 2014, **14**(23-24):2631-2632.

121. Rivers RC, Kinsinger C, Boja ES, Hiltke T, Mesri M, Rodriguez H: **Linking Cancer Genome to Proteome: NCI's Investment into Proteogenomics**. *Proteomics* 2014.

122. Faulkner S, Dun MD, Hondermarck H: **Proteogenomics: emergence and promise**. *Cellular and molecular life sciences : CMLS* 2015, **72**(5):953-957.

123. Menschaert G, Fenyo D: **Proteogenomics from a bioinformatics angle: A growing field**. *Mass spectrometry reviews* 2015.

124. Polyakova A, Kuznetsova K, Moshkovskii S: **Proteogenomics meets cancer immunology: mass spectrometric discovery and analysis of neoantigens**. *Expert review of proteomics* 2015, **12**(5):533-541.

125. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM: **Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation**. *Annual review of analytical chemistry* 2016, **9**(1):521-545.

126. Ma S, Ren J, Fenyo D: **Breast Cancer Prognostics Using Multi-Omics Data**. *AMIA Jt Summits Transl Sci Proc* 2016, **2016**:52-59.

127. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S *et al*: **Proteogenomic characterization of human colon and rectal cancer**. *Nature* 2014, **513**(7518):382-387.

128. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou JY, Petyuk VA, Chen L, Ray D *et al*: **Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer**. *Cell* 2016, **166**(3):755-765.

129. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F *et al*: **Proteogenomics connects somatic mutations to signalling in breast cancer**. *Nature* 2016, **534**(7605):55-62.

130. Huang KL, Li S, Mertins P, Cao S, Gunawardena HP, Ruggles KV, Mani DR, Clauser KR, Tanioka M, Usary J *et al*: **Proteogenomic integration reveals therapeutic targets in breast cancer xenografts**. *Nat Commun* 2017, **8**:14864.

131. Castellana N, Bafna V: **Proteogenomics to discover the full coding content of genomes: A computational perspective**. *Journal of proteomics* 2010, **73**(11):2124-2135.

132. Li H, Joh YS, Kim H, Paek E, Lee SW, Hwang KB: **Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification**. *BMC genomics* 2016, **17**(Suppl 13):1031.

133. Eriksson J, Fenyo D: **The statistical significance of protein identification results as a function of the number of protein sequences searched**. *Journal of proteome research* 2004, **3**(5):979-982.

134. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation**. *Proteomics* 2004, **4**(1):59-77.

135. Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD *et al*: **Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation**. *Genome research* 2007, **17**(9):1362-1377.

136. Oshiro G, Wodicka LM, Washburn MP, Yates JR, 3rd, Lockhart DJ, Winzeler EA: **Parallel identification of new genes in Saccharomyces cerevisiae**. *Genome research* 2002, **12**(8):1210-1220.

137. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O *et al*: **A high-quality catalog of the Drosophila melanogaster proteome**. *Nature biotechnology* 2007, **25**(5):576-583.

138. Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S: **Genome-scale proteomics**

**reveals Arabidopsis thaliana gene models and proteome dynamics**. *Science* 2008, **320**(5878):938-941.

139. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(52):21034-21038.
140. Xing X-B, Li Q-R, Sun H, Fu X, Zhan F, Huang X, Li J, Chen C-L, Shyr Y, Zeng R *et al*: **The discovery of novel protein-coding features in mouse genome based on mass spectrometry data**. *Genomics* 2011, **98**(5):343-351.
141. Khatun J, Yu Y, Wrobel JA, Risk BA, Gunawardena HP, Secrest A, Spitzer WJ, Xie L, Wang L, Chen X *et al*: **Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions**. *BMC genomics* 2013, **14**:141.
142. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: **Peptidomic discovery of short open reading frame-encoded peptides in human cells**. *Nature chemical biology* 2013, **9**(1):59-64.
143. Li J, Su Z, Ma ZQ, Slebos RJ, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B: **A bioinformatics workflow for variant peptide detection in shotgun proteomics**. *Molecular & cellular proteomics : MCP* 2011, **10**(5):M110 006536.
144. Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B: **Protein identification using customized protein sequence databases derived from RNA-Seq data**. *Journal of proteome research* 2012, **11**(2):1009-1017.
145. Sheynkman GM, Shortreed MR, Frey BL, Smith LM: **Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq**. *Molecular & cellular proteomics : MCP* 2013, **12**(8):2341-2353.
146. Wang X, Zhang B: **customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search**. *Bioinformatics* 2013, **29**(24):3235-3237.
147. Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, Griffin TJ, Smith LM: **Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations**. *BMC genomics* 2014, **15**:703.
148. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyo D, Zhang B, Mani DR: **Methods, Tools and Current Perspectives in Proteogenomics**. *Molecular & cellular proteomics : MCP* 2017, **16**(6):959-981.
149. Askenazi M, Ruggles KV, Fenyo D: **PGx: Putting Peptides to BED**. *Journal of proteome research* 2016, **15**(3):795-799.
150. Keegan S, Cortens JP, Beavis RC, Fenyo D: **g2pDB: A Database Mapping Protein Post-Translational Modifications to Genomic Coordinates**. *Journal of proteome research* 2016, **15**(3):983-990.
151. Mashl RJ, Scott AD, Huang KL, Wyczalkowski MA, Yoon CJ, Niu B, DeNardo E, Yellapantula VD, Handsaker RE, Chen K *et al*: **GenomeVIP: a cloud platform for genomic variant discovery and interpretation**. *Genome research* 2017, **27**(8):1450-1459.
152. Pang CN, Tay AP, Aya C, Twine NA, Harkness L, Hart-Smith G, Chia SZ, Chen Z, Deshpande NP, Kaakoush NO *et al*: **Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing**. *Journal of proteome research* 2014, **13**(1):84-98.
153. Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Kall L, Lehtio J, Lukasse P, Moerland PD, Griffin TJ: **Multi-omic data analysis using Galaxy**. *Nature biotechnology* 2015, **33**(2):137-139.

154. Krasnov GS, Dmitriev AA, Kudryavtseva AV, Shargunov AV, Karpov DS, Uroshlev LA, Melnikova NV, Blinov VM, Poverennaya EV, Archakov AI *et al*: **PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics**. *Journal of proteome research* 2015, **14**(9):3729-3737.

155. Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, Mandyam RA, Nones K, Pearson JV, Grimmond SM: **PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization**. *Journal of proteome research* 2015, **14**(5):2255-2266.

156. Li Y, Wang X, Cho JH, Shaw TI, Wu Z, Bai B, Wang H, Zhou S, Beach TG, Wu G *et al*: **JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells**. *Journal of proteome research* 2016, **15**(7):2309-2320.

157. Zhang K, Fu Y, Zeng WF, He K, Chi H, Liu C, Li YC, Gao Y, Xu P, He SM: **A note on the false discovery rate of novel peptides in proteogenomics**. *Bioinformatics* 2015, **31**(20):3249-3253.

158. Bruford EA, Lane L, Harrow J: **Devising a Consensus Framework for Validation of Novel Human Coding Loci**. *Journal of proteome research* 2015, **14**(12):4945-4948.

159. Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, Choudhary JS, Harrow J: **Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow**. *Nat Commun* 2016, **7**:11778.

160. Wang X, Codreanu SG, Wen B, Li K, Chambers MC, Liebler DC, Zhang B: **Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity**. *Molecular & cellular proteomics : MCP* 2018, **17**(3):422-430.

161. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**. *Genome research* 2008, **18**(1):188-196.

162. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nature biotechnology* 2011, **29**(7):644-652.

163. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucleic acids research* 2003, **31**(19):5654-5666.

164. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training**. *Genome research* 2008, **18**(12):1979-1990.

165. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: a web server for gene finding in eukaryotes**. *Nucleic acids research* 2004, **32**(Web Server issue):W309-312.

166. Korf I: **Gene finding in novel genomes**. *BMC bioinformatics* 2004, **5**:59.

167. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE: **Web Apollo: a web-based genomic annotation editing platform**. *Genome Biol* 2013, **14**(8):R93.

168. Gioti A, Nystedt B, Li W, Xu J, Andersson A, Averette AF, Munch K, Wang X, Kappauf C, Kingsbury JM *et al*: **Genomic insights into the atopic eczema-associated skin commensal yeast Malassezia sympodialis**. *MBio* 2013, **4**(1):e00572-00512.

169. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research* 2010, **38**(16):e164.

170. Lin MF, Jungreis I, Kellis M: **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions**. *Bioinformatics* 2011, **27**(13):i275-282.

171. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling**. *Science* 2009, **324**(5924):218-223.

172. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, Schreiber S, Platzer M, Krawczak M, Hampe J *et al*: **Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting**. *Genome research* 2012, **22**(11):2208-2218.

173. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T *et al*: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**(7493):462-470.

174. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome research* 2005, **15**(8):1034-1050.

175. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ *et al*: **Expressed pseudogenes in the transcriptional landscape of human cancers**. *Cell* 2012, **149**(7):1622-1634.

176. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M *et al*: **The GENCODE pseudogene resource**. *Genome Biol* 2012, **13**(9):R51.

177. Paulovich AG, Billheimer D, Ham AJ, Vega-Montoto L, Rudnick PA, Tabb DL, Wang P, Blackman RK, Bunk DM, Cardasis HL *et al*: **Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance**. *Molecular & cellular proteomics : MCP* 2010, **9**(2):242-254.

178. Boja ES, Rodriguez H: **Proteogenomic convergence for understanding cancer pathways and networks**. *Clinical proteomics* 2014, **11**(1):22.

179. Olsen L, Campos B, Winther O, Sgroi DC, Karger BL, Brusic V: **Tumor antigens as proteogenomic biomarkers in invasive ductal carcinomas**. *BMC medical genomics* 2014, **7 Suppl 3**:S2.

180. Woo S, Cha SW, Na S, Guest C, Liu T, Smith RD, Rodland KD, Payne S, Bafna V: **Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data**. *Proteomics* 2014, **14**(23-24):2719-2730.

181. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, Franci C, Cheung TK, Fritsche J, Weinschenk T *et al*: **Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing**. *Nature* 2014, **515**(7528):572-576.

182. Huang CH, Kuo CJ, Liang SS, Chi SW, Hsi E, Chen CC, Lee KT, Chiou SH: **Onco-proteogenomics identifies urinary S100A9 and GRN as potential combinatorial biomarkers for early diagnosis of hepatocellular carcinoma**. *BBA clinical* 2015, **3**:205-213.

183. Shukla HD, Mahmood J, Vujaskovic Z: **Integrated proteo-genomic approach for early diagnosis and prognosis of cancer**. *Cancer letters* 2015, **369**(1):28-36.

184. Locard-Paulet M, Pible O, Gonzalez de Peredo A, Alpha-Bazin B, Almunia C, Burlet-Schiltz O, Armengaud J: **Clinical implications of recent advances in proteogenomics**. *Expert review of proteomics* 2016, **13**(2):185-199.

185. Mann M, Wilm M: **Error-tolerant identification of peptides in sequence databases by peptide sequence tags**. *Analytical chemistry* 1994, **66**(24):4390-4399.

186. Tabb DL, Saraf A, Yates JR, 3rd: **GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model**. *Analytical chemistry* 2003, **75**(23):6415-6421.

187. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL: **TagRecon: high-throughput mutation identification through sequence tagging**. *Journal of proteome research* 2010, **9**(4):1716-1726.

188. Taylor JA, Johnson RS: **Sequence database searches via de novo peptide sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 1997, **11**(9):1067-1075.

189. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA: **De novo peptide sequencing via tandem mass spectrometry**. *Journal of computational biology : a journal of computational molecular cell biology* 1999, **6**(3-4):327-342.

190. Taylor JA, Johnson RS: **Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry**. *Analytical chemistry* 2001, **73**(11):2594-2604.

191. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G: **PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 2003, **17**(20):2337-2342.

192. Frank A, Pevzner P: **PepNovo: de novo peptide sequencing via probabilistic network modeling**. *Analytical chemistry* 2005, **77**(4):964-973.

193. Pan C, Park BH, McDonald WH, Carey PA, Banfield JF, VerBerkmoes NC, Hettich RL, Samatova NF: **A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry**. *BMC bioinformatics* 2010, **11**:118.

194. Tran NH, Zhang X, Xin L, Shan B, Li M: **De novo peptide sequencing by deep learning**. *Proceedings of the National Academy of Sciences of the United States of America* 2017.

195. Kim S, Gupta N, Bandeira N, Pevzner PA: **Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra**. *Molecular & cellular proteomics : MCP* 2009, **8**(1):53-69.

196. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA: **Identification of post-translational modifications by blind search of mass spectra**. *Nature biotechnology* 2005, **23**(12):1562-1567.

197. The M, Kall L: **MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics**. *Journal of proteome research* 2016, **15**(3):713-720.

198. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP: **A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides**. *Nature biotechnology* 2015, **33**(7):743-749.

199. Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP: **A tissue-specific atlas of mouse protein phosphorylation and expression**. *Cell* 2010, **143**(7):1174-1189.

200. Panizza E, Branca RMM, Oliviusson P, Orre LM, Lehtio J: **Isoelectric point-based fractionation by HiRIEF coupled to LC-MS allows for in-depth quantitative analysis of the phosphoproteome**. *Scientific reports* 2017, **7**(1):4513.

201. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M: **Global, in vivo, and site-specific phosphorylation dynamics in signaling networks**. *Cell* 2006, **127**(3):635-648.

202. Woo EM, Fenyo D, Kwok BH, Funabiki H, Chait BT: **Efficient identification of phosphorylation by mass spectrometric phosphopeptide fingerprinting**. *Analytical chemistry* 2008, **80**(7):2419-2425.

203. Mayya V, Han DK: **Phosphoproteomics by mass spectrometry: insights, implications, applications and limitations**. *Expert review of proteomics* 2009, **6**(6):605-618.

204. Lind SB, Artemenko KA, Pettersson U: **A strategy for identification of protein tyrosine phosphorylation**. *Methods* 2012, **56**(2):275-283.

205. Olsen JV, Mann M: **Status of large-scale analysis of post-translational modifications by mass spectrometry**. *Molecular & cellular proteomics : MCP* 2013, **12**(12):3444-3452.

206. Kallsten M, Bergquist J, Zhao H, Konzer A, Lind SB: **A comparative study of phosphopeptide-selective techniques for a sub-proteome of a complex biological sample**. *Anal Bioanal Chem* 2016, **408**(9):2347-2356.

207. Wang X, Slebos RJ, Chambers MC, Tabb DL, Liebler DC, Zhang B: **proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data**. *Molecular & cellular proteomics : MCP* 2016, **15**(3):1164-1175.

208. Menschaert G, Wang X, Jones AR, Ghali F, Fenyo D, Olexiouk V, Zhang B, Deutsch EW, Ternent T, Vizcaino JA: **The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data**. *Genome Biol* 2018, **19**(1):12.

209. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

210. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis**. *Curr Protoc Bioinformatics* 2014, **47**:11 12 11-34.

211. Olexiouk V, Menschaert G: **proBAMconvert: A Conversion Tool for proBAM/proBed**. *Journal of proteome research* 2017, **16**(7):2639-2644.