Clinical implications of recent advances in proteogenomics

Marie Locard-Paulet, Olivier Pible, Anne Gonzalez de Peredo, Béatrice Alpha-Bazin, Christine Almunia, Odile Burlet-Schiltz, J. Armengaud

▶ **To cite this version:**

Marie Locard-Paulet, Olivier Pible, Anne Gonzalez de Peredo, Béatrice Alpha-Bazin, Christine Almunia, et al.. Clinical implications of recent advances in proteogenomics. Expert Review of Proteomics, 2016, 13 (2), pp.185-199. 10.1586/14789450.2016.1132169 . hal-03080146

HAL Id: hal-03080146
https://hal.science/hal-03080146v1

Submitted on 19 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Publisher: Taylor & Francis

Journal: *Expert Review of Proteomics*

DOI: 10.1586/14789450.2016.1132169

Review

Clinical implications of recent advances in proteogenomics

Marie Locard-Paulet[1,2], Olivier Pible[3], Anne Gonzalez de Peredo[1,2], Béatrice Alpha-Bazin[3], Christine Almunia[3], Odile Burlet-Schiltz[1,2], Jean Armengaud[3*]

[1]CNRS, IPBS (*Institut de Pharmacologie et Biologie Structurale), 205 route de Narbonne, 31077 Toulouse, France.*

[2]*Université de Toulouse, UPS, IPBS, 31077 Toulouse, France.*

[3]*CEA-Marcoule, DSV/IBITEC-S/SPI/Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200 Bagnols-sur-Cèze, France.*

**Author to whom correspondence should be addressed: Jean Armengaud, CEA-Marcoule, DSV/IBITEC-S/SPI/Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200 Bagnols-sur-Cèze, France; jean.armengaud@cea.fr; Tel: +00 33 4 66 79 68 02; Fax: +00 33 4 66 79 19 05.*

Running title: Clinical proteogenomics

ACCEPTED MANUSCRIPT

Abstract

Proteogenomics, the alliance of proteomics, transcriptomics, genomics and bioinformatics, was first proposed for refining genome annotation using experimental data acquired on gene products. With high-throughput analysis of proteins made possible with next-generation tandem mass spectrometers, proteogenomics is greatly improving human genome annotation *per se*, and is helping to decrypt the numerous gene and protein modifications occurring during development, aging, illness and cancer progression. Further efforts are required to obtain a comprehensive picture of human genes, their products, functions, and drift over time or in reaction to microbiota and pathogen stimuli. This should be performed not only to obtain a general overview of the human population, but also to gain specific information at the individual level. This review focuses on the clinical implications of proteogenomics: novel biological insights into fundamental biology, better characterization of pathogens and parasites, discovery of novel diagnostic approaches for cancer, and personalized medicine.

Keywords:

Proteogenomics, Cancer, Onco-proteogenomics, Shotgun proteomics, Pathogens, Virulence factors, Personalized medicine

ACCEPTED MANUSCRIPT

1. Introduction.

The numerous applications of genomics in human health have led to considerable changes in clinical laboratories, where quantitative polymerase chain reaction and next-generation sequencing facilities are now widespread. Mass spectrometry-based analytical tools have been used in clinics for a long time, mainly for the identification and quantification of small molecules related to toxicology and endocrinology [1]. A novel increase in interest for such approaches can be explained by the unrivaled discriminative performance of mass spectrometry, granting the higher selectivity required for protein-based assays. While antibody-based detection of proteins remains widespread, this is restricted by the availability of antibodies and their limits in terms of specificity. Novel strategies for preparing medical samples and acquiring and interpreting mass spectrometry data are being developed leading to the identification and quantification of distinct proteoforms through specific and sensitive assays [2]. As an example, the evolution of diagnostics in microbiology over the past five years has been impressive, with the possibility to identify pathogenic bacteria within a few minutes using matrix-assisted laser desorption/ionization – time of flight (MALDI-TOF) mass spectrometry [3]. This approach is based on the measurement of the masses of low molecular weight, basic and abundant proteins from pathogens previously isolated on agar plates or grown in liquid blood cultures. These recorded mass profiles can then be matched against species-specific profiles present in curated databases. Such technical advances have contributed to the demonstration of the simplicity and advantages of using mass spectrometry-based approaches in the development of diagnostics, paving the way for more use of mass spectrometry in clinics.

With regard to diagnostics, omics approaches may be considered redundant in terms of information or even antagonist technologies, as they require specific instruments and

ACCEPTED MANUSCRIPT

expertise. However, concerning discovery and validation, omics are highly complementary [4]. Here, we describe how the alliance of proteomics, transcriptomics, genomics, and bioinformatics contributes to the elucidation of novel biological insights into fundamental biology, a better characterization of pathogens, and the discovery of novel diagnostic approaches for cancer. Although this alliance, named proteogenomics, was first devoted to improving genome annotation [5-9], novel applications have rapidly flourished. In the present review, we focus on the clinical implications of the most recent advances in proteogenomics.

2. Proteogenomics comes to the clinical scene.

Proteomics is heavily dependent on protein sequence databases. A cornerstone concept in biology is that genetic information together with its dynamic expression through transcription and protein regulation result in cellular behaviors and pathologies. Hence, human gene sequences and their environments have been studied for many years. The human genome was established as a result of worldwide collaborations [10,11]. Its analysis shed light on several genetic causes of disease and contributed to the improvement of our knowledge regarding cell defense systems and organismal responses to injury, toxic compounds and pathogens. Genome annotation is the result of sequencing data analysis, based on knowledge of proteins collected over time, and the basic transcription and translation rules established several decades ago. With the predicted protein coding sequences to hand, bottom-up proteomics can be applied on a genome-wide scale. This approach basically consists of digesting proteins into smaller peptides, measuring their precise masses together with the masses of their fragmented products by tandem mass spectrometry, and then matching these lists of parent and fragment peptide masses to theoretical protein sequences. Novel genomes can now be quickly

ACCEPTED MANUSCRIPT

sequenced and are automatically annotated through bioinformatic pipelines that apply basic translation rules, as well as the conservation of protein sequences in evolutionarily related organisms. However, genome annotation with such pipelines is far from perfect, and is still subject to debate [12].

The limits of genomics, transcriptomics and proteomics. Errors in genome annotation can result in overcalling or missing open reading frames, bad prediction of numerous translational starts, and splicing errors (Armengaud J, 2009). These errors can subsequently lead to wrong protein identifications in mass spectrometry analyses. Several studies have shown the incorrect annotation of a significant number (10–20%) of N-termini from annotated bacterial polypeptides [13]. As the gene-calling task in eukaryotes is more complex than in these microorganisms because of introns, an even higher rate of annotation errors can be predicted for complex biologic species. Indeed, in protozoan parasites such as *Toxoplasma gondii* and *Neospora caninum*, the assessment of splicing and alternative splicing of mRNA remains a daunting task [14]. Moreover, it has been shown experimentally that conservation of N-terminal protein sequences derived from comparative genomics does not always apply [15]. In fact, the specific traits of a given strain or subspecies should be explained at the molecular level by both protein sequence specificities and abundance changes. Sequence polymorphisms and mutations, gene duplications, acquisition of novel protein sequences, N-terminal polypeptide changes and splicing differences may all contribute to protein sequence differences between organisms. In addition, reversible variations in expression are also involved, such as those associated with chromosome spatial organization, which is currently being explored through methods derived from the chromosome conformation capture (3C) technique [16,17], and also post-translational modifications. All of these events are responsible for the presence of not yet described proteoforms that define subtle phenotypic differences between individuals [2]. As a result, direct derivation of gene structure from a

ACCEPTED MANUSCRIPT

closely related genome may be a source of mistakes. It thus becomes clear that DNA- and RNA-derived data should be combined with proteomic analysis in order to obtain a more comprehensive picture of molecular mechanisms.

Combining omics strengths with proteogenomics. Two decades ago, a pioneering idea was proposed: mining proteome data in order to correct genome annotation [18,19]. Peptide sequences certified by mass spectrometry could be mapped back onto the genome sequence to correct translational starts, firmly establish splicing events, or highlight the presence of unannotated protein coding sequences. This proteogenomic mapping was introduced by Jaffe et al. [7]. With the amazing technical evolution of tandem mass spectrometry, the throughput of proteomics has become adequate for large genome-wide surveys [20], allowing the proteogenomic mapping of numerous bacteria, archaea, protozoan parasites, fungi, plants, and animals. This strategy has also proved useful to improve the annotation of the human genome. However, even with the newest mass spectrometry tools, only the most abundant components of the so-called "expressed proteome" are detected. Because of a proteome sequence coverage that remains low, the approach is not yet sufficiently comprehensive. Over time, proteogenomics has evolved, taking advantage of RNA sequencing (RNAseq) and advanced sample preparation [21,22]. The presence of introns in eukaryotic genomes introduces an additional level of difficulty when predicting protein sequences. This issue has partially been circumvented by mapping splicing events from mRNA. Indeed, sequencing cDNA obtained from mature mRNA transcripts is directly informative. This has become a routine procedure for non-model organisms for which the genome has not yet been established [23-25] or is extremely difficult to annotate [26]. Moreover, proteogenomics can also be applied to the interpretation of proteomic data from a draft genome or transcriptome, without the need for a time-consuming genome annotation. In this case, the proteogenomic approach allows for a quick focus on the key proteins of a given metabolic or physiologic trait highlighted by the

ACCEPTED MANUSCRIPT

comparison of samples [27]. As depicted in Figure 1, proteomics can now be intimately combined with genomics and transcriptomics, together with the help of advanced bioinformatics, in order to get a more precise picture of molecular processes and diseases. This approach allows for the analysis of protein dynamics, providing information about cellular or subcellular localization and post-translational modifications. Proteogenomics has now come on to the clinical scene. The human body can be analyzed at a more comprehensive cellular and molecular level. Pathogens can be better characterized. Novel candidate biomarkers corresponding to specific proteoforms can be proposed for diagnostics. In the following chapters, we will review the different applications of proteogenomics and their perspectives.

- Insert Figure 1 here -

An expanding proteogenomic toolbox. Numerous methodological improvements in proteogenomics have already been published. For example, specific labeling of the N-termini of proteins and detection of the products by high resolution tandem mass spectrometry has been successful for systematic N-terminomics [13,28-31]. In terms of annotating the N-termini of human proteomes, novel specific datasets have been recorded [32,33], showing alternative translation initiation sites, pervasive endoproteolytic processing, and stabilization of protein fragments *in vivo* by an extensive post-translational Nα-acetylation. The integration of RNAseq and proteomic data has confirmed human coding genes [34] and contributed to the detection of gene mutations [35]. This alliance of transcriptomics and proteomics is also encompassed within the boundary of proteogenomics [36]. Protocols to enrich specific subproteomes and to improve proteome coverage by chromatography and mass spectrometry are also being developed [33]. In order to facilitate data interpretation, ribosome-protected mRNA fragments may be extracted and sequenced. This so-called “ribosome profiling” has enabled an in-depth characterization of alternative translation start sites [22]. This

ACCEPTED MANUSCRIPT

methodology has been applied to the human HCT116 cell line, for which nine previously unannotated protein products could be identified, as well as several alternatively spliced isoforms and protein variants [37]. Most of the proteins for which no experimental evidence has been recorded are membrane proteins. Kitata et al. [38] specifically explored membrane-enriched fractions of eleven non-small cell lung cancer cell lines, thereby identifying 178 previously unrecorded membrane-associated proteins. This work exemplifies the advantages of mass spectrometry approaches to broaden proteome coverage through analysis of specific subproteomes. Further evidence for the existence of a given protein may be obtained using *in vitro* transcription/translation strategies developed to devise precise conditions and parameters that are then applied for the identification of missing proteins in human cells and tissues using targeted mass spectrometry [39]. Last but not least, numerous bioinformatic pipelines or tools have become available to assist this inventory of missing proteins or splice variants by proteogenomics [40-46], while metrics for high confidence identification of "missing" proteins are just emerging [47]. Besides the common core proteome, more specific sets of proteins with highly variable sequences require dedicated efforts. For example, immunoglobulin light chain amyloidosis caused by the accumulation of clonal proteins can be probed by proteomics if peptide sequence variability is taken into consideration, because each sequence is unique to each patient [48]. A traditional database search strategy is inefficient in such a quest because of the lack of peptide sequences in reference databases comprehensively covering the variable region of immunoglobulin light chains. However, the successful strategy consisted of peptide mapping using an augmented protein database constructed from mRNAs from a cohort of Alzheimer's disease patients [48].

ACCEPTED MANUSCRIPT

3. Boosting the hunt for human missing proteins and splice variants.

The "Human Proteome Project" (HPP) conducted under the auspices of the Human Proteome Organization aims to improve the annotation of the human genome and to generate a comprehensive map of human proteins, their interactions, and their regulation [49]. Proteogenomics embraces the validation of the existence of "missing proteins", which comprise as-yet undetected protein products of annotated gene sequences as well as the detection of hitherto unannotated proteins and the correction of gene structures. A pioneering study interrogating tandem mass spectrometry data against an exhaustive theoretical protein database generated from a 6-frame translation of the entire human genome allowed the identification of previously unexpected peptides corresponding to splice variants in blood samples [50]. Amid the different projects supported by HPP, tandem mass spectrometry data from numerous laboratories were gathered in a unique database in order to get a global picture of the human proteome, which includes the identification of expressed proteins, splice variants and post-translational modifications: the PeptideAtlas compendium. Using a normalized Trans Proteomic Pipeline, the consistent analysis of mass spectrometry data from different sources allowed the mapping of at least one specific peptide to each of 11,868 SwissProt entries in 2012 [51]; 12,934 in 2013 [52]; and 14,070 in July 2015 [53]. This database has proven to be a prime tool of choice for developing targeted mass spectrometry strategies, and further confirms previously unreported peptide sequences. It also provides insights derived from sampling origins, such as which proteins are abundant in a given tissue, and at which developmental stage. Another database, Proteomics DB, enables the retrieval of proteomic information for a large set of human tissues, cell lines, and body fluids [54]. The latest one-shot extensive draft map of the human proteome validated 17,294 genes, accounting for 84% of annotated protein-coding genes [55]. Further analysis of large shotgun

ACCEPTED MANUSCRIPT

proteomic datasets resulted in a better annotation of specific human subproteomes: the mitochondrial proteome [33], hippocampus and brain proteome [56,57], liver proteome [58], and spermatozoan proteome [59], as well as several human chromosomes: chromosome 1 [60], chromosomes 2 and 14 [61], chromosome 4 [62], chromosome 7 [63], chromosome 9 [64], chromosome 12 [65,66], chromosome 16 [67], chromosome 17 [68], chromosome 18 [69,70], and chromosome 22 [71]. **Table 1** lists the main outcomes of these studies. Such a quest for missing proteins demands regular updates taking advantage of novel shotgun proteomic data [12], but also novel genome and RNAseq data. A strong focus on the annotation of disease-related proteins is extremely fruitful, as shown recently with the analysis of clinical specimens from patients with gliomas, Alzheimer's disease, and Parkinson's disease [65]. Being more precise with single nucleotide polymorphisms and splice variants is also achievable by means of proteogenomics [43,72-74].

- Insert Table 1 here -

4. Recent advances in human fundamental biology.

From gene expression to protein quantities. A broader definition of proteogenomics has emerged. Proteogenomics is not restricted to genome reannotation but can be aimed at functional analysis [21]. Most of the recent insights have been derived from a strategy coupling RNAseq-inferred protein identification with mass spectrometry quantification. Among others, an ultra-large proteomic analysis of 86 human colon and rectum tumor samples was performed using customized sequence databases from patient-matched RNAseq data [75]. This work identified numerous previously unreported single amino acid variants and allowed proteogenomic subtyping of colorectal cancer specimens. Interestingly, the

ACCEPTED MANUSCRIPT

availability of quantitative information at the protein as well as RNA levels demonstrated a modest transcript-protein correlation, thereby confirming the distinct RNA and protein quantities observed in previous measurements [76,77]. Indeed, protein quantity is a result of the balance between gene expression and protein degradation, which can be modulated by RNA editing, post-translational modifications, specific protein cleavage or degradation. Figure 2 shows the advantages of integrating different omics data regarding the numerous proteoforms present in cells and throughout the development of an individual.

- Insert Figure 2 here -

Novel proteogenomic insights into the immunopeptidome and the surfaceome. The immunopeptidome comprises all the small fragments of proteins that are displayed by cells on their surface in order to be checked by immune cells. The immune cells are in charge of discarding cells that are damaged, diseased, or contaminated by intracellular pathogens, as well as pathogens themselves, which can all be distinguished by non-self presented peptides. A novel atlas has been proposed to improve our knowledge of the immunopeptidome [78], allowing the development of improved personalized strategies for immune-based therapies. Deciphering the full repertoires of peptides bound to HLA molecules, in both health and disease, may be a daunting task [79], but novel strategies for recording and extracting mass spectrometry data make this possible. Recent work on cancer cell lines and primary cells has yielded data for an impressive number of HLA peptides, validating protein degradation as a key factor for HLA presentation in addition to identifying specific peptide sequences from a human colon cancer cell line [80]. A wider concept, the cell surface protein repertoire (the so-called surfaceome) is of pharmacological interest for improving drug addressing or defining new cancer vaccine targets. Such an atlas is currently being nourished with proteogenomics-derived information [81].

ACCEPTED MANUSCRIPT

Regulatory networks for understanding cellular mechanisms and diseases. Proteogenomics can help to define novel therapeutic targets based on a comprehensive overview of the molecular players in cells (Armengaud, 2010). As an example, high-throughput phosphoproteomics allows the analysis of signal-induced phosphoregulations that can occur in perturbed contexts such as diseases. Recently, a global analysis of kinase signaling networks upon EGF and IGF1 stimulation in models of acquired drug resistance demonstrated how the analysis of global phosphorylation responses to stimuli could provide information on cancer resistance mechanisms. Cells chronically treated with PI3K or mTORC1/2 inhibitors present different kinase responses to stimulation [82]. Importantly, a substantial advantage of mass spectrometry is its possible combination with isotopomeric labeling [83,84]. This allows the analysis of co-cultured samples through deconvolution of signals resulting from direct cell-cell interaction and can be applied to study the impact of specific microenvironments in disease development. A study comparing contact-initiated phosphosignaling in interacting EphB2- and ephrin-B1-expressing cells allowed the definition of the complex asymmetric and nonautonomous processes regulating Eph-ephrin signaling. In this case, cells were labeled with stable-isotope amino acids to differentiate phosphopeptides regulated in two interacting cell populations. The phosphorylation-based signaling of both cell types was analyzed in a system including multiple membrane receptors (as opposed to working with a recombinant ligand for stimulation), exemplifying the study of directional cell signaling in dynamic multicell population contexts [85]. These studies increase in interest if all the cellular protein players and the different proteoforms are well defined, thus highlighting the interest of applying proteogenomics whenever possible.

Omics-based personalized monitoring. Genomic and proteomic studies have traditionally relied on the use of reference databases built on the general population, thereby masking individual variations. A significant potential of proteogenomics is the ability to apply

ACCEPTED MANUSCRIPT

personalized analysis by combining DNA or RNA and proteomic data from a single patient. The first large personal omics profiling was performed on a single individual over a period of 14 months [86]. Blood samples were analyzed with a combination of genomics, transcriptomics, proteomics, metabolomics, and even an autoantibody profile was carried out, producing an invaluable dataset called the "Snyderome" [87]. Dynamic changes in molecular and biological pathways could be compared over this period, where the patient alternated between a healthy state, two viral infections and the onset of diabetes. The whole genome sequence of the subject was determined as a genomic baseline. By implementing this pilot study of personalized medicine, individual-specific single nucleotide variants and edited transcripts were shown to be translated into proteins. How gene and protein isoforms vary across different environmental conditions was described in this integrative personal omics profile, a concept that was abbreviated as "iPOP". This work highlights definitively the possible future use of proteogenomics for personalized medicine, and the benefit of complementary tools such as genomics, transcriptomics, proteomics, and metabolomics. To interpret proteomic data from a single patient without the need for deep sequencing of the genome, the use of databases compiling all polymorphism events occurring at the population level is an attractive alternative. Because proteomics brings information on the abundance of proteins, their polypeptide structure and post-translational modifications, focus on these molecular players should be a central pillar in personalized medicine.

ACCEPTED MANUSCRIPT

5. Onco-proteogenomics, the alliance of proteomics and genomics for a better understanding of cancer pathways and for novel diagnostic tools.

The application of proteogenomics in oncology began several years ago, and its success recently gave birth to a specific subfield dubbed "onco-proteogenomics" by Helmy and colleagues [88]. Numerous key studies fueled this important medical field [88-93]. The discovery of aberrant cancer-specific peptides, including products of gene fusion, odd splicing variants and novel expressed genes not present in the normal human proteome catalogue, can be performed by individual proteogenomic analysis where transcriptome and proteome data from each patient are combined. Initiatives such as the Clinical Proteomic Tumor Analysis Consortium (CPTAC, http://proteomics.cancer.gov) provide a fully integrated account of DNA, RNA, and protein abnormalities found in several tumor types [94]. The quest for biomarkers to detect early stage cancer has been a long journey, exemplified by a proteomic study of blood samples from patients with early stage ovarian cancer and of healthy women, which led to the proposal of three candidate biomarkers [95]. Of these, two were truncated forms of proteins that were only identified via the mass spectrometry approach: a truncated form of transthyretin and a cleavage fragment of the inter-α-trypsin inhibitor heavy chain H4. Proteins responsible for breast cancer susceptibility and dissemination have also been investigated with increasingly more powerful approaches, as exemplified by the latest study combining shotgun and targeted proteomics [96]. The advantages of proteogenomics to study the blood peptidome/degradome profile of cancer was illustrated with samples from breast cancer patients compared with healthy individuals [97]. A very similar set of plasma proteins was observed in the patient and control samples but strikingly different degradation patterns were observed, with 71 protein substrates degraded into 839 distinct peptides in the breast cancer patients while 50 proteins were degraded into 425 peptides in the healthy plasma

ACCEPTED MANUSCRIPT

samples. Although no major differences were observed in the cleavage specificities between the two conditions, the proteins cleaved were not the same. These changes, not identified by conventional bottom-up proteomics, potentially provided unique signatures that may be of diagnostic utility. Many studies have already demonstrated the importance of cancer-driving mutations of proteins such as RAS [98] and BRAF [99]. These have been the basis of diagnostic and cancer therapeutic developments for many years, paving the way towards targeted therapies where kinase inhibitors can reduce cancer burden. High-throughput technologies widen the window of analysis from a handful of proteins surrounding cancer drivers to large functional networks. Such studies shed light on the cellular consequences of oncogenic mutations and have shown that signaling pathways can be specifically dysregulated in defined pathological contexts. This has broadened the therapeutic options, since a pathway (or several pathways) can be targeted through many components in opposition to single mutated oncogenes or tumor suppressor genes that can be difficult to target [100]. Several proteogenomic approaches have been applied on cancer samples. Potential markers of colorectal cancer were proposed after the in-depth analysis of three human colon cancer cell lines [101]. Comparison of paired colorectal cancer and nontumorigenic tissues specified cancer-associated proteins and deregulated pathways such as focal adhesion, cytoskeleton or Rho signaling [102]. Human colon and rectum cancers were also categorized to delineate five proteomic subtypes and to prioritize candidate driver genes [75]. Advanced proteogenomics relying on proteomics and RNAseq data was able to identify multiple peptide mutations in colon cancer, as well as immunoglobulin gene variations/rearrangements, thus indicating a tumor immune response [103]. The genetic aberrations of 32 tumor antigens in invasive ductal carcinomas were documented by proteogenomic analyses, highlighting a quicker identification than with large-scale screening of genomic analyses [104]. A recent study explored the tumor microenvironment to assess exosome oncogenicity [105].

ACCEPTED MANUSCRIPT

A screen for alternative splicing via proteogenomics identified a novel protein isoform derived from cancer-related splicing variants in a highly metastatic gastric cancer cell line [106]. Proteogenomics is efficient for the identification of aberrant cancer peptides [107] and gene mutations [35] that could be neoantigens deserving further attention [108]. The system-wide approaches presented here illustrate the advantages of proteogenomics in the context of biomedical research by identifying potential biomarkers, together with providing mechanistic insights into cancer-regulated networks and cancer progression. The possible general strategies for personalized onco-proteogenomic diagnostics are illustrated in Figure 3.

- Insert Figure 3 here -

6. Revisiting human pathogens, parasites, and toxins with comprehensive proteogenomic approaches.

A better knowledge of the behavior of our age-old enemies – bacterial pathogens, viruses, and parasites – is vital to counteract them. Systems biology relies on the most comprehensive description of all parts of the system and their interactions. Proteogenomics is helpful in defining the ultimate missing pieces in terms of coding genes and proteoforms [6]. Although most of the methodological advances in proteogenomics have been proposed and optimized with nonpathogenic models [30,109-111], numerous proteogenomic studies have been performed on pathogens in order to better annotate their genomes and describe their cellular mechanisms. As a result, virulence factors and their modulation are now better understood [112,113], and a more comprehensive view of the interactions between a given pathogen and its host can be drawn [114]. The proteogenomic analysis of 15 million MS/MS spectra

ACCEPTED MANUSCRIPT

recorded from the bacterium *Yersinia pestis KIM* proteome allowed six genes to be discovered, numerous translational starts to be corrected, and an ultra-rare start codon, AUU, to be documented [115]. A key enzyme for virulence was shown to be erroneously annotated in terms of the translational start. This protein, Yersiniabactin thioesterase, participates in the biosynthesis of a siderophore, an important virulence factor involved in iron acquisition from the host. Such work was further complemented with novel data acquired on other *Y. pestis* strains [116]. The *Bacillus anthracis* and *Streptococcus pyogenes* proteogenomes were also probed through a large survey [117], as well as those of *Shigella flexneri* [118], *Helicobacter pylori* [119], enterotoxigenic *Escherichia coli* [120] and the mycoplasma, *Spiroplasma melliferum* [121]. The genomes of a pathogenic yeast, *Cryptococcus neoformans* [122], and a highly opportunistic pathogen, *Candida glabrata* [123], have also been reannotated based on proteogenomic results. Because these eukaryotic pathogen genomes are rich in introns, the validation of gene models by identification of the corresponding proteins is of great interest. Proteogenomics has proven its efficiency in such tasks by validating 83% of the predicted protein-coding genes in the latest organism. Proteogenomics was also found to be useful for defining biomarkers for whole-cell MALDI-TOF identification of pathogens such as *Neisseria meningitidis* [124] and *Francisella tularensis* [125].

Genome annotation of complex protozoan parasites can also be improved by proteogenomics. This has been exemplified with *Leishmania donovani*, which is responsible for the most severe form of leishmaniasis [126-129]. This analysis resulted in a solid background for conducting further experimental studies on this highly relevant parasite model [130]. Two other protozoan parasites, *Toxoplasma gondii* and *Neospora caninum,* have been analyzed with the help of RNAseq and proteomics data [14]. These studies highlighted the importance of maintaining active curation efforts to improve genome resources. A proteogenomics

ACCEPTED MANUSCRIPT

strategy has also been used to discover and characterize a novel bioactive peptide from the venom of a marine cone snail [131], and to assess susceptibility to endotoxin in animals [132].

Expert commentary

In its *sensu stricto* definition, proteogenomics combines genomics and proteomics to improve genome annotation and discover novel proteoforms. From a wider viewpoint, proteogenomics integrates different omics-related data in order to consolidate the gene structure and function of key proteins in a given biological system, and to add new information that cannot be derived directly from DNA/RNA-based technologies, such as abundance, post-translational modifications, and protein localization. For instance, the composition of the extracellular matrix, which is a crucial factor in the behavior of cancers, should be analyzed from a proteome perspective rather than through gene expression methodologies. Proteogenomics was first successfully applied to better annotate the genomes of numerous bacterial pathogens. Over the last decade, the proteogenomics toolbox has been enriched with methods for sample preparation, as well as numerous bioinformatic algorithms to facilitate data interpretation. It has also gained throughput with the technical advances of tandem mass spectrometry and RNAseq technology. The most recent results were obtained for the human genome annotation, which has been refined by means of large-scale studies in the framework of the Human Proteome Project, as well as for the genome annotation of several important human pathogens and parasites. Better multi-omics personalized monitoring is expected with these results in hand, together with higher-level integration of data acquired from the same individual that can be obtained by means of proteogenomics. Characterization of novel proteoforms produced during cancer progression through alternative splicing or mutations has

ACCEPTED MANUSCRIPT

given rise to a new specific field, called onco-proteogenomics. Developments in this area should lead to interesting novel concepts in terms of clinical diagnosis and prognosis. These clinical implications require large investments in proteogenomics-inspired multi-omics monitoring of large cohorts.

Five-year view

The analytical potential of the current generation of mass spectrometers is considerable. In the next five years, major improvements should be proposed in tandem mass spectrometry, leading to instruments with higher sensitivity, higher dynamic range, and higher *m/z* scan speed. For example, a new fragmentation mode, EThcD, has recently been introduced, combining electron transfer dissociation (ETD) and beam-type collision-induced dissociation (HCD). Such novelty holds great promise for the sequencing of large proteins with post-translational modifications and for *de novo* sequencing [133]. Moreover, the development of new mass spectrometry acquisition modes, especially data-independent acquisition [134,135], should increase the coverage of low abundance proteoforms. A result of this technological trend will be more comprehensive proteogenomic studies, and this should reinforce scientific outcomes and medical implications. Developments in bioinformatics and statistics should take into account the giant size of datasets that will need to be analyzed, as well as the gigantic size of databases. The best experts in the new "big data" field may be attracted by such a challenge, where most concepts are still to be developed in order to allow and speed up data interpretation. To this aim, removing the noisiest or least informative data from the datasets and databases is probably a winning strategy. Tremendous efforts have been already made to better annotate the human genome, but to date almost two proteins out of ten predicted remain

ACCEPTED MANUSCRIPT

missing. Exploiting the current tandem mass spectrometry datasets has already provided some interesting refinements in terms of missing proteins and splice variants. However, the current version of the human proteome is far from perfect, with numerous protein-coding genes still hypothetical. Apart from the identification of crystal-clear evidence of their existence, the documentation and cataloguing of functional and quantitative data for the corresponding gene products will bring some clues about function. The next challenge is to integrate in proteogenomics a new stage of molecular information such as metabolomics [136] or subcellular microscopic localization.

Novel, challenging proteogenomics projects for the personalized monitoring of individuals throughout life could be proposed. In this case, the main challenge will be that of informatics for the storage and treatment of data throughout an individual's lifespan, acquired with heterogeneous methodologies, given that technological improvements will occur frequently over time, with important breakthroughs allowing important scale changes. Long-term illness, exposure to toxicants, and aging could be monitored with such a multi-omics integrated approach. Onco-proteogenomics should rapidly become mature in terms of methodology and should lead to a better knowledge of cancer progression. Onco-proteogenomics could give rise to a paradigm shift in terms of cancer-related biomarkers, which could most probably be a set of proteoforms determined for each individual and each cancer type that could appear or change in abundance under specific conditions, rather than a given biomarker for the whole population. In such a case, the validation of the discovery approach would be more important than the biomarkers themselves. Of note, such proteogenomic biomarkers may not be amenable to targeted antibody- or PCR-based detection. Rather, nucleic acid sequencers and tandem mass spectrometers might be common in the near future in clinical and oncological diagnostic laboratories. Better knowledge of pathogens and parasites is also likely to be obtained by proteogenomics, leading to novel fundamental advances in fighting these agents.

ACCEPTED MANUSCRIPT

Key issues

● Proteogenomics combines the strengths of different omics approaches. Based on the six-frame translation of genome or RNAseq transcriptome sequences, proteogenomics allows the quick identification of key proteins from a shotgun proteomic dataset. Better annotation of translation starts and splicing events can also be obtained from the peptides recorded by tandem mass spectrometry.

● Proteogenomics is today the customary approach for discovering novel human genes and documenting the different proteoforms that could be produced from each of the ~20,000 human protein-coding genes. To date, approximately 18% of human genes have not been characterized through the detection by mass spectrometry of any peptide sequence evidence. Advances in covering the whole human proteome are being achieved through a multinational proteogenomics consortium.

● Omics-based personalized monitoring is starting to be implemented, representing an important breakthrough in personalized medicine. Proteogenomics allows individual polymorphisms to be taken into account when analyzing proteomic data. Although the cost of long-term studies on large cohorts restrains its generalization, proteogenomics approaches show attractive perspectives.

● Onco-proteogenomic approaches have been implemented for a better understanding of cancer pathways. Because of the important genomic drift of cancer tissues, novel proteoforms not yet present in human protein sequence databases can be detected by tandem mass spectrometry when searching unassigned MS/MS spectra. These unexpected peptides or proteoforms could be used as novel diagnostic biomarkers, or at least for personalized monitoring.

● Numerous human pathogens, largely bacteria, have been better characterized in terms of protein sequence patrimony by proteogenomics. Complex protozoan parasites are also under

ACCEPTED MANUSCRIPT

study to improve their genome annotation and to better describe their cellular mechanisms. New knowledge could emerge from proteogenomics-inspired characterization of microbe-host interactions, and consequently novel drugs targeting the most sensitive molecular players could be proposed.

Financial & competing interests disclosure

This study was supported by the Commissariat à l'Energie Atomique et aux Energies Alternatives and the Agence Nationale de la Recherche (ANR-12-BSV6-0012-01 & ANR-14-CE21-0006-02). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

References

Papers of special note have been highlighted as:
● of interest
●● of considerable interest

1. Strathmann FG, Hoofnagle AN. Current and future applications of mass spectrometry to the clinical laboratory. *Am J Clin Pathol*, 136(4), 609-616 (2011).
2. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. *Nat Methods*, 10(3), 186-187 (2013).
3. Lavigne JP, Espinal P, Dunyach-Remy C, Messad N, Pantel A, Sotto A. Mass spectrometry: a revolution in clinical microbiology? *Clin Chem Lab Med*, 51(2), 257-270 (2013).
4. Armengaud J. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr Opin Microbiol*, 12(3), 292-300 (2009).
5. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief Funct Genomic Proteomic*, 7(1), 50-62 (2008).
6. Armengaud J. Proteogenomics and systems biology: quest for the ultimate missing parts. *Expert Rev Proteomics*, 7(1), 65-77 (2010).

ACCEPTED MANUSCRIPT

7. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, 4(1), 59-77 (2004).
- A pioneering paper for proteogenomics where peptide-to-nucleotide sequence mapping was introduced.

8. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*, 11(11), 1114-1125 (2014).
- A recent review on methodologies and concepts in proteogenomics.

9. Renuse S, Chaerkady R, Pandey A. Proteogenomics. *Proteomics*, 11(4), 620-630 (2011).
10. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931-945 (2004).
11. Venter JC, Adams MD, Myers EW *et al.* The sequence of the human genome. *Science*, 291(5507), 1304-1351 (2001).
12. Horvatovich P, Lundberg EK, Chen YJ *et al.* Quest for Missing Proteins: Update 2015 on Chromosome-Centric Human Proteome Project. *J Proteome Res*, (2015).
13. Hartmann EM, Armengaud J. N-terminomics and proteogenomics, getting off to a good start. *Proteomics*, 14(23-24), 2637-2646 (2014).
14. Krishna R, Xia D, Sanderson S *et al.* A large-scale proteogenomics study of apicomplexan pathogens-Toxoplasma gondii and Neospora caninum. *Proteomics*, 15(15), 2618-2628 (2015).
15. Bland C, Hartmann EM, Christie-Oleza JA, Fernandez B, Armengaud J. N-Terminal-oriented proteogenomics of the marine bacterium roseobacter denitrificans Och114 using N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography. *Mol Cell Proteomics*, 13(5), 1369-1381 (2014).
16. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*, 295(5558), 1306-1311 (2002).
17. Le TB, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, 342(6159), 731-734 (2013).
18. Shevchenko A, Jensen ON, Podtelejnikov AV *et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A*, 93(25), 14440-14445 (1996).
19. Yates JR, 3rd, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, 67(18), 3202-3210 (1995).
20. Armengaud J. Microbiology and proteomics, getting the best of both worlds! *Environ Microbiol*, 15(1), 12-23 (2013).
21. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. Non-model organisms, a species endangered by proteogenomics. *J Proteomics*, 105, 5-18 (2014).
22. Menschaert G, Van Criekinge W, Notelaers T *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics*, 12(7), 1780-1790 (2013).
23. Chocu S, Evrard B, Lavigne R *et al.* Forty-four novel protein-coding loci discovered using a proteomics informed by transcriptomics (PIT) approach in rat male germ cells. *Biol Reprod*, 91(5), 123 (2014).
24. Luge T, Kube M, Freiwald A, Meierhofer D, Seemuller E, Sauer S. Transcriptomics assisted proteomic analysis of Nicotiana occidentalis infected by Candidatus Phytoplasma mali strain AT. *Proteomics*, 14(16), 1882-1889 (2014).

ACCEPTED MANUSCRIPT

25. Trapp J, Geffard O, Imbert G *et al.* Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. *Mol Cell Proteomics*, 13(12), 3612-3625 (2014).
26. de Groot A, Roche D, Fernandez B *et al.* RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium Deinococcus deserti. *Genome Biol Evol*, 6(4), 932-948 (2014).
27. Rubiano-Labrador C, Bland C, Miotello G *et al.* Proteogenomic insights into salt tolerance by a halotolerant alpha-proteobacterium isolated from an Andean saline spring. *J Proteomics*, 97, 36-47 (2014).
28. Armengaud J. The power of positive thinking in quantitative proteomics. *Proteomics*, (2015).
29. Bertaccini D, Vaca S, Carapito C, Arsene-Ploetze F, Van Dorsselaer A, Schaeffer-Reiss C. An improved stable isotope N-terminal labeling approach with light/heavy TMPP to automate proteogenomics data validation: dN-TOP. *J Proteome Res*, 12(6), 3063-3070 (2013).
30. Bland C, Bellanger L, Armengaud J. Magnetic immunoaffinity enrichment for selective capture and MS/MS analysis of N-terminal-TMPP-labeled peptides. *J Proteome Res*, 13(2), 668-680 (2014).
31. Kleifeld O, Doucet A, Prudova A *et al.* Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat Protoc*, 6(10), 1578-1611 (2011).
32. Lange PF, Huesgen PF, Nguyen K, Overall CM. Annotating N termini for the human proteome project: N termini and Nalpha-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *J Proteome Res*, 13(4), 2028-2044 (2014).
33. Vaca Jacome AS, Rabilloud T, Schaeffer-Reiss C *et al.* N-terminome analysis of the human mitochondrial proteome. *Proteomics*, 15(14), 2519-2524 (2015).
34. Sun H, Chen C, Shi M *et al.* Integration of mass spectrometry and RNA-Seq data to confirm human ab initio predicted genes and lncRNAs. *Proteomics*, 14(23-24), 2760-2768 (2014).
35. Sun H, Chen C, Lian B *et al.* Identification of HPV integration and gene mutation in HeLa cell line by integrated analysis of RNA-Seq and MS/MS data. *J Proteome Res*, 14(4), 1678-1686 (2015).
36. Bischoff R, Permentier H, Guryev V, Horvatovich P. Genomic variability and protein species - improving sequence coverage for proteogenomics. *J Proteomics*, (2015).
37. Koch A, Gawron D, Steyaert S *et al.* A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*, 14(23-24), 2688-2698 (2014).

● A very attractive approach to discover translation start sites based on the RNAseq of mRNAs engaged in protein synthesis and straightforwardly purified through ribosome profiling.

38. Kitata RB, Dimayacyac-Esleta BR, Choong WK *et al.* Mining Missing Membrane Proteins by High-pH Reverse-Phase StageTip Fractionation and Multiple Reaction Monitoring Mass Spectrometry. *J Proteome Res*, (2015).
39. Horvatovich P, Vegvari A, Saul J *et al.* In Vitro Transcription/Translation System: A Versatile Tool in the Search for Missing Proteins. *J Proteome Res*, (2015).
40. Islam MT, Garg G, Hancock WS, Risk BA, Baker MS, Ranganathan S. Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the "missing" human proteome. *J Proteome Res*, 13(1), 76-83 (2014).

ACCEPTED MANUSCRIPT

41. Jagtap P, Goslinga J, Kooren JA *et al.* A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, 13(8), 1352-1357 (2013).
42. Jagtap PD, Johnson JE, Onsongo G *et al.* Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *J Proteome Res*, 13(12), 5898-5908 (2014).
43. Krasnov GS, Dmitriev AA, Kudryavtseva AV *et al.* PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. *J Proteome Res*, (2015).
44. Nagaraj SH, Waddell N, Madugundu AK *et al.* PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization. *J Proteome Res*, 14(5), 2255-2266 (2015).
45. Pang CN, Tay AP, Aya C *et al.* Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J Proteome Res*, 13(1), 84-98 (2014).
46. Tabas-Madrid D, Alves-Cruzeiro J, Segura V *et al.* Proteogenomics Dashboard for the Human Proteome Project. *J Proteome Res*, (2015).
47. Omenn GS, Lane L, Lundberg EK, Beavis RC, Nesvizhskii AI, Deutsch EW. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J Proteome Res*, (2015).
48. Dasari S, Theis JD, Vrana JA *et al.* Proteomic detection of immunoglobulin light chain variable region peptides from amyloidosis patient biopsies. *J Proteome Res*, 14(4), 1957-1967 (2015).
49. Marko-Varga G, Omenn GS, Paik YK, Hancock WS. A first step toward completion of a genome-wide characterization of the human proteome. *J Proteome Res*, 12(1), 1-5 (2013).
50. Fermin D, Allen BB, Blackwell TW *et al.* Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol*, 7(4), R35 (2006).
51. Farrah T, Deutsch EW, Hoopmann MR *et al.* The state of the human proteome in 2012 as viewed through PeptideAtlas. *J Proteome Res*, 12(1), 162-171 (2013).
52. Farrah T, Deutsch EW, Omenn GS *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J Proteome Res*, 13(1), 60-75 (2014).
53. Deutsch EW, Sun Z, Campbell D *et al.* State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *J Proteome Res*, (2015).
54. Wilhelm M, Schlegl J, Hahne H *et al.* Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), 582-587 (2014).
55. Kim MS, Pinto SM, Getnet D *et al.* A draft map of the human proteome. *Nature*, 509(7502), 575-581 (2014).
56. Fernandez-Irigoyen J, Zelaya MV, Perez-Valderrama E, Santamaria E. New insights into the human brain proteome: Protein expression profiling of deep brain stimulation target areas. *J Proteomics*, (2015).
57. Kang MG, Byun K, Kim JH *et al.* Proteogenomics of the human hippocampus: The road ahead. *Biochim Biophys Acta*, 1854(7), 788-797 (2015).
58. Chen C, Liu X, Zheng W, Zhang L, Yao J, Yang P. Screening of missing proteins in the human liver proteome by improved MRM-approach-based targeted proteomics. *J Proteome Res*, 13(4), 1969-1978 (2014).

ACCEPTED MANUSCRIPT

59. Jumeau F, Com E, Lane L *et al.* Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J Proteome Res*, (2015).
60. Zhang C, Li N, Zhai L *et al.* Systematic analysis of missing proteins provides clues to help define all of the protein-coding genes on human chromosome 1. *J Proteome Res*, 13(1), 114-125 (2014).
61. Carapito C, Lane L, Benama M *et al.* Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J Proteome Res*, (2015).
62. Chen LC, Liu MY, Hsiao YC *et al.* Decoding the disease-associated proteins encoded in the human chromosome 4. *J Proteome Res*, 12(1), 33-44 (2013).
63. Ranganathan S, Khan JM, Garg G, Baker MS. Functional annotation of the human chromosome 7 "missing" proteins: a bioinformatics approach. *J Proteome Res*, 12(6), 2504-2510 (2013).
64. Ahn JM, Kim MS, Kim YI *et al.* Proteogenomic analysis of human chromosome 9-encoded genes from human samples and lung cancer tissues. *J Proteome Res*, 13(1), 137-146 (2014).
65. Gupta MK, Jayaram S, Madugundu AK *et al.* Chromosome-centric human proteome project: deciphering proteins associated with glioma and neurodegenerative disorders on chromosome 12. *J Proteome Res*, 13(7), 3178-3190 (2014).
66. Manda SS, Nirujogi RS, Pinto SM *et al.* Identification and characterization of proteins encoded by chromosome 12 as part of chromosome-centric human proteome project. *J Proteome Res*, 13(7), 3166-3177 (2014).
67. Segura V, Medina-Aunon JA, Mora MI *et al.* Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome. *J Proteome Res*, 13(1), 158-172 (2014).
68. Liu S, Im H, Bairoch A *et al.* A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J Proteome Res*, 12(1), 45-57 (2013).
69. Ponomarenko EA, Kopylov AT, Lisitsa AV *et al.* Chromosome 18 transcriptoproteome of liver tissue and HepG2 cells and targeted proteome mapping in depleted plasma: update 2013. *J Proteome Res*, 13(1), 183-190 (2014).
70. Zgoda VG, Kopylov AT, Tikhonova OV *et al.* Chromosome 18 transcriptome profiling and targeted proteome mapping in depleted plasma, liver tissue and HepG2 cells. *J Proteome Res*, 12(1), 123-134 (2013).
71. Pinto SM, Manda SS, Kim MS *et al.* Functional annotation of proteome encoded by human chromosome 22. *J Proteome Res*, 13(6), 2749-2760 (2014).
72. Rappsilber J, Ryder U, Lamond AI, Mann M. Large-scale proteomic analysis of the human spliceosome. *Genome Res*, 12(8), 1231-1245 (2002).
73. Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics*, 12(8), 2341-2353 (2013).
●● A pioneering high-throughput analysis by tandem mass spectrometry of peptides signing splice-junctions facilitated with RNA-seq data.
74. Zhu Y, Hultin-Rosenberg L, Forshed J, Branca RM, Orre LM, Lehtio J. SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol Cell Proteomics*, 13(6), 1552-1562 (2014).
75. Zhang B, Wang J, Wang X *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518), 382-387 (2014).

ACCEPTED MANUSCRIPT

●● A large dataset acquired on 95 tumor samples by shotgun proteomics and corresponding to 7,526 protein groups was analyzed to understand which protein coding alterations are expressed at the protein level. The authors clearly show that in their samples mRNA abundance does not reliably predict protein abundance.
76. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3), 1720-1730 (1999).
77. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 13(4), 227-232 (2012).
78. Caron E, Espona L, Kowalewski DJ *et al.* An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife*, 4 (2015).
79. Admon A, Bassani-Sternberg M. The Human Immunopeptidome Project, a suggestion for yet another postgenome next big thing. *Mol Cell Proteomics*, 10(10), O111 011833 (2011).
80. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics*, 14(3), 658-673 (2015).
81. Bausch-Fluck D, Hofmann A, Bock T *et al.* A mass spectrometric-derived cell surface protein atlas. *PLoS One*, 10(3), e0121314 (2015).
82. Wilkes EH, Terfve C, Gribben JG, Saez-Rodriguez J, Cutillas PR. Empirical inference of circuitry and plasticity in a kinase signaling network. *Proc Natl Acad Sci U S A*, 112(25), 7719-7724 (2015).
83. Ong SE, Blagoev B, Kratchmarova I *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5), 376-386 (2002).
84. Zhu H, Pan S, Gu S, Bradbury EM, Chen X. Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Commun Mass Spectrom*, 16(22), 2115-2123 (2002).
85. Jorgensen C, Sherman A, Chen GI *et al.* Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. *Science*, 326(5959), 1502-1509 (2009).
86. Chen R, Mias GI, Li-Pook-Than J *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6), 1293-1307 (2012).
87. Snyder M. Q & A: the Snyderome. *Genome Biol*, 13(3), 147 (2012).
● Comments from Michael Snyder on the first large scale OMICS personalized survey conducted by his research team over a 14-month period.
88. Helmy M, Sugiyama N, Tomita M, Ishihama Y. Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. *Geome Biology*, 11(Suppl 1), P17 (2010).
● Introduction of proteogenomics in oncology and first proposal for the wording "onco-proteogenomics".
89. Alfaro JA, Sinha A, Kislinger T, Boutros PC. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nat Methods*, 11(11), 1107-1113 (2014).
●● An interesting state-of-the-art review on onco-proteogenomics.
90. Boja ES, Rodriguez H. Proteogenomic convergence for understanding cancer pathways and networks. *Clin Proteomics*, 11(1), 22 (2014).
91. Rivers RC, Kinsinger C, Boja ES, Hiltke T, Mesri M, Rodriguez H. Linking cancer genome to proteome: NCI's investment into proteogenomics. *Proteomics*, 14(23-24), 2633-2636 (2014).

ACCEPTED MANUSCRIPT

92. Ntai I, LeDuc RD, Fellers RT *et al.* Integrated Bottom-up and Top-down Proteomics of Patient-derived Breast Tumor Xenografts. *Mol Cell Proteomics*, (2015).
93. Huang CH, Kuo CJ, Liang SS *et al.* Onco-proteogenomics identifies urinary S100A9 and GRN as potential combinatorial biomarkers for early diagnosis of hepatocellular carcinoma. *BBA Clinical*, 3(June), 205-213 (2015).
94. Edwards NJ, Oberti M, Thangudu RR *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res*, 14(6), 2707-2713 (2015).
95. Zhang Z, Bast RC, Jr., Yu Y *et al.* Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res*, 64(16), 5882-5890 (2004).
96. Sjostrom M, Ossola R, Breslin T *et al.* A Combined Shotgun and Targeted Mass Spectrometry Strategy for Breast Cancer Biomarker Discovery. *J Proteome Res*, 14(7), 2807-2818 (2015).
97. Shen Y, Tolic N, Liu T *et al.* Blood peptidome-degradome profile of breast cancer. *PLoS One*, 5(10), e13133 (2010).
98. Barbacid M. ras genes. *Annu Rev Biochem*, 56, 779-827 (1987).
99. Davies H, Bignell GR, Cox C *et al.* Mutations of the BRAF gene in human cancer. *Nature*, 417(6892), 949-954 (2002).
100. Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1), 2-8 (2010).
101. Fanayan S, Smith JT, Lee LY *et al.* Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. *J Proteome Res*, 12(4), 1732-1742 (2013).
102. Sethi MK, Thaysen-Andersen M, Kim H *et al.* Quantitative proteomic analysis of paired colorectal cancer and non-tumorigenic tissues reveals signature proteins and perturbed pathways involved in CRC progression and metastasis. *J Proteomics*, 126, 54-67 (2015).
103. Woo S, Cha SW, Bonissone S *et al.* Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J Proteome Res*, (2015).
●● The authors describe an integrative proteogenomic method to identify multiple variant peptides with advanced bioinformatics. Their methodology using customized mining of RNA-seq data applies really well for immunoglobulin gene variations & rearrangements.
104. Olsen L, Campos B, Winther O, Sgroi DC, Karger BL, Brusic V. Tumor antigens as proteogenomic biomarkers in invasive ductal carcinomas. *BMC Med Genomics*, 7 Suppl 3, S2 (2014).
105. Keerthikumar S, Gangoda L, Liem M *et al.* Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. *Oncotarget*, 6(17), 15375-15396 (2015).
106. Hatakeyama K, Ohshima K, Fukuda Y *et al.* Identification of a novel protein isoform derived from cancer-related splicing variants using combined analysis of transcriptome and proteome. *Proteomics*, 11(11), 2275-2282 (2011).
107. Woo S, Cha SW, Na S *et al.* Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics*, 14(23-24), 2719-2730 (2014).
108. Polyakova A, Kuznetsova K, Moshkovskii S. Proteogenomics meets cancer immunology: mass spectrometric discovery and analysis of neoantigens. *Expert Rev Proteomics*, 1-9 (2015).
● An interesting review on the application of proteogenomics for cancer immunology. Perspectives of this field are really interesting in terms of clinics.

ACCEPTED MANUSCRIPT

109. Baudet M, Ortet P, Gaillard JC *et al.* Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Mol Cell Proteomics*, 9(2), 415-426 (2010).
110. Gallien S, Perrodou E, Carapito C *et al.* Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res*, 19(1), 128-135 (2009).
111. Gupta N, Benhamida J, Bhargava V *et al.* Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res*, 18(7), 1133-1142 (2008).
112. Clair G, Roussi S, Armengaud J, Duport C. Expanding the known repertoire of virulence factors produced by Bacillus cereus through early secretome profiling in three redox conditions. *Mol Cell Proteomics*, 9(7), 1486-1498 (2010).
113. Madeira JP, Alpha-Bazin B, Armengaud J, Duport C. Time dynamics of the Bacillus cereus exoproteome are shaped by cellular oxidation. *Front Microbiol*, 6, 342 (2015).
114. Durmus S, Cakir T, Ozgur A, Guthke R. A review on computational systems biology of pathogen-host interactions. *Front Microbiol*, 6, 235 (2015).
115. Payne SH, Huang ST, Pieper R. A proteogenomic update to Yersinia: enhancing genome annotation. *BMC Genomics*, 11, 460 (2010).
116. Schrimpe-Rutledge AC, Jones MB, Chauhan S *et al.* Comparative omics-driven genome annotation refinement: application across Yersiniae. *PLoS One*, 7(3), e33903 (2012).
117. Venter E, Smith RD, Payne SH. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One*, 6(11), e27587 (2011).
118. Zhao L, Liu L, Leng W, Wei C, Jin Q. A proteogenomic analysis of Shigella flexneri using 2D LC-MALDI TOF/TOF. *BMC Genomics*, 12, 528 (2011).
119. Muller SA, Findeiss S, Pernitzsch SR *et al.* Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. *J Proteomics*, 86, 27-42 (2013).
120. Pettersen VK, Steinsland H, Wiker HG. Improving Genome Annotation of Enterotoxigenic Escherichia coli TW10598 by a Label-Free Quantitative MS/MS Approach. *Proteomics*, (2015).
121. Alexeev D, Kostrjukova E, Aliper A *et al.* Application of Spiroplasma melliferum proteogenomic profiling for the discovery of virulence factors and pathogenicity mechanisms in host-associated spiroplasmas. *J Proteome Res*, 11(1), 224-236 (2012).
122. Nagarajha Selvan LD, Kaviyil JE, Nirujogi RS *et al.* Proteogenomic analysis of pathogenic yeast Cryptococcus neoformans using high resolution mass spectrometry. *Clin Proteomics*, 11(1), 5 (2014).
123. Prasad TS, Harsha HC, Keerthikumar S *et al.* Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. *J Proteome Res*, 11(1), 247-260 (2012).
124. Suarez S, Ferroni A, Lotz A *et al.* Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. *J Microbiol Methods*, 94(3), 390-396 (2013).
125. Durighello E, Bellanger L, Ezan E, Armengaud J. Proteogenomic biomarkers for identification of Francisella species and subspecies by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry. *Anal Chem*, 86(19), 9394-9398 (2014).
126. Nirujogi RS, Pawar H, Renuse S *et al.* Moving from unsequenced to sequenced genome: reanalysis of the proteome of Leishmania donovani. *J Proteomics*, 97, 48-61 (2014).

ACCEPTED MANUSCRIPT

127. Pawar H, Kulkarni A, Dixit T, Chaphekar D, Patole MS. A bioinformatics approach to reanalyze the genome annotation of kinetoplastid protozoan parasite Leishmania donovani. *Genomics*, 104(6 Pt B), 554-561 (2014).
128. Pawar H, Renuse S, Khobragade SN *et al.* Neglected tropical diseases and omics science: proteogenomics analysis of the promastigote stage of Leishmania major parasite. *OMICS*, 18(8), 499-512 (2014).
129. Pawar H, Sahasrabuddhe NA, Renuse S *et al.* A proteogenomic approach to map the proteome of an unsequenced pathogen - Leishmania donovani. *Proteomics*, 12(6), 832-844 (2012).
130. Jamdhade MD, Pawar H, Chavan S *et al.* Comprehensive proteomics analysis of glycosomes from Leishmania donovani. *OMICS*, 19(3), 157-170 (2015).
131. Robinson SD, Safavi-Hemami H, Raghuraman S *et al.* Discovery by proteogenomics and characterization of an RF-amide neuropeptide from cone snail venom. *J Proteomics*, 114, 38-47 (2015).
132. Chemonges S, Tung JP, Fraser JF. Proteogenomics of selective susceptibility to endotoxin using circulating acute phase biomarkers and bioassay development in sheep: a review. *Proteome Sci*, 12(1), 12 (2014).
133. Frese CK, Altelaar AF, van den Toorn H *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal Chem*, 84(22), 9668-9673 (2012).
134. Malmstrom L, Bakochi A, Svensson G *et al.* Quantitative proteogenomics of human pathogens using DIA-MS. *J Proteomics*, (2015).

- A stimulating study where data-independent mass spectrometry acquisition has been performed for 34 clinical isolates of *Streptococcus pyogenes*. The authors developed a proteogenomics strategy for explaining virulence modulation with genomics and proteomics integrated data.

135. Sajic T, Liu Y, Aebersold R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics Clin Appl*, 9(3-4), 307-321 (2015).
136. Faulkner S, Dun MD, Hondermarck H. Proteogenomics: emergence and promise. *Cell Mol Life Sci*, 72(5), 953-957 (2015).

ACCEPTED MANUSCRIPT

Table 1. Main features of proteogenomic annotation of the human proteome.

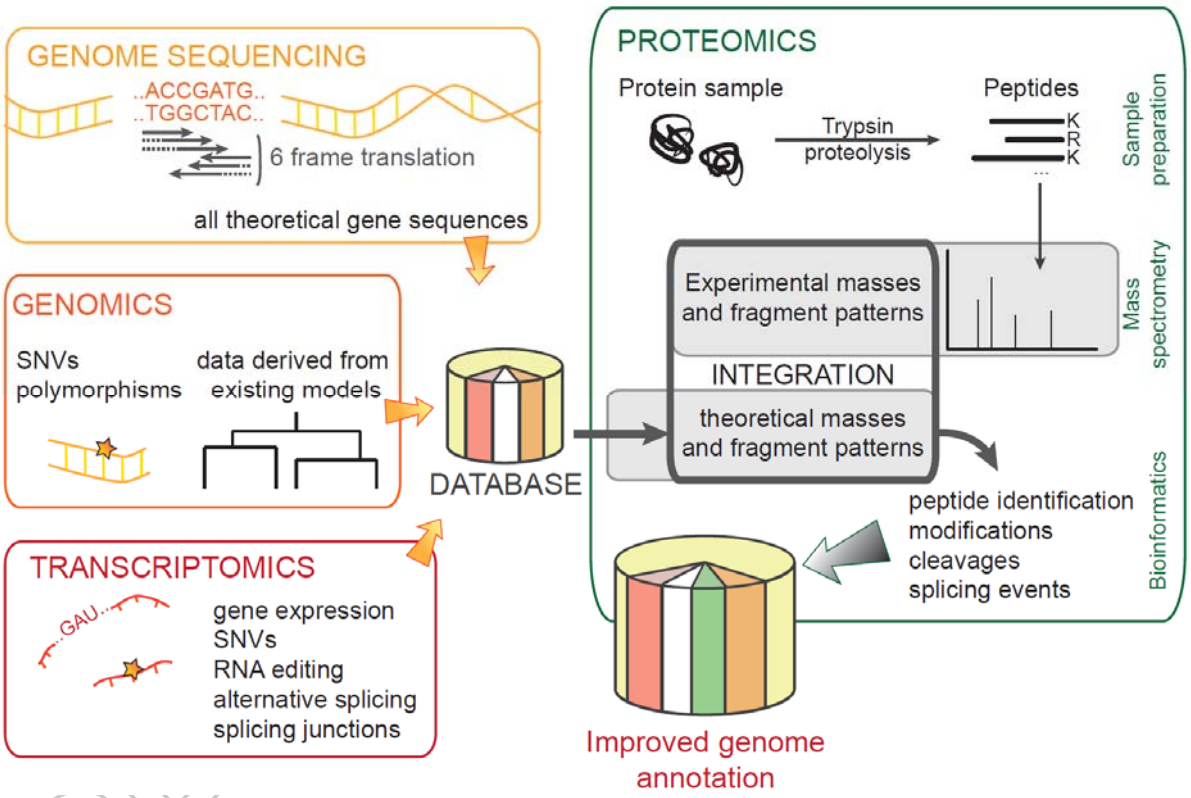
Subproteome	Description of the study	Scale of the study	Significant output	References
Mitochondrial proteome	N-terminome analysis of human mitochondria-enriched samples from U937 human monocytic cells.	Identification of 2,714 unique proteins, including 810 mitochondrial.	Identification of the N-termini of 693 unique proteins (283 mitochondrial); confirmation of 120 already annotated processing cleavage sites; characterization of 302 new cleavage sites.	[33]
Hippocampus and brain proteome	Proteomic analysis of 2 subcortical brain areas and correlation of their proteomic expression profiles with human brain transcriptome data available at Allen Brain Atlas.	Identification of 4,235 unique proteins.	Identification of 894 proteins with brain-wide expression not previously identified by MS; Identification of 112 proteins with region-specific expression.	[56]
Liver proteome	MRM screen of proteins identified in the liver with low protein evidence (one peptide only) in peptide atlas.	Monitoring of 1,110 transitions corresponding to 185 proteins.	Presence of 57 proteins confirmed with one peptide; identification of 7 proteins with no previous MS information.	[58]
Spermatozoa proteome	Catalogue of missing proteins in total protein extracts from isolated human spermatozoa.	Identification of 11,281 peptides corresponding to 1,547 proteins.	Identification of 89 missing proteins; confirmation of the presence of 3 uncertain proteins.	[59]
From the chromosome-centric human proteome project				
Chromosome 1	OMICs-integrated analysis of three human liver cell lines presenting different lung metastatic potentials.	Identification of 1,308 (out of 1,719) proteins.	Mass spectrometric evidence for 60 additional chromosome 1 gene products.	[60]
Chromosomes 2 and 14	Analysis of 40 human samples in order to identify missing proteins encoded by chromosomes 2 and 14.	85,326 dat files searched against a database containing 30,952 unique peptide sequences	Detection of 83 unique peptides from 58 proteins that were not previously validated.	[61]
Chromosome 4	Exploration of proteomics data on chromosome 4 for cancer biomarker identification.	Analysis of 757 protein-coding genes and their experimental evidence at the protein level.	Identification of 141 chromosome 4-encoded proteins as cancer cell-secretable/shedable proteins and 54 chromosome 4-encoded proteins that have been classified as cancer-associated proteins.	[62]
Chromosome 7	Functional annotation of the proteins of chromosome 7 that do not have evidence at the proteomic, antibody, or structural levels.	Bioinformatics analysis to gain insights into the 170 "missing" proteins of chromosome 7.	Identification of 90 “missing” proteins; putative functional annotations for 27 proteins.	[63]
Chromosome 9	Bioinformatic and proteogenomic analysis to catalogue chromosome 9-encoded proteins from normal tissues, lung cancer cell lines and lung cancer tissues.	Identification of 75% of the human chromosome 9 genes.	Identification of 46 missing proteins, 15 being detected only in lung cancer tissues; identification of 21 SNPs and 4 mutations containing peptides from normal human cells/tissues and lung cancer cell lines, respectively.	[64]

ACCEPTED MANUSCRIPT

Chromosome 12	Mass spectrometry profiling of 30 different histologically normal human tissues and cell types in order to refine chromosome 12 genomic annotation.	Identification of 1,535 proteins encoded by 836 genes on human chromosome 12.	Identification of 89 proteins for which prior experimental evidence was missing; confirmation of the start sites of ~200 proteins by identifying protein N-terminal acetylated peptides; identification of alternative start sites for 11 proteins that were not annotated in public databases; identification of 12 novel protein coding sequences.	[66]
Chromosome 12	Functional analysis of the proteome encoded by chromosome 12 from databases and patient samples in order to detect proteins potentially involved in gliomas and major neurological conditions.	Identification of 1,066 protein coding genes.	Identification of 171 proteins defined as "missing"; first mass spectrometric evidence for two "missing" proteins; identification of 103 differentially expressed proteins with secretory potential.	[65]
Chromosome 16	Proteogenomic analysis of chromosome 16 and development of SRM methods for quantification of chromosome 16-encoded proteins.	Coverage of 41% of chromosome 16 proteins.	SRM identification of 49 known proteins and recombinant forms of 24 "missing" proteins.	[67]
Chromosome 17	Proteogenomic analysis of chromosome 18.	Identification of 1,169 protein-coding genes.	List of 59 "missing" proteins as well as 201 proteins that have inconclusive mass spectrometric identifications.	[68]
Chromosome 18	SRM and RNAseq profiling of proteins encoded by chromosome 18 in plasma, liver and HepG2 cells, followed by copy number estimation.	277 proteins targeted by SRM.	Identification of 209 proteins in the plasma, 168 in the liver, with over 50% overlap; 27 proteins remained undetected; estimation of protein copy numbers for 228 master proteins, including quantitative data on 164 proteins in plasma, 171 in the HepG2 cell line, and 186 in liver tissue.	[69,70]
Chromosome 22	Proteogenomic analysis of chromosome 22.	Proteomic profiling specifically focused on the 442 RefSeq gene entries corresponding to chromosome 22-encoded genes.	Protein evidence for 367 genes including 47 proteins that are currently annotated as "missing" proteins; confirmation of the translation start sites of 120 chromosome 22-encoded proteins; evidence of novel coding regions.	[71]

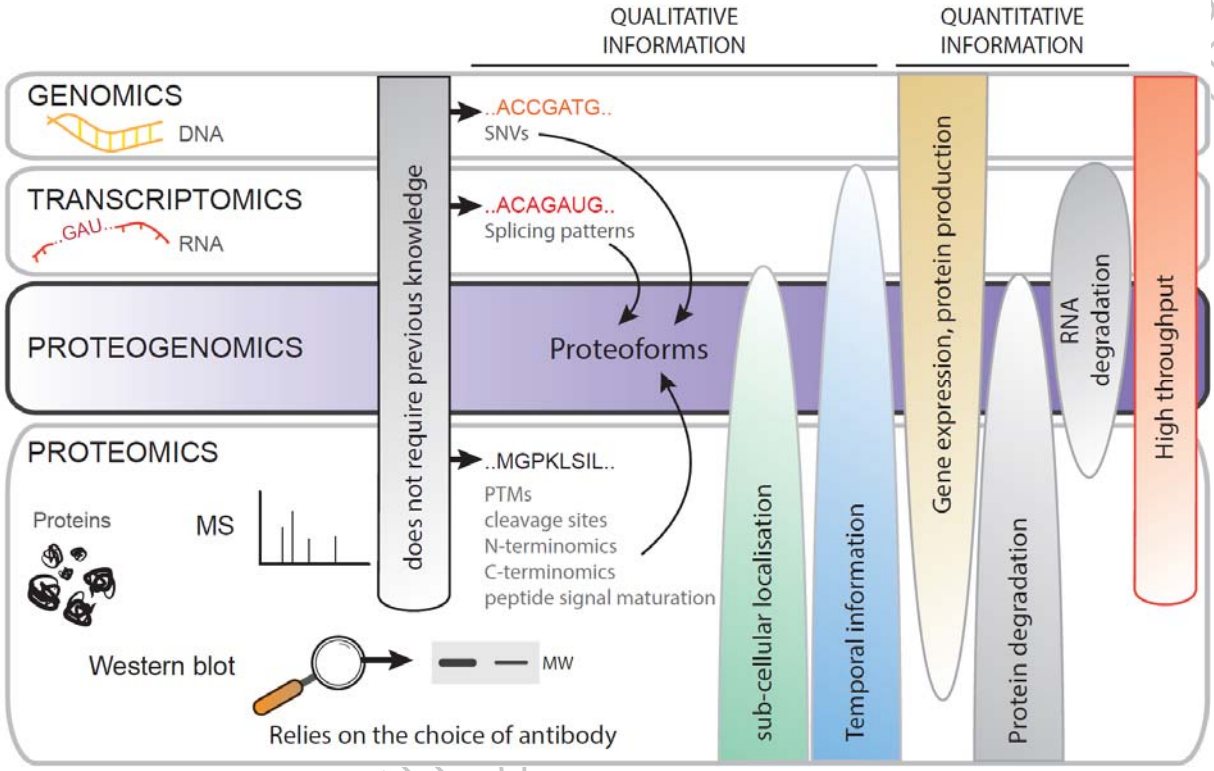
ACCEPTED MANUSCRIPT

Figure 1. Overview of the proteogenomics pipeline. A database of theoretical protein sequences can be created based on genome sequencing data (6-frame translation of all sequenced contigs), transcriptomic data (6-frame translation, alternative splicing events, etc.), and genomic data (single nucleotide variations and more complex polymorphisms). This database allows the interpretation of the mass spectrometry data acquired with standard proteomic approaches. The proteogenomic integration of these experimental data leads to an improved genome annotation, and an improved protein sequence database.



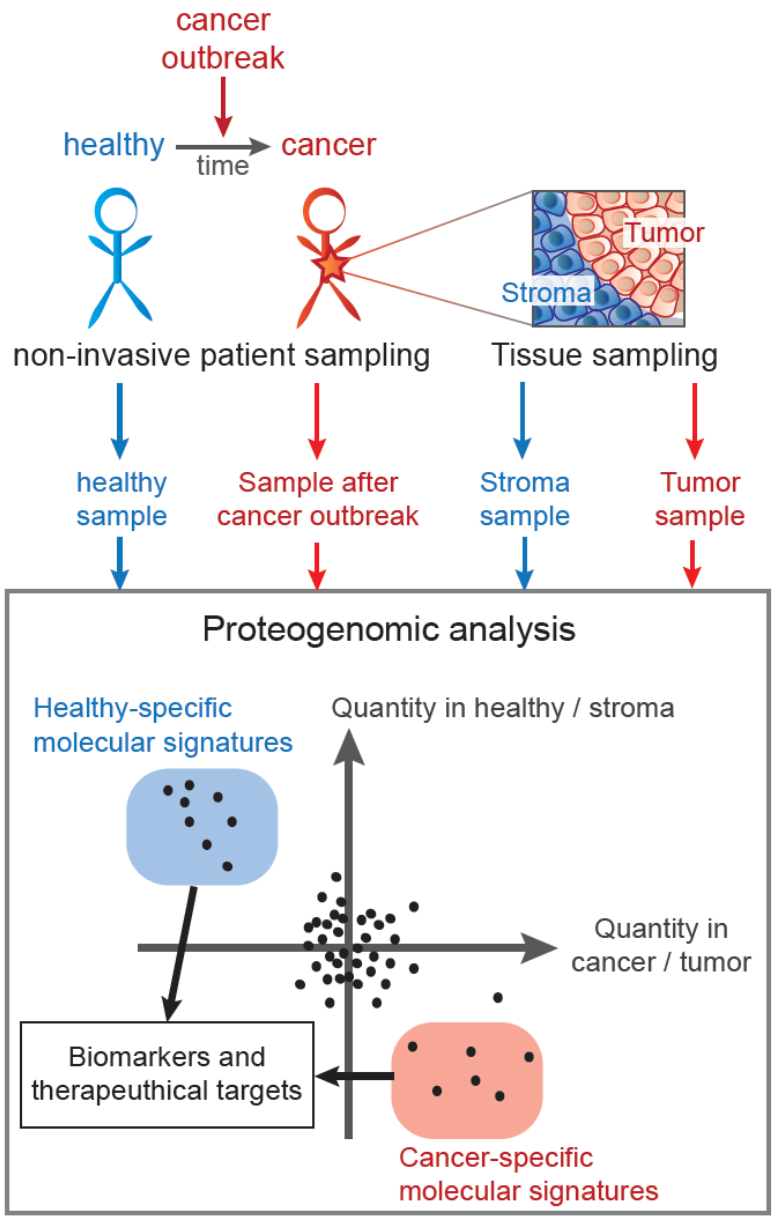
ACCEPTED MANUSCRIPT

Figure 2. Combining omics strengths. The advantages of genomics, transcriptomics and proteomics can be combined in proteogenomics, both in qualitative and quantitative terms.



ACCEPTED

Figure 3. General strategy in personalized onco-proteogenomics. Potential markers and therapeutic targets are obtained by identifying novel proteoforms by proteogenomics and their quantitation in healthy and non-healthy samples.



ACCEPTED MANUSCRIPT