Mutant Proteogenomics

6

Ákos Végvári

Abstract

Identification of mutant proteins in biological samples is one of the emerging areas of proteogenomics. Despite the fact that only a limited number of studies have been published up to now, it has the potential to recognize novel disease biomarkers that have unique structure and desirably high specificity. Such properties would identify mutant proteoforms related to diseases as optimal drug targets useful for future therapeutic strategies. While mass spectrometry has demonstrated its outstanding analytical power in proteomics, the most frequently applied bottom-up strategy is not suitable for the detection of mutant proteins if only databases with consensus sequences are searched. It is likely that many unassigned tandem mass spectra of tryptic peptides originate from single amino acid variants (SAAVs). To address this problem, a couple of protein databases have been constructed that include canonical and SAAV sequences, allowing for the observation of mutant proteoforms in mass spectral data for the first time. Since the resulting large search space may compromise the probability of identifications, a novel concept was proposed that included identification as well as verification strategies. Together with transcriptome based approaches, targeted proteomics appears to be a suitable method for the verification of initial identifications in databases and can also provide quantitative insights to expression profiles, which often reflect disease progression. Important applications in the field of mutant proteoform identification have already highlighted novel biomarkers in large-scale investigations.

Á. Végvári (✉)
Clinical Protein Science & Imaging, Department of Medical Bioengineering, Biomedical Center, Lund University, Lund, Sweden

Department of Pharmacology & Toxicology, University of Texas Medical Branch, Galveston, TX, USA
e-mail: akos.vegvari@BME.LTH.SE; akvegvar@UTMB.EDU

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_6

Keywords

Non-synonymous single nucleotide polymorphism • Single amino acid variant • Mutation • Bottom-up proteomics • Mass spectrometry • Tandem mass spectra • Protein databases • Mutant protein identification

6.1 Introduction

The technological development in the field of mass spectrometry (MS) based proteomics, in particular bottom-up proteomics, offers a powerful approach to verify newly identified open reading frames (ORFs) and genes (Pandey and Pevzner 2014), which is in line with the Human Genome Organization (HUGO) project. The Human Protein Organization (HUPO) has analogously planned to outline similar goals, including the completion of the human protein catalogue (Paik et al. 2012; Omenn et al. 2015). To facilitate the identification of proteins, databases have been created and maintained, adding and improving protein sequences continuously in these repositories. The most frequently used protein databases (*e.g.*, UniProt/SwissProt and neXtProt) were designed to include the most common forms of human proteins, hence the definition of consensus or canonical protein sequences were established. Due to the interplay between genomic and proteomics research, these protein databases were extended with known splicing isoforms (alternative splicing variants or ASVs), increasing the number of entries by a factor of 2 to 3 (for instance, the neXtProt 2015-09-01 release has 20,066 consensus and 21,932 ASVs). Additionally, great attention has been given to post-translational modifications (PTMs), such as phosphorylation, ubiquitination, glycosylation, etc. Information on PTMs provides a huge amount of novel information about proteoforms with various functions that could be useful in description of disease progression (Nørregaard Jensen 2004). Interestingly, however, little attention has been shown towards mutant proteins, although they represent a level of variability of molecular forms between ASVs and PTMs. In particular, single amino acid variants (SAAVs) are of interest, since:

1. Genetic information/data is often available
2. They may be the ultimate markers as their unique sequence may alter the function
3. They may complicate the quantification of given proteins

The importance of finding non-synonymous single nucleotide polymorphisms (nsSNPs) does not only lie in the discovery of variations in the amino acid sequences that have functional consequences but also to provide information regarding the genetic, and possibly phenotypic, variability within the population of samples (Salisbury et al. 2003). Secondary validation by sequencing of corresponding genomic DNA has confirmed the presence of the predicted single nucleotide polymorphisms (SNPs) in 8 out of 10 SNP-peptides. In their study, Bunger *et al.* highlighted the usefulness of interpreting unassigned spectra as polymorphisms (Bunger et al. 2007). Although, DNA genotyping scans have perhaps the greatest utility in defining the haplotype structure on a genome-wide scale, proteins are a major functional component of most disease progressions. Therefore, information gained from being able to reliably monitor SNP products in proteomic data allows more functional inference to be assigned to expressed alleles. In this regard, the utility of detecting expressed SNPs in proteomics assays will integrate protein profiling with genome information. Such analysis will reveal differential allelic expressions that can be correlated to phenotypic variation between individuals. Recent interest in differential allelic expression has been driven by the discovery that 45–56 % of heterozygous alleles in humans are differentially expressed by a factor of two or more.

6.2 Theoretical Considerations

By definition, a gene mutation is a permanent alteration of the DNA region recognized as a gene. A gene mutation can consist of a number of variations that can be characterized as a single nucleotide variation, a longer sequence changes, all the way up to a large change in a segment of a chromosome that involves multiple genes. On the other hand, from a biological point of view, gene mutations can be divided into two types: hereditary (germline) and acquired (somatic), depending on whether germ or somatic cells are the holders of the mutated DNA. In addition to the most common form of genes (wild type), almost all genes have typical variations due to the high frequency of mutations. If a genetic alteration occurs in more than 1 % of the entire population, such a variation is called a polymorphism. SNPs are frequent in the human genome: as many as 3.1 million SNPs have been found by the International HapMap Consortium (Frazer et al. 2007).

The consequences of a SNP in a gene on the production of the protein can be synonymous, nonsense, and non-synonymous, resulting in unaltered (*i.e.*, fully functional), truncated and mutant protein sequences, respectively (Fig. 6.1). Non-synonymous SNPs certainly contribute to the complexity of the proteome but can also provide significant insight into genetic variability when comparing individuals in a population.

While many of the polymorphisms are harmless and responsible for common variations in humans, some of them can contribute to the risk of disease progression. It is known that somatic mutations can drive cancer development and their accumulation in the mitochondrial DNA is associated with an increased risk of some age-related disorders, such as cardiovascular and neurodegenerative diseases (Taylor and Turnbull 2005). Variations can contribute to an increased likelihood to develop certain diseases on the basis of the genetic makeup of an individual, which is often regarded as genetic predisposition or susceptibility. For instance, we can mention single amino acid variants of the *BRCA1* and *BRCA2* genes that can indicate a significantly increased risk of developing breast and ovarian cancer. *BRCA1* and *BRCA2* are human genes that produce tumor suppressor proteins, which help to repair damaged DNA. Upon mutation their protein products are not made, or do not function correctly, which results in impaired DNA repair, which can in turn lead to an increased probability to develop additional genetic alterations in the cells and eventually cause cancer. The understanding of the underlying biological mechanisms at the molecular level has resulted in an improved clinical diagnosis, monitoring these mutations in risk groups of patients. The level of elevated breast cancer risk may vary as mutations on other genes, like *BARD1* and *BRIP1*, are typically also associated with the disease.

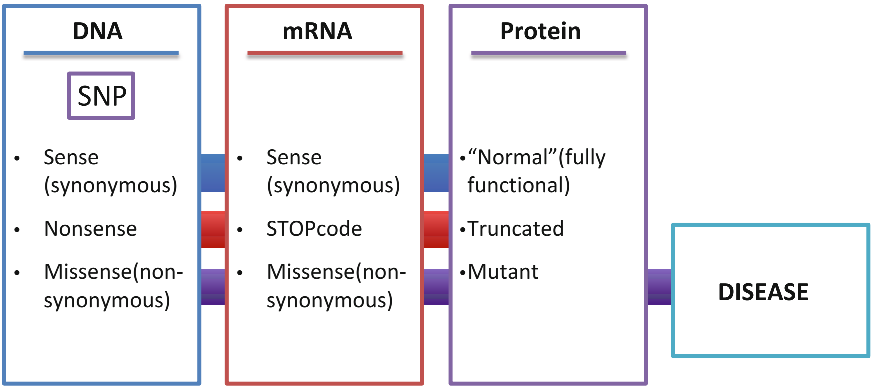

Fig. 6.1 Schematic presentation of single nucleotide polymorphisms in the coding regions of DNA. The non-synonymous SNPs play the most important role in biological mechanisms as their protein products with altered structure and function can contribute to disease progression

Furthermore, protein aggregation has been recognized in neurological disease development, such as Jacob-Krautzfeld disease (associated with prions), Alzheimer's disease and other amyloid diseases (*e.g.*, familial amyloidotic polyneuropathy). A single amino acid alteration may significantly influence protein association and dissociation. For instance, in the case of the familial amyloidotic polyneuropathy (FAP), transthyrein monomer units can aggregate into fibrils that can lead to death upon deposition in the heart and lung. The SAAVs of transthyrein, *p.V30M* and *p.L55P*, facilitate the dissociation of aggregates, while the *p.T119M* proteoform inhibits tetramer dissociation (Hammarström et al. 2003). The main reason for this inhibition is that position 119 is located at the dimer interface. However, low-molecular-weight compounds could be developed to efficiently inhibit transthyrein aggregation by binding to the tetrameric form.

In general, the specific structure of mutant proteoforms allows for the development of efficient drugs. The Pharmaceuticals and Medical Devices Agency in Japan was the first in the world who approved personalized medicine therapy in the case of non-small cell lung cancer, recommending the application of tyrosine kinase inhibitors such as gefitinib and erlotinib. The epidermal growth factor receptor (EGFR) is overexpressed in the cells of certain types of human carcinomas (*e.g.*, in lung and breast cancers), which leads to an inappropriate activation of the anti-apoptotic Ras signaling cascade, eventually causing uncontrolled cell proliferation. It was found that a mutation in the EGFR tyrosine kinase domain is responsible for activating the anti-apoptotic pathways in non-small cell lung cancers (Sordella et al. 2004). These somatic mutations are occur more commonly in lung adenocarcinomas of individuals of Asian descent, women, and non-smokers, rendering erlotinib/gefitinib treatments exceptionally efficient.

It must be noted that the identification of a certain gene as a disease marker is difficult and it is even less probable that this one gene is responsible for disease development alone, indicating that an overall alteration in the genetic profile is expected to be more indicative. Characteristically, variations in individual genes may slightly increase the risk of disease development but the combination of mutations on multiple genes can result in a significant level of risk (Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement 2005). Today, it is generally accepted that the variations in many genes, with their small individual effects, may underlie the susceptibility to many common diseases when combined, this includes diseases like diabetes, obesity, cardiovascular disease and cancer.

Additionally to the understanding of the multifunctional nature of many diseases, and to the tremendously successful genetic investigations on a large number of samples in populations, it is important to emphasize the fact that genes do not have biological function, only their products, the expressed proteins. Therefore, it is necessary to study not only DNA variations but also their corresponding mRNA and their coded proteins, which together can indicate expression activities. Furthermore, the interplay between these principle "omics" areas seems to be inevitable to fully reveal the biology of diseases, unfortunately, the current status of these fields does not always facilitate their integration.

6.3 Methodologies for Detection of Mutant Proteins

6.3.1 Mass Spectrometry Based Proteomics of Mutant Proteoforms

Mass spectrometry based proteomic platforms have the capacity to identify a great number of proteins simultaneously in a single analysis. However, protein identifications by the most commonly applied bottom-up proteomics approach largely rely on protein sequence databases. Due to the fact that such databases are comprised of consensus sequences, *i.e.*, the most frequently observed proteoforms, mass spectra of SAAVs and many other modifications of the analyzed peptides will be unassigned. In principle, known ASVs can be identified searching databases, like neXtProt and UniProt, if isoform specific tryptic peptides could be generated prior to MS analysis. Alternative digestion strategies, using a combination of proteases for the generation of specific peptide sequences may improve ASV identifications. However, the identifications of SAAVs in high quality MS/MS data sets can be easily achieved developing searchable databases that include altered amino acid sequences (Nesvizhskii et al. 2006). Genomic information of nsSNPs can be translated and incorporated into protein sequence databases that can confirm genomic based data in existing tandem mass spectra, and eventually observe novel proteoforms in biological samples.

6.3.1.1 Databases for Identification of SAAVs

The first reported database designed for mutant protein identification was based on the International Protein Sequence (IPI) database that was widely used, holding some 70,000 human protein entries (Schandorff et al. 2007). The inclusion of SAAVs and single amino acid conflicts reported in the SwissProt databases posed the problem of increased sequence redundancy that could eventually compromise the confidence of identifications. Therefore, sequential variations of proteins were attached to the consensus protein sequences with an addition of letter “J” (as a “spacer” to recognize the extra information in entries) between each peptide with the mutation site flanked with a tryptic peptide at both ends. The resulting MSIPI database was completed with a decoy database and published together with each new IPI release by EBI until its final version (v3.67) that held 87,062 human protein entries.

Alternatively, another protein database (K-SNPdb) was created with 125,622 tryptic peptides, which included sequences of the altered amino acids (Bunger et al. 2007). The construction was based on filtering the NCBI dbSNP for nsSNPs (Sherry et al. 2001) that exceeded ten million SNPs throughout the entire human genome. The number of nsSNPs out of all coding region SNPs was about 65,000. In addition to the filtering, manual allocation with the protein accession numbers, the location of nsSNPs and the amino acid changes were derived from the NCBI protein database, creating a fasta file with paired reference and alternative alleles. In order to improve the false discovery rate, a decoy database (FalseSNPdb) with the same number of peptides and identical masses to the peptides of K-SNPdb was composed from IPI entries. This strategy granted the identification of 629 SAAVs, of which 36 were not present in the protein databases of NCBI and IPI.

In an attempt to collect comprehensive sequential data about human mutant proteins, in particular about those involved in cancer, oncogenesis and tumor progression, a novel database (CanProVar) was created (Li et al. 2010). Information on protein variations from public resources, including the Human Proteome Initiative (HPI) (O’Donovan et al. 2001), the Catalogue of Somatic Mutations in Cancer (COSMIC) (Bamford et al. 2004), the Online Mendelian Inheritance in Man (OMIM) (Hamosh et al. 2005), the Cancer Genome Atlas (TCGA) (Comprehensive genomic characterization defines human glioblastoma genes and core pathways 2008) and two large-scale cancer genome sequencing studies (Greenman et al. 2007; Sjoblom et al. 2006), were integrated with a special recognition of cancer related variations (crVAR). The final version of CanProVar holds 41,541 non-cancer specific and 11,445 cancer related variations (http://bioinfo.vanderbuilt.edu/canprovar/). Most importantly, this collection of human mutant proteins is searchable on-line, offering extremely useful data linked to cancer samples, additional data sources, publications along with functional information on gene ontology annotations and interaction partners (Fig. 6.2).

As an outcome of the CanProVar project, an effective identification workflow with multiple search engine options and a tool designed for the correction of the false discovery rate (FDR), was proposed and demonstrated using CRC cell line data (Li et al. 2011). Notably, this downloadable MS-CanProVar database was completed with the Ensemble protein database (v53), 148

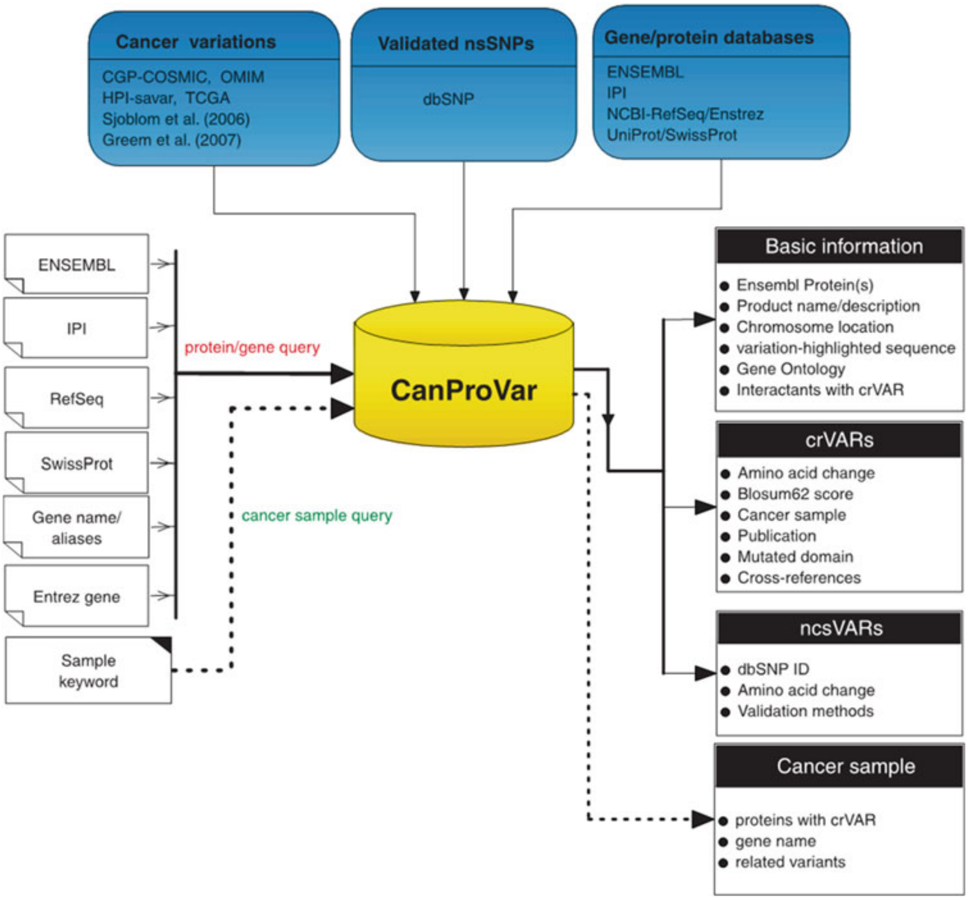

Fig. 6.2 The system architecture of CanProVar protein database that provides a large number of SAAV sequences (Reproduced with permission from Li et al. 2010) Copyright (2010) John Wiley and Sons

contaminant sequences and their reversed sequences as decoys (total searchable entry number is 290,440). Missense variations, non-sense variations and single amino acid deletions and insertions were all included in the database. To address the increasing redundancy posed by the inclusion of mutant sequences, they were shortened to a tryptic peptide with the mutation site flanked by two peptides. As a result, the addition of these new peptides increased the databased size by only 3.4 %.

The MS-CanProVar database formed the basis of a recent approach that combined it with the unique variant sequences in UniProt human polymorphisms (release 2011-12-14), resulting in a total of 87,745 SAAVs (Song et al. 2014). The Swiss-CanSAAVs database contains a total of 161,747 downloadable entries (http://bioanalysis.dicp.ac.cn/proteomics/Publications/SSAV/SAAV-Database.htm) with minimized redundancy using only tryptic mutant peptide sequences flanked by two additional peptides. A customized database, the Human Protein Mutant Database (HPMD), was also created by extracting and combining sequential information of known disease mutations from OMIM, the Protein Mutant Database (PMD) (Kawabata et al. 1999), the Systematic Platform for Identifying Mutated Proteins (SysPIMP) (Xi et al. 2009) and UniProt (Magrane and Consortium 2011) (see Fig. 6.3) (Mathivanan et al. 2012). To improve the FDR, sequence redundancy was decreased by limiting the peptide sequences to 101 amino acids with the mutation site in the center position

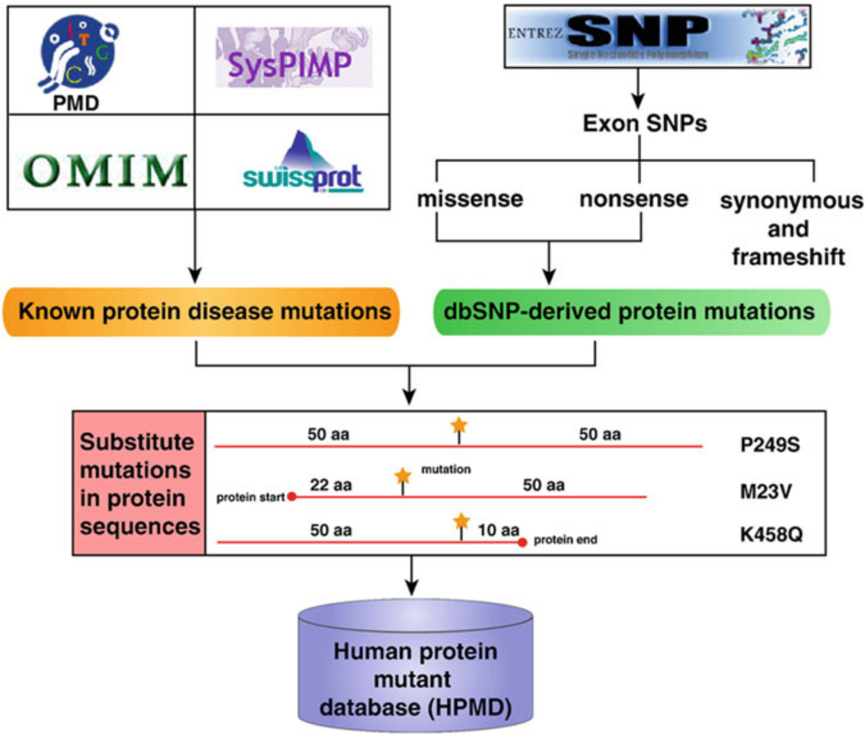
PMD
SysPIMP
OMIM
swissprot
ENTREZ SNP
Exon SNPs
missense
nonsense
synonymous and frameshift
Known protein disease mutations
dbSNP-derived protein mutations
Substitute mutations in protein sequences
50 aa
50 aa
P249S
mutation
protein start
22 aa
50 aa
M23V
50 aa
10 aa
protein end
K458Q
Human protein mutant database (HPMD)

Fig. 6.3 Construction of Human Protein Mutant Database (HPMD) for MS based protein mutation search. Schematic of the construction of HPMD is shown. Known protein disease mutations downloaded from OMIM, PMD, SysPIMP and UniProt were combined along with the missense and nsSNPs from dbSNP. The mutations were substituted in protein sequences to form peptides (maximum 101 amino acids) with mutations. The mutations were fixed to the center (51st residue) unless and until the mutation is localized close to the start or end of the protein sequence. The database composed 171,919 mutations (31,479 – known disease mutations and 140,440 – dbSNP) (Reproduced with permission from Mathivanan et al. 2012) Copyright (2012) Elsevier

(51st residue). This strategy has a drawback as it excludes the possible identification of mutant proteins by a miss-cleaved tryptic peptide.

Through the increasing access to RNA-Seq data, another path has opened to generate extended protein sequence databases, which is the translation of transcriptomic information to amino acids. The combination of shotgun proteomics with next generation sequencing (NGS) technologies has shown to be an effective approach to gain multilevel information and knowledge about cellular systems (Chen et al. 2012). To facilitate translations of RNA-Seq sequences to protein levels, an elegant bioinformatics tool was introduced recently allowing for the generation of customized protein databases (Wang and Zhang 2013). The R package of customProDB can easily create improved protein databases from RNA-Seq data with identified single nucleotide variations, short insertions and deletions as well as novel junctions between exons. The customProDB was an integrated and important part of the newly developed proteogenomic dashboard (dasHPPboard) intended to facilitate the protein mapping efforts of HPP (Tabas-Madrid et al. 2015).

6.3.2 Concept for Identification of Mutant Proteins

The strategy to identify novel mutant proteoforms in biological samples was designed using high quality shotgun proteomic tandem mass spectra for database search by existing search engine algorithms (Lichti et al. 2015). The key element of the approach was the unique set of database entries that describes all SAAV sequences, translated from known genomic studies (Ensembl). Using a custom made software tool, new protein sequences were generated to include a point mutation in each new entry that thus differed in a single amino acid from the consensus protein. The mutant protein database (MuPdb) included 2.3 million SAAVs, excluding titin (Q8WZ42). The sequence redundancy was greatly reduced keeping the tryptic peptides with the mutation site surrounded by two missed cleavages at both termini. The resulting *in silico* derived proteoforms were denoted following the neXtProt nomenclature, including the access codes but also adding information about the nature and the position of the mutation (such as NX_P07288-SNP-L-132-I).

The MuPdb was rendered as a combination of consensus (40,548 entries of UniProtKB) and mutant proteoform sequences of chromosome 19 (132,264 entries), together with 115 common contaminant sequences (cRAP) in standardized. *fasta* format, in order to be used with various search engines, including Proteome Discoverer, Mascot and PEAKS. To address the challenge that the large search space represents, a custom decoy database was created using PEAKS. Additionally, manual validation of tandem mass spectra was performed following blast searches for uniqueness. An initial identification and validation on glioblastoma stem cells (GSC) revealed many SAAVs. Interestingly, a thoroughly investigated mutation (*p.T186R*) of branched-chain aminotransferase 2 (BCAT2) was confirmed (Lichti et al. 2015). This and other newly observed SAAVs in GSC samples were further validated at the transcript level and by SRM-assays designed for suitable SAAV peptides.

This concept was generalized for the identification of SAAVs in any biological sample as presented in Fig. 6.4. Following searches of high quality MS/MS data in the custom made mutant database, the initial findings need to be verified. Currently, this step consists of both targeted proteomics and transcriptomic methods, a combination of which is sufficiently powerful to provide novel biomarkers and drug targets in future applications.

The SRM-MS analysis for verification of mutant proteoforms, targeting the most potential tryptic peptides specific to mutation sites and their corresponding wild type sequences, can also be performed. Synthetic heavy isotope labeled peptides with corresponding sequences can be spiked into the clinical samples for unambiguous identification of mutant proteoforms. In addition, to provide qualitative confirmation and quantification, the ratios between wild type and mutant forms can be determined in heterozygous expression.

Furthermore, RNA sequence analysis can be performed with biological samples and the transcript data can be used to verify newly identified mutant proteins. This approach can provide considerable reference information and verify SAAVs. The quantitative readout of RNA-Seq results can also be correlated with observed levels of expressed mutant proteins for confirmation.

6.3.3 Targeted Proteomics of SAAVs

One of the most popular mass spectrometry technologies, selected reaction monitoring (SRM), can be successfully applied to identify and quantify specific peptides within the digested samples of complex mixtures (Feng and Picotti 2016). In addition, the SRM methodology is inherently easy to multiplex, allowing for the development of multiple protein assays that offer high sensitivity and throughput. When applying the stable isotope technology, uniformly ^{13}C-^{15}N-labeled proteins can be quantified in blood plasma at levels of 100 ng/mL. However, in many cases, additional enrichment steps are required for the identification of proteins present at lower concentrations in human samples like plasma or serum. Targeted enrichment, with or without antibodies, has been introduced to improve the detection sensitivity.

Recently, several approaches combining immunoaffinity with SRM using stable isotope peptides, such as SISCAPA-SRM (Anderson et al. 2004), immuno-SRM (Whiteaker et al. 2007; Whiteaker et al. 2011), and mass spectro-

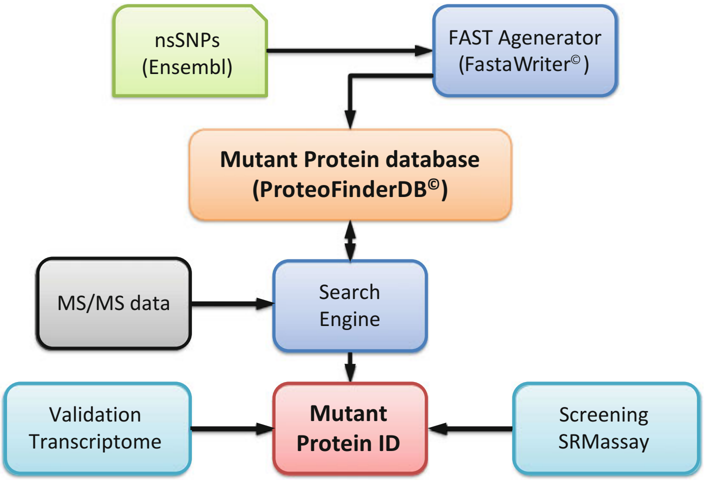

Fig. 6.4 The general concept of identification of SAAVs by tandem mass spectra searching a specialized protein database containing more than two million SAAV sequences

metric immunoassay (MSIA) (Lopez et al. 2010), have significantly improved the limit of detection of low abundant protein biomarkers present in plasma. Using the MSIA method, the lowest detection level (LOQ) of plasma proteins is 16–31 pg/ml (Lopez et al. 2010). Immunoprecipitation (IP) is a logical strategy to enrich mutant proteoforms in combination with targeted proteomics techniques, such as SRM, because antibodies can be generated against most proteins of interest and SRM-MS does not require absolute specificity for the antigens or for the mutations of interest. Additionally, IP can remove the most abundant proteins, including cytoskeletal proteins, immunoglobulins, and serum albumin from biological samples (Anderson et al. 2004).

Since antibodies are not always available and can be expensive to develop, antibody-free enrichment of target proteins was recently demonstrated for the quantitation of low abundant plasma proteins at concentrations in the 50–100 pg/mL range (Shi et al. 2012).

6.3.4 Quantifications of SAAVs

The SRM technology offers precise and efficient quantifications of known proteins, targeting their characterized proteotypic peptides in biological samples. Utilizing the high specificity of SRM-MS, multiplexing can be easily achieved, providing a relatively high throughput methodology. However, the requirement of isotope labeled peptide standards for relative or absolute quantification may constrain this approach, considering the difficulties to synthesize certain peptide sequences, the related costs may also be a limitation.

Nevertheless, quantitative analysis of mutant proteoforms in studies of disease progression can certainly provide important insights into the ratio of allele-specific gene expressions, which has been shown to be closely associated with variations in individuals (Yan et al. 2002; Montgomery et al. 2010). It has been also demonstrated that a SAAV of a single allele could in fact be expressed in either a homozygous or heterozygous manner (Végvári et al. 2013). The frequency of the mutant prostate specific antigen (PSA_*p.L132I*) form agreed well with population based genomic data, although more systematic and large-scale studies are required to understand how universal this finding is (Fig. 6.5). There are indications that disease progression may be monitored by the level of mutant proteoforms in heterozygous expressions, which can improve our understanding of, for instance, cancer biology.

An encouraging investigation was designed to utilize SRM based targeted proteomics for detection and quantification of selected SAAVs in 290 clinical plasma samples collected from Asian patients with both obesity and diabetes (Su et al.

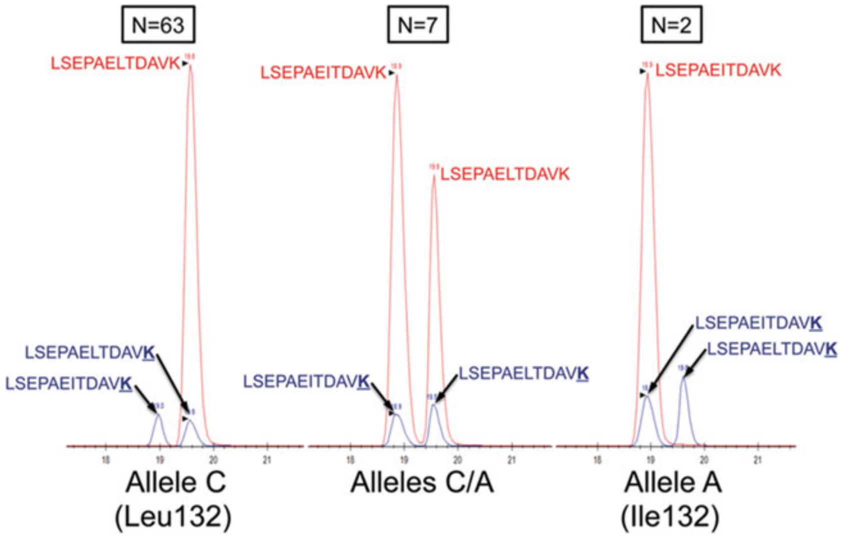

Fig. 6.5 Detection of three possible combinations of allele expressions in examples of SRM-MS analyses in clinical samples. Endogenous signals of LSEPAELTDAVK and LSEPAEITDAVK are shown in *red*, and their corresponding heavy-isotope labeled internal standard signals are in *blue* (Reproduced with permission from Végvári et al. 2013)

2011). Following initial identifications of SAAVs in patient groups, including healthy controls, key proteotypic tryptic peptides were monitored and their corresponding complement component proteins (C7, factor H and C5) were determined by absolute quantification. The results indicated that the homozygous expressions of wild type C7_*p.T587* and factor H_*p.V62* were over represented in the selected Asian patient groups, while the C7_*p.T587P* was significantly higher in control samples. Additionally, no homozygous expression was found in any individuals, which agreed well with previous studies (Gimelbrant et al. 2007). Similarly, heterozygous expression profiles of these SAAVs were determined with sufficient accuracy. This study has proven the novel concept of SRM-based quantitative analysis, which indicated that the levels of heterozygous and homozygous SAAVs in patient populations have significant associations with certain disease traits.

A cost effective quantification approach was developed for the large-scale study of SAAVs in clinical samples (Song et al. 2014). The stable isotope dimethyl-labeling methodology (Kovanich et al. 2012) could be adopted to quantify a total of 282 unique SAAV peptides in combined CID and HCD tandem mass spectra (Fig. 6.6). Because leucine (Leu) and isoleucine (Ile) are isobaric, SAAVs with altered Leu or Ile were excluded in the final results. The initial identification of SAAVs was performed using searches of mass spectra against a custom made protein database, holding 87,745 amino acid variant sequences and 73,910 UniProt canonical protein entries (Swiss-CanSAAVs). Notably, the mutant sequences were shortened to a tryptic peptide with two missed cleavage at both ends in order to reduce sequential redundancy and thus improve the false discovery rate (FDR). Furthermore, the Swiss-CanSAAVs database was concatenated with the reversed sequences allowing for FDR analysis.

6.4 Applications and Their Biological Findings

Up to today, little attention has been paid to the functional link between mutant proteins and diseases (Wang et al. 2011). Cancer research has recently found that solid tumors typically produce 20–100 mutant genes with non-synonymous

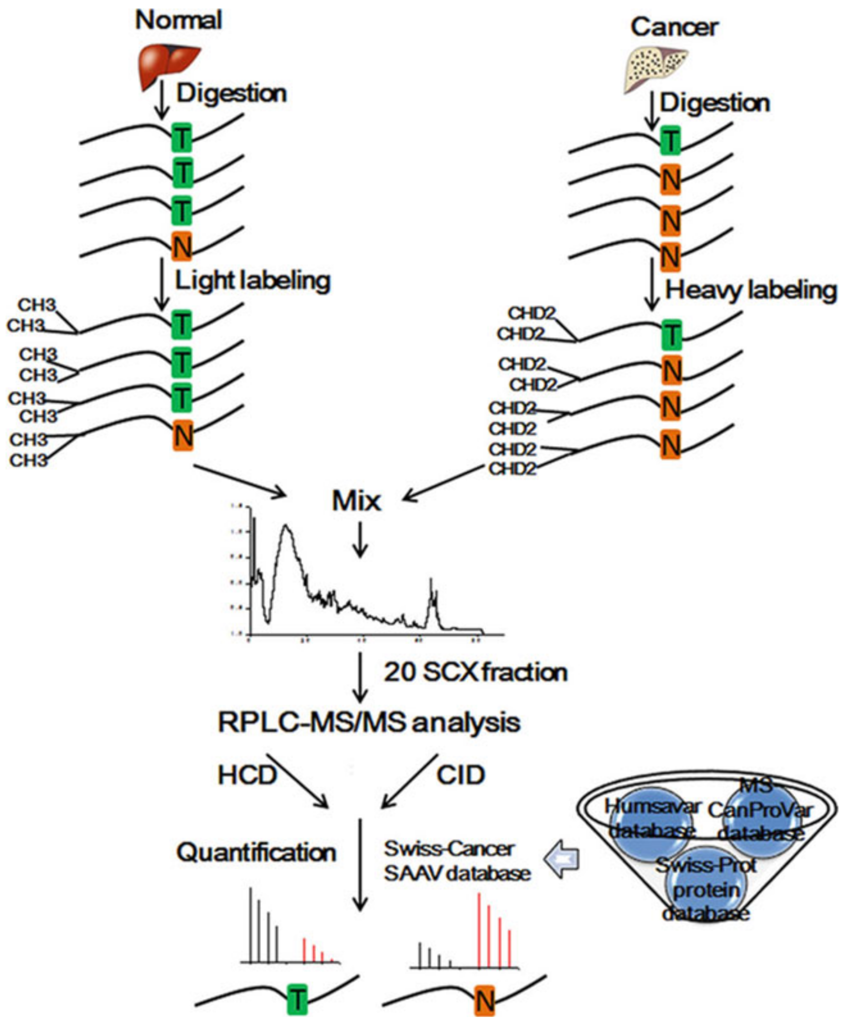

Fig. 6.6 Workflow for the large-scale quantitative analysis of SAAVs between hepatocellular carcinoma and normal human liver tissues (Reproduced with permission from Song et al. 2014. Copyright (2014) American Chemical Society)

alterations, as DNA-based sequencing studies have revealed (Wood et al. 2007). Recent studies on certain cancer forms have shown that mutant proteins can be associated with disease and even be the cause of disease (Bozic et al. 2010; Haber and Settleman 2007). Based on the occurrence and biological function of these mutant proteins, two classes were suggested:

1. “Drivers” that can initiate and are responsible for tumor genesis
2. “Passengers” are not directly associated with malignant differentiation (Haber and Settleman 2007; Bozic et al. 2010).

Importantly, the altered genetic codes of nsSNPs were found to be tightly associated with physiological and pathological traits of individuals (Sun et al. 2008). Additionally, the ratio of allele-specific gene expressions in heterozygous state is also associated with various traits of individuals (Montgomery et al. 2010) and the quantitative relationship of the wild type/mutant proteins can also indicate disease traits (Yan et al. 2002; Montgomery et al. 2010). Consequently, both qualitative and quantitative data about the structures and the functional proteins of individuals with SAAVs are required for comprehensive analysis.

Identification and quantification of mutant proteoforms were performed searching tandem mass spectra against a custom database that consisted of 87,745 amino acid variant sequences and all canonical protein entries of UniProtKB (Song et al. 2014). The approach was applied on profiling mutant proteomes in hepatocellular carcinoma (HCC) and healthy human tissue samples identifying 282 unique SAAV sites. Importantly, a significant increase of carbamoyl phosphate synthase (CPS1) *p.T1406N* and HIV-1 TAT-interactive protein 2 (HTATIP2) *p.S197R* mutations were quantified in HCC samples, which could be associated with cancer progression (Song et al. 2014). Similarly significant alteration of mutant proteomes was detected in serum samples from patients with pancreatic cancer and quantified using a isobaric labeling method (Nie et al. 2014). As a result, a novel biomarker panel was suggested, including α-1-antichymotrypsin (AACT), thrombospondin-1 (THBS1) and a mutant form of serotransferrin (TF_*p.V448I*), that could differentiate pancreatic cancer from healthy controls and chronic pancreatitis.

6.5 Future Perspective

Many genomic studies have produced a large amount of high quality data originating from population wide investigations. While the association of genes with certain diseases, identifying germline and somatic mutations, is very useful, the actual expression profiles of their wild type and mutant products is at least as important, given that proteins are the functional players in biology. Because the altered biology of cells, characteristic of disordered progresses, is driven by proteins, functional data should be generated by taking snapshots of expression profiles in healthy and patient samples. MS-based proteomics provides a unique tool to assess the expression profiles of mutant proteins in body fluids and tissue samples, which can be identified as lead candidates of optimal disease biomarkers. The qualitative and quantitative analyses of these proteoforms could thus result in novel diagnostic and prognostic values.

The fact that SAAVs can be identified in tandem mass data by unique peptide sequences, which are absent from typical protein databases, makes their observation difficult to confirm. Theoretically, certain SAAV and also ASV specific peptides may be identical with tryptic sequences of other consensus proteins. Additionally, mutant proteins can have altered biological activities, making these proteoforms functionally new molecules. It may be suggested that the definition of proteins might be improved with a more functional view. Of course, as the identification is strictly based on structural information, such a new protein definition would not be directly supportive if databases are confined to consensus sequences only. However, a more progressive view and protein definition should facilitate the development of identification strategies to identify mutant proteins in large-scale clinical studies.

References

Anderson, N. L., Anderson, N. G., Haines, L. R., Hardie, D. B., Olafson, R. W., & Pearson, T. W. (2004). Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *Journal of Proteome Research, 3*(2), 235–244.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., & Wooster, R. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer, 91*(2), 355–358.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., & Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences, 107*(43), 18545–18550.

Bunger, M. K., Cargile, B. J., Sevinsky, J. R., Deyanova, E., Na, Y., Hendrickson, R. C., & Stephenson, J. L. (2007). Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *Journal of Proteome Research, 6*, 2331–2340.

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y. K., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D., Euskirchen, G., Lacroute, P., Bettinger, K., Boyle, A. P., Kasowski, M., Grubert, F., Seki, S., Garcia, M., Whirl-Carrillo, M., Gallardo, M., Blasco, M. A., Greenberg, P. L., Snyder, P., Klein, T. E., Altman, R. B., Butte, A. J., Ashley, E. A., Gerstein, M., Nadeau, K. C., Tang, H., & Snyder, M. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell, 148*(6), 1293–1307.

Comprehensive genomic characterization defines human glioblastoma genes and core pathways (2008). *Nature, 455*(7216), 1061–1068.

Feng, Y., & Picotti, P. (2016). Selected reaction monitoring to measure proteins of interest in complex samples: A practical guide. *Methods in Molecular Biology (Clifton, NJ), 1394*, 43–56.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449 (7164):851–861.

Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: Recommendation statement. (2005). Annals of internal medicine 143 (5):355–361

Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., & Chess, A. (2007). Widespread monoallelic expression on human autosomes. *Science, 318*(5853), 1136–1140.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y. E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M. H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., & Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153–158.

Haber, D. A., & Settleman, J. (2007). Cancer: Drivers and passengers. *Nature, 446*(7132), 145–146.

Hammarström, P., Wiseman, R. L., Powers, E. T., & Kelly, J. W. (2003). Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science (New York, NY), 299*(5607), 713–716.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 33*(Database issue), D514–517.

Kawabata, T., Ota, M., & Nishikawa, K. (1999). The protein mutant database. *Nucleic Acids Research, 27*(1), 355–357.

Kovanich, D., Cappadona, S., Raijmakers, R., Mohammed, S., Scholten, A., & Heck, A. J. (2012). Applications of stable isotope dimethyl labeling in quantitative proteomics. *Analytical and Bioanalytical Chemistry, 404*(4), 991–1009.

Li, J., Duncan, D. T., & Zhang, B. (2010). CanProVar: A human cancer proteome variation database. *Human Mutation, 31*(3), 219–228.

Li, J., Su, Z., Ma, Z. Q., Slebos, R. J. C., Halvey, P., Tabb, D. L., Liebler, D. C., Pao, W., & Zhang, B. (2011). A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Molecular & Cellular Proteomics, 10*(5), M110.006536–M006110.006536.

Lichti, C. F., Mostovenko, E., Wadsworth, P. A., Lynch, G. C., Pettitt, B. M., Sulman, E. P., Wang, Q., Lang, F. F., Rezeli, M., Marko-Varga, G., Végvári, Á., & Nilsson, C. L. (2015). Systematic identification of single amino acid variants in Glioma stem-cell-derived chromosome 19 proteins. *Journal of Proteome Research, 14*(2), 778–786.

Lopez, M. F., Rezai, T., Sarracino, D. A., Prakash, A., Krastins, B., Athanas, M., Singh, R. J., Barnidge, D. R., Oran, P., Borges, C., & Nelson, R. W. (2010). Selected reaction monitoring-mass spectrometric immunoassay responsive to parathyroid hormone and related variants. *Clinical Chemistry, 56*(2), 281–290.

Magrane, M., & Consortium, U. (2011). UniProt knowledgebase: A hub of integrated protein data. *Database* 2011, bar009.

Mathivanan, S., Ji, H., Tauro, B. J., Chen, Y.-S, & Simpson, R. J. (2012). Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *Journal of Proteomics, 76*, 141–149.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., & Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature, 464*(7289), 773–777.

Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., Baginsky, S., & Aebersold, R. (2006). Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Molecular & Cellular Proteomics, 5*(4), 652–670.

Nie, S., Yin, H., Tan, Z., Anderson, M. A., Ruffin, M. T., Simeone, D. M., & Lubman, D. M. (2014). Quantitative analysis of single amino acid variant peptides associated with pancreatic cancer in serum by an isobaric labeling quantitative method. *Journal of Proteome Research, 13*(12), 6058–6066.

Nørregaard Jensen, O. (2004). Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology, 8*(1), 33–41.

O'Donovan, C., Apweiler, R., & Bairoch, A. (2001). The human proteomics initiative (HPI). *Trends in Biotechnology, 19*(5), 178–181.

Omenn, G. S., Lane, L., Lundberg, E. K., Beavis, R. C., Nesvizhskii, A. I., & Deutsch, E. W. (2015). Metrics for the Human Proteome Project 2015: Progress on the human proteome and guidelines for high-confidence protein identification. *Journal of Proteome Research, 14*(9), 3452–3460.

Paik, Y. K., Omenn, G. S., Uhlen, M., Hanash, S., Marko-Varga, G., Aebersold, R., Bairoch, A., Yamamoto, T., Legrain, P., Lee, H. J., Na, K., Jeong, S. K., He, F., Binz, P. A., Nishimura, T., Keown, P., Baker, M. S., Yoo, J. S., Garin, J., Archakov, A., Bergeron, J., Salekdeh, G. H., & Hancock, W. S. (2012). Standard guidelines for the chromosome-centric human proteome project. *Journal of Proteome Research, 11*(4), 2005–2013.

Pandey, A., & Pevzner, P. A. (2014). Proteogenomics. *Proteomics, 14*(23–24), 2631–2632.

Salisbury, B. A., Pungliya, M., Choi, J. Y., Jiang, R., Sun, X. J., & Stephens, J. C. (2003). SNP and haplotype variation in the human genome. *Mutation Research, 526*, 53–61.

Schandorff, S., Olsen, J. V., Bunkenborg, J., Blagoev, B., Zhang, Y., Andersen, J. S., & Mann, M. (2007). A mass spectrometry-friendly database for cSNP identification. *Nature Methods, 4*, 465–466.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research, 29*, 308–311.

Shi, T., Fillmore, T. L., Sun, X., Zhao, R., Schepmoes, A. A., Hossain, M., Xie, F., Wu, S., Kim, J. S., Jones, N., Moore, R. J., Pasa-Tolic, L., Kagan, J., Rodland, K. D., Liu, T., Tang, K., Camp, D. G., 2nd, Smith, R. D., & Qian, W. J. (2012). Antibody-free, targeted mass-spectrometric approach for quantification of proteins at low picogram per milliliter levels in human plasma/serum. *Proceedings of the National Academy of Sciences of the United States of America, 109*(38), 15395–15400.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., & Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science, 314*(5797), 268–274.

Song, C., Wang, F., Cheng, K., Wei, X., Bian, Y., Wang, K., Tan, Y., Wang, H., Ye, M., & Zou, H. (2014). Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. *Journal of Proteome Research, 13*(1), 241–248.

Sordella, R., Bell, D. W., Haber, D. A., & Settleman, J. (2004). Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. *Science, 305*(5687), 1163–1167.

Su, Z. D., Sun, L., Yu, D. X., Li, R. X., Li, H. X., Yu, Z. J., Sheng, Q. H., Lin, X., Zeng, R., & Wu, J. R. (2011). Quantitative detection of single amino acid polymorphisms by targeted proteomics. *Journal of Molecular Cell Biology, 3*(5), 309–315.

Sun, T., Zhou, Y., Yang, M., Hu, Z., Tan, W., Han, X., Shi, Y., Yao, J., Guo, Y., Yu, D., Tian, T., Zhou, X., Shen, H., & Lin, D. (2008). Functional genetic variations in cytotoxic T-lymphocyte antigen 4 and susceptibility to multiple types of cancer. *Cancer Research, 68*(17), 7025–7034.

Tabas-Madrid, D., Alves-Cruzeiro, J., Segura, V., Guruceaga, E., Vialas, V., Prieto, G., Garcia, C., Corrales, F. J., Albar, J. P., & Pascual-Montano, A. (2015). Proteogenomics dashboard for the human proteome project. *Journal of Proteome Research, 14*(9), 3738–3749.

Taylor, R. W., & Turnbull, D. M. (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics, 6*(5), 389–402.

Végvári, Á., Sjödin, K., Rezeli, M., Malm, J., Lilja, H., Laurell, T., & Marko-Varga, G. (2013). Identification of a novel proteoform of prostate specific antigen (SNP-L132I) in clinical samples by multiple reaction monitoring. *Molecular & Cellular Proteomics, 12*(10), 2761–2773.

Wang, X., & Zhang, B. (2013). CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics (Oxford, UK), 29*(24), 3235–3237.

Wang, Q., Chaerkady, R., Wu, J., Hwang, H. J., Papadopoulos, N., Kopelovich, L., Maitra, A., Matthaei, H., Eshleman, J. R., Hruban, R. H., Kinzler, K. W., Pandey, A., & Vogelstein, B. (2011). Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences, 108*(6), 2444–2449.

Whiteaker, J. R., Zhao, L., Zhang, H. Y., Feng, L. C., Piening, B. D., Anderson, L., & Paulovich, A. G. (2007). Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. *Analytical Biochemistry, 362*(1), 44–54.

Whiteaker, J. R., Zhao, L., Abbatiello, S. E., Burgess, M., Kuhn, E., Lin, C., Pope, M. E., Razavi, M., Anderson, N. L., Pearson, T. W., Carr, S. A., & Paulovich, A. G. (2011). Evaluation of large scale quantitative proteomic assay development using peptide affinity-based mass spectrometry. *Molecular & Cellular Proteomics, 10*(4), M110.005645.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., & Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science (New York, NY), 318*(5853), 1108–1113.

Xi, H., Park, J., Ding, G., Lee, Y. H., & Li, Y. (2009). SysPIMP: The web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. *Nucleic Acids Research, 37*(Database issue), D913–920.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science, 297*(5584), 1143.