

Reconciling proteomics with next generation sequencing

Teck Yew Low^{1,2} and Albert JR Heck^{1,2}

Both genomics and proteomics technologies have matured in the last decade to a level where they are able to deliver system-wide data on the qualitative and quantitative abundance of their respective molecular entities, that is DNA/RNA and proteins. A next logical step is the collective use of these technologies, ideally gathering data on matching samples. The first large scale so-called proteogenomics studies are emerging, and display the benefits each of these layers of analysis has on the other layers to together generate more meaningful insight into the connection between the phenotype/physiology and genotype of the system under study. Here we review a selected number of these studies, highlighting what they can uniquely deliver. We also discuss the future potential and remaining challenges, from a somewhat proteome biased perspective.

Addresses

¹ Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

² Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

Corresponding author: Heck, Albert JR (A.J.R.Heck@uu.nl)

Current Opinion in Chemical Biology 2016, 30:14–20

This review comes from a themed issue on **Omics**

Edited by **Daniel Nomura, Alan Saghatelian** and **Eranthie Weerapana**

<http://dx.doi.org/10.1016/j.cbpa.2015.10.023>

1367-5931/© 2015 Elsevier Ltd. All rights reserved.

manner, omics strategies such as transcriptomics, epigenomics, proteomics and metabolomics have emerged and developed into different, albeit reasonable, levels of maturity. Nevertheless, integrating these data modalities creates a next challenge, essential to understand biology in a system-wide manner.

Considering the vast scope of multi-omics, in this review we focus our discussion on multi-omics integration with an emphasis on proteomics, one layer that in particular has attained increasing maturity over the past decade. Proteomics refers to the *en masse* characterization or measurement of the structure and function of proteins; their abundance, post-translational modifications (PTMs) state, and interactions with other proteins or biomolecules [1]. The core analytical technology used in the analysis of proteomes is mass spectrometry. It is largely implemented in a so-called bottom-up workflow, whereby initially lysates are proteolysed into peptides, separated by chromatography and then measured and sequenced by mass spectrometry (MS). The resulting peptide fragmentation spectra are matched to sequences in genome and/or protein databases. In the field of multi-omics integration, the first major challenge is to ensure that different data modalities can relate to each other. In proteomics, this bottleneck lies in the protein databases used for peptide-to-spectrum matching. Such databases are typically constructed from reference genomes. This poses a limit as to how accurately a proteome can be measured as many biological specimens are obtained from non-model organisms lacking reference genome. Besides, due to mutations, gene sequences are inherently polymorphic, while decoding DNA to proteins may involve co-transcriptional or post-transcriptional modifications that introduce multiple isoforms/proteoforms that are not found in the reference genome.

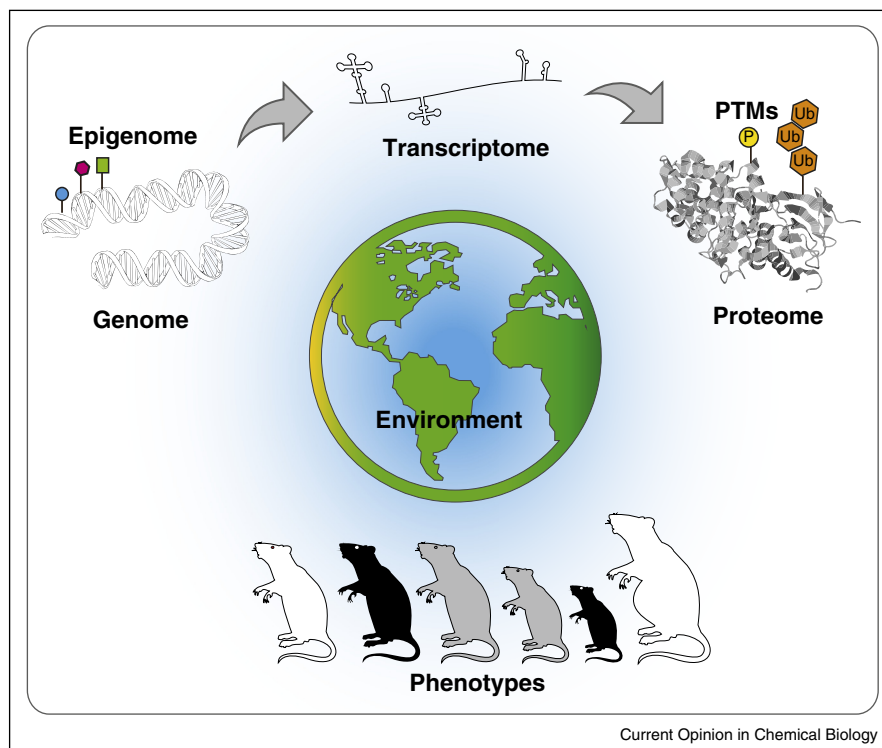
Introduction

In essence, genomes form the blueprints of life. This genetic blueprint transforms *via* multiple layers of biomolecular conversions, involving RNA and proteins, being influenced by interactions with the environment, into specific phenotypes of cells, tissues and organisms (Figure 1). The phenotypes, in return, determine how likely a genome is inherited when subjected to natural selection. Although whole genomes can now be sequenced in parallel, thorough understanding of the resulting biology can only be accomplished by integrative studies of the interplays and dynamics among the different molecular layers. To probe each layer in a system-wide

Proteogenomics: towards more precise genomes and proteomes

Strategically, since a gene, its transcripts and proteins are derived from the same template, it is obvious that genome, transcriptome and proteome data, obtained by a diverse array of platforms, algorithms and expertise should foremost be standardized. In recent years, next-generation sequencing (NGS) [2], with its high throughput and depth, has rendered it economically and logistically feasible to interface genomics directly with proteomics using the matched sample as reference [3^{••}]. This emerging field of proteogenomics (Figure 2) is generally associated with annotating newly sequenced genomes, with six-frame

Figure 1



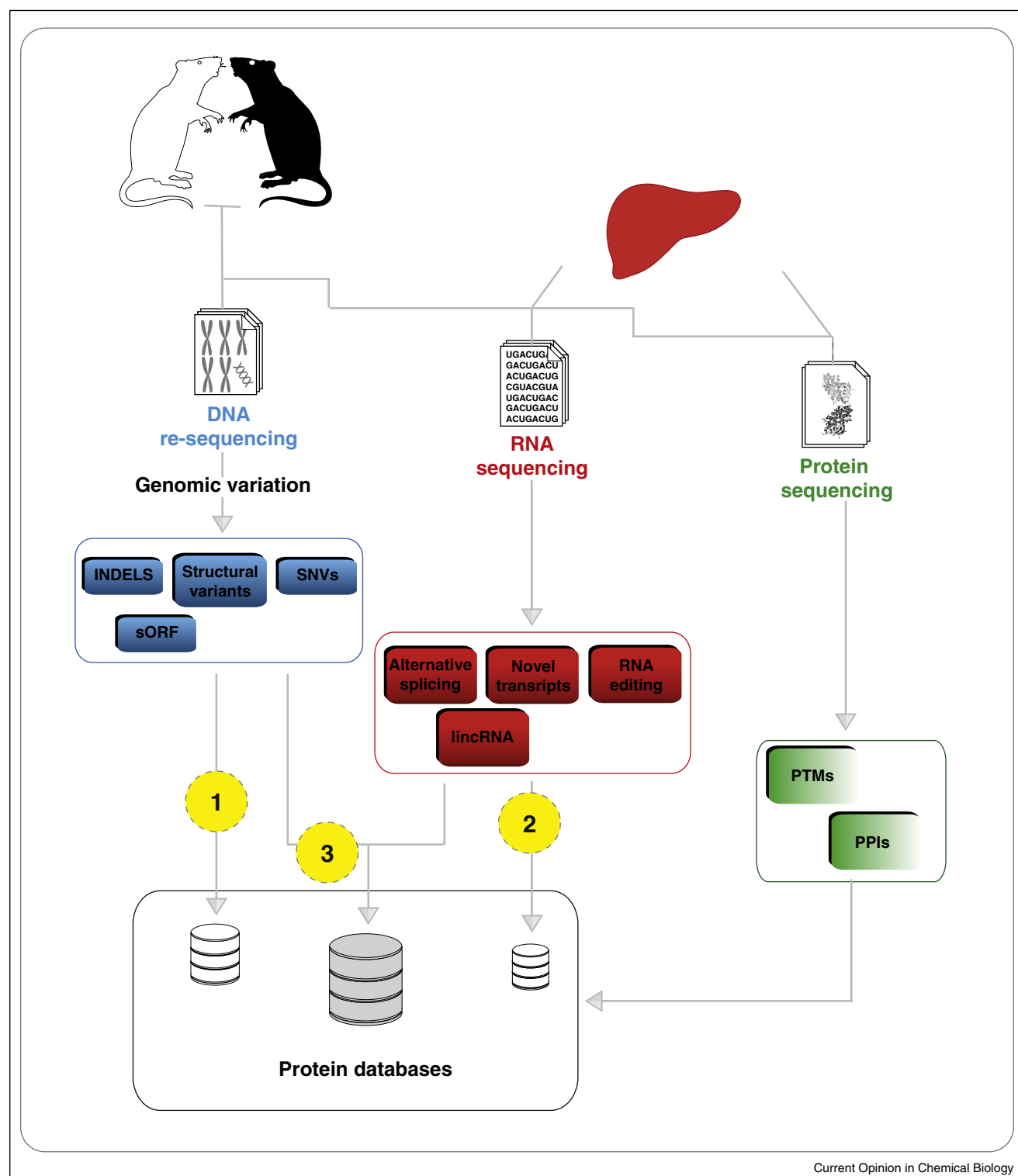
Interaction of genomes, environments and phenotypes. The phenotypes of organisms are largely determined by the interaction of the genomes and the environment. Save for spontaneous mutations, genome sequences themselves are largely static. The transmission of conversion of encoded genetic information *via* layers of bio-molecules such as epigenomes, transcriptomes, proteomes and post-translational modifications help an organisms react to external stimuli and respond to environmental selection pressures.

translation [4] or *ab initio* gene-prediction algorithms [5] to transform genome data *in silico* to gene models, which are then validated by MS-based proteomics data. Illustratively, these techniques have been successfully demonstrated in confirming splice-junctions in maize [5], pseudo-genes in mouse [6] and small open reading frames (sORF) in human cell lines [7]. Meanwhile, high-throughput genotyping can be performed by re-sequencing the genomes or exomes of multiple organisms and mapped to an annotated reference, so as to detect genome-encoded variants, including single nucleotide polymorphisms/variants (SNP/SNV), small insertions and deletions (indels), structural variants or copy number variants (CNV). By incorporating these variants that are protein-coding in a protein database, Lichti *et al.* identified 17 proteins on chromosome 19 carrying single amino acid variants (SAAVs) in glioma stem cells [8]. Woo *et al.* searched MS-based data acquired from ovarian carcinomas against databases constructed from The Cancer Genome Atlas (TCGA) repository, and identified 524 novel peptides including doubly mutated peptides, frame-shifts, and non-sample-recruited mutations [9].

NGS also forms the technical backbone for RNA sequencing (RNAseq) and ribosome profiling (RIBOseq).

RNAseq quantifies all sequenced transcripts, that is mRNAs and non-coding RNAs. Similar to genome sequencing, RNAseq detects protein-coding variants, while additionally captures co-transcriptionally/post-transcriptionally modified RNAs that result from alternative events in transcription initiation, splicing, poly-adenylation [10] and RNA editing [11]. Incorporating RNAseq data helps discriminating proteoforms further. Low *et al.* performed parallel genomics, transcriptomics and in-depth proteomics (using 5 proteases for proteolysis) on liver tissues from two strains of rats and identified with proteomics, 20 out of the 196 non-synonymous RNA editing events captured by RNAseq. At the same time they resolved 15 and 13 protein splice-isoforms that were unique for one of the two strains, respectively [12[•]]. In a proteogenomic study of human colon and rectal cancer by the Clinical Proteomics Tumor Analysis Consortium (CPTAC), a separate protein database was generated from RNAseq for each of the 87 tumor samples. This allowed CPTAC to discover 64 and 101 somatic variants documented previously by TCGA and COSMIC; *versus* 526 likely germline variants from dbSNP database [13[•]]. Nonetheless, the greatest merit is perhaps the *de novo* assembly of full transcriptome without reference

Figure 2



A schematic of proteogenomics workflow. Key to proteogenomics study is building an accurate reference protein database for peptide-to-spectrum matching (PSM) in proteomics. (1.) Such databases are normally derived from the genome sequences of a single reference organism in the form of reference genome. By genotyping multiple animals, genome variants such as SNPs and indels can be detected, and can be used to build a database, which is specific to an individual or a strain of animals. (2.) It is also possible to build protein databases from RNAseq data. RNAseq-derived protein databases do not only contain genome variants but also encompass variants that are introduced at the transcript level such as alternative splicing or RNA-editing. (3.) To build a comprehensive protein database, both DNAseq and RNAseq data are considered. Although this scheme is quite generic it has been adopted from the proteogenomics study of Low *et al.* [12*].

genomes, as this allows protein databases to be constructed relatively cheaply from RNAseq data especially for non-model organisms [14]. As proof of this principle, Evans *et al.* demonstrated a proteomics informed by transcriptomics (PIT) technique that identified over 3,700 distinct proteins from human cells infected with adenovirus [15]. Wühr *et al.* constructed a reference proteome by combining heterogeneous sources of *Xenopus laevis* RNAseq data, and were in this way able to identify more than 11,000 proteins from unfertilized *Xenopus* eggs [16^{*}].

RIBOseq determines the sequences and quantifies, at single-nucleotide resolution, ribosome-bound mRNA fragments that are protected from nuclease digestion [17]. It does not provide any full-length transcript sequence, but rather, every sequenced fragment infers the position of an *in vivo* translating ribosome; and the density of these footprints reveals the number of ribosomes translating that transcript. RIBOseq measures RNAs but actually quantifies protein synthesis and therefore can be considered as a *de facto* proteogenomics technique. By using inhibitors targeting translational initiation, ribosomes can be arrested at the initiation sites. With RIBOseq, translation initiation sites (TIS), as well as alternative TISs, non-canonical (non-AUG) TISs or upstream ORFs (uORFs) [18^{**}] can be identified. RIBOseq data can also be incorporated into existing reference databases, as demonstrated by Menschaert *et al.* who searched such databases with N-terminal proteomics data, resulting in 2.5% increase in protein identification including 16 N-terminally extended protein variants introduced by alternative TISs, besides four translated uORFs [19]. Besides, more than 1700 unique alternative protein N-termini identified by the same group in human and murine cellular proteomes [20].

Challenges, hypes and opportunities in proteogenomics; learning from each other

Proteogenomics faces challenges that arise due to the sheer scale of genomic data enlarging the database size, affecting the false discovery rates of peptide-to-spectrum matching. These challenges have been reviewed in detail [3^{**},21,22], and will not be elaborated here. Nonetheless, accompanying these challenges are opportunities for developing bioinformatics tools that (i.) generate a peptide database from a genome and matches peptides to MS/MS spectra [23], (ii.) integrate transcriptomic and proteomic datasets to identify novel proteins and discriminate protein-coding and non-coding genes [24], (iii.) provide genome-wide visualization of ribosome occupancy and a TIS calling algorithm for generating protein synthesis-based sequence databases [25] or (iv.) databases that combine mutation information collected from several public resources such as COSMIC, IARC P53, OMIM, and UniProtKB [26].

Importantly, the proteomics community needs to recognize that NGS has also contributed to several anomalies

that are controversial or even artifactual. Firstly, the vast majority of eukaryotic genomes are pervasively transcribed. However, it is unclear how many newly found RNAs have functions and how many are byproducts of functional, or spurious, transcription events [27]. Secondly, although about 10,000 non-canonical RNA editing events were initially reported [28], this has been later disputed due to lack of evidence [29]. Finally, a recent RIBOseq study suggests that ribosomes occupy many supposedly noncoding regions of transcriptomes, including 5' UTRs and long noncoding RNAs (lncRNAs), whereby these pervasive translations outside protein-coding sequences were supported by multiple lines of evidence [30]. Especially perplexing examples are the lncRNAs, as RIBOseq studies pertaining to their translation to functional proteins are conflicting [31,32]; while mass spectrometry provided independent evidence for lncRNA-derived peptides [7,33^{*},34^{*}]. As such, discoveries inspired by NGS, though exciting, need to be evaluated critically, and it will be advantageous when different omics communities collaborate more closely for exchange of information, exploiting different technology platforms for confirmation.

Towards integrative biological insights by quantitative proteogenomics

Beyond improving proteoform discrimination, the ultimate objective of multi-omics should lie in providing biological insights into the correlation of a selected phenotype to a particular genotype [35]. It does so by first collecting and integrating information about the identities of biomolecules and their dynamic changes within certain spatial, temporal and environmental constraints. It is not hard to see that a quantitative strategy is required for quantifying changes in the abundance, localities or modifications of these biomolecules; so that their inter-relationship can be mathematically represented and modeled.

Organisms display predominantly quantitative traits — complex yet measurable phenotypes that are cumulatively influenced by multiple genes and the environment. The variations in quantitative traits, such as height or weight, can be statistically correlated to genomic loci that influence them ('quantitative trait loci' or QTL) using QTL mapping or genome-wide association studies (GWAS) [36]. Likewise, the level of expression of transcripts, proteins and metabolites can also be correlated to genomic loci to give respectively, statistically significant associating eQTLs [37], pQTLs [38] and mQTLs [39]. Exemplary, Wu *et al.* performed combined QTL analyses by quantifying the transcriptome, a subset of the metabolome and liver proteome from 40 strains of the BXD mouse genetic reference population on two diverse diets [40]. They discovered dozens of eQTLs, pQTLs and mQTLs that are linked to metabolic phenotypes, while Dhtkd1 was identified as a primary regulator of 2-aminoadipate. Low *et al.*, with quantitative proteomics and RNAseq, demonstrated that CYP17A1, the most down-regulated gene in

the liver of the hypersensitive SHR rat, is also a previously reported top-hit for hypertension in human GWAS [12[•]]. Subsequent eQTL analysis of CYP17A1 in recombinant inbred HXB/BXH rats revealed a very strong cis-effect, which was later traced to a single point mutation at the transcription start site. Lundby *et al.* performed tissue-specific quantitative interaction proteomics to map a network of five genes involved in long QT syndrome (LQTS). The LQTS protein-protein interaction network was subsequently integrated with 35 GWAS loci from the corresponding QT-interval variation, a quantitative trait, to identify candidate genes [41[•]].

While the genetic basis for gene expression or protein/metabolite synthesis can be explained by QTLs, another relationship that intrigues many biologists pertains to how well transcripts and proteins correlate in abundance [42]. Cellular physiology is mainly regulated by the absolute concentrations and activity of proteins. The concentrations of proteins in steady state are determined by transcription and mRNA decay, as well as translation and protein degradation. However, their respective share of contribution towards the regulation of protein abundance remains unclear. Historically, mRNA abundance has been shown to correlate poorly with that of the proteins ($R^2 \leq 0.4$); and efforts by experimentalists to disentangle these have invariably concluded with translation and protein turnover being the main determinants of protein abundance [13[•],43,44]. However, recent re-analysis of these data by statisticians revealed that this poor correlation can be attributed to measurement errors that underestimated the low-abundance proteins and transcription; and after correcting for these errors, as high as 81% of variance in protein levels can be explained by mRNA levels [45,46^{••}]. In other studies, proteins-to-mRNA ratios were found to remain remarkably conserved for orthologous genes within two strains of *Pseudomonas* [47], in the livers tissues of two strains of rats [12[•]], and across 12 different human tissues [34[•]]. These observations led to the speculation that translation rate is an encoded characteristic of a transcript, and thus by regulating mRNA levels, one effectively controls the amount of proteins synthesized [34[•]]. Regardless of the underlying principles, these latest results imply the feasibility to, directly or indirectly, use mRNA levels as proxies to estimate the levels of proteins in steady state.

Direct interactions within and between different molecular layers

Biomolecules rarely function alone, but interact with one another to form modules and complex networks. Already, for many transcriptomics and proteomics datasets, co-expression analyses are performed to discover genes having similar expression patterns across multiple conditions, suggestive of functional relationship such as physical interaction [48]. On the other hand, omics methods are increasingly made available to directly assess physical

interactions. For example, yeast-two-hybrid or affinity purification coupled to mass spectrometry (AP-MS) have become the methods of choice to probe protein-protein interactions [49], while ChIP-seq is used to capture DNA regions that are bound by transcription factors [50]. For instance, quantitative mass spectrometry has been applied by Spruijt *et al.* to probe readers binding to 5-methylcytosine (mC) and 5-hydroxymethylcytosine (hmC) [51]; and by Butter *et al.* to screen 12 SNPs encoding IL2RA for differential transcription factor binding, and found differential, allele-specific binding of the transcription factors [52]. To systematically detect interactions between RNA and RNA-binding proteins (RBPs), Kwon *et al.* employed interactome capture, which involved UV cross-linking of RBP to RNA in living cells, oligo (dT) capture and mass spectrometry. Consequently, they identified 555 proteins constituting the mRNA interactome in mouse embryonic stem cells [53]. These technologies, although still rather immature, are important for defining the interplay between the different molecular levels present in the cell.

Conclusions and future outlook

Several omics technologies have reached maturation in terms of throughput, sensitivity, reproducibility and speed. This has resulted in a number of large-scale multi-omics projects being completed or still in the planning [54–56]. The next step would be to integrate these omics data so as to understand biological systems [57^{••},58]. Apart from huge, multi-national consortium-led efforts, we also witness multi-omics being increasingly applied in biomedical research in smaller scales [59]. Though the strength of multi-omics remain analytical and observational, powerful engineering disciplines have since emerged for genome-editing [60], designing gene circuits or networks to reprogram organisms [61] or even synthesizing whole chromosomes [62]. It is foreseeable that multi-omics information can be used to improve such synthetic systems; *vice versa*, synthetic systems can provide a much more maneuverable and testable platforms (in comparison to, say, the recombinant inbred rat panels) for validating multi-omics results — as the next stage of integration.

Acknowledgements

This work has been supported by the Netherlands Proteomics Centre, the Netherlands Organization for Scientific Research (NWO) supporting the Roadmap embedded large scale proteomics facility *Proteins@Work* (project 184.032.201) and by the PRIME-XS project grant agreement number 262067 supported by the European Community's Seventh Framework Programme (FP7/2007-2013) to AJRH.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Altelaar AFM, Munoz J, Heck AJR: **Next-generation proteomics: towards an integrative view of proteome dynamics.** *Nat Rev Genet* 2013, **14**:35–48.

2. Reuter JA, Spacek DV, Snyder MP: **High-throughput sequencing technologies.** *Mol Cell* 2015, **58**:586-597.
3. Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies.** *Nat Methods* 2014, **11**:1114-1125.
One of the most comprehensive and up-to-date review for proteogenomics, covering research strategies in different aspects as well as discussion of their advantages and drawbacks.
4. Khatun J, Yu Y, Wrobel JA, Risk BA, Gunawardena HP, Secret A, Spitzer WJ, Xie L, Wang L, Chen X *et al.*: **Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions.** *BMC Genomics* 2013, **14**:141.
5. Castellana NE, Shen Z, He Y, Walley JW, Cassidy CJ, Briggs SP, Bafna V: **An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*.** *Mol Cell Proteomics* 2014, **13**:157-167.
6. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS *et al.*: **Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and resurrected pseudogenes in the mouse genome.** *Genome Res* 2011, **21**:756-767.
7. Slavoff SA, Mitchell AJ, Schwaib AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: **Peptidomic discovery of short open reading frame-encoded peptides in human cells.** *Nat Chem Biol* 2013, **9**:59-64.
8. Lichti CF, Mostovenko E, Wadsworth PA, Lynch GC, Pettitt BM, Sulman EP, Wang Q, Lang FF, Rezeli M, Marko-Varga G *et al.*: **Systematic identification of single amino acid variants in glioma stem-cell-derived chromosome 19 proteins.** *J Proteome Res* 2015, **14**:778-786.
9. Woo S, Cha SW, Na S, Guest C, Liu T, Smith RD, Rodland KD, Payne S, Bafna V: **Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data.** *Proteomics* 2014, **14**:2719-2730.
10. De Klerk E, 't Hoen PAC: **Alternative mRNA transcription, processing, and translation: insights from RNA sequencing.** *Trends Genet* 2015, **31**:128-139.
11. Wulff B-E, Sakurai M, Nishikura K: **Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing.** *Nat Rev Genet* 2011, **12**:81-85.
12. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hübner N, van Breukelen B, Mohammed S *et al.*: **Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis.** *Cell Rep* 2013, **5**:1469-1478.
An integrated genomics, transcriptomics and proteomics study of rat liver tissues from two genetic backgrounds; serves as a good example for studying the links between genotypes and phenotypes.
13. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S *et al.*: **Proteogenomic characterization of human colon and rectal cancer.** *Nature* 2014, **513**:382-387.
One of the largest proteogenomics efforts performed by the CPTAC on actual tumour tissue samples obtained from TCGA.
14. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29**:644-652.
15. Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA: **De novo derivation of proteomes from transcriptomes for transcript and protein identification.** *Nat Methods* 2012, **9**:1207-1211.
16. Wühr M, Freeman RM, Presler M, Horb ME, Peshkin L, Gygi SP, Kirschner MW: **Deep proteomics of the *Xenopus laevis* egg using an mRNA-derived reference database.** *Curr Biol* 2014, **24**:1467-1475.
This paper describes an interface that can be used to derive protein databases from RNAseq data from multiple sources.
17. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS: **The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments.** *Nat Protoc* 2012, **7**:1534-1550.
18. Ingolia NT: **Ribosome profiling: new views of translation, from single codons to genome scale.** *Nat Rev Genet* 2014, **15**:205-213.
A very informative and detailed discussion of the past and future applications of ribosome profiling by a pioneer in the field.
19. Menschaert G, Van Crielinge W, Notelaers T, Koch A, Crappé J, Gevaert K, Van Damme P: **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* 2013, **12**:1780-1790.
20. Van Damme P, Gawron D, Van Crielinge W, Menschaert G: **N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men.** *Mol Cell Proteomics* 2014, **13**:1245-1261.
21. Alfaro JA, Sinha A, Kislinger T, Boutros PC: **Onco-proteogenomics: cancer proteomics joins forces with genomics.** *Nat Methods* 2014, **11**:1107-1113.
22. Wang X, Liu Q, Zhang B: **Leveraging the complementary nature of RNA-Seq and shotgun proteomics data.** *Proteomics* 2014, **14**:2676-2687.
23. Risk BA, Spitzer WJ, Giddings MC: **Peppy: proteogenomic search software.** *J Proteome Res* 2013, **12**:3019-3025.
24. Gascoigne DK, Cheetham SW, Cattenoz PB, Clark MB, Amaral PP, Taft RJ, Wilhelm D, Dinger ME, Mattick JS: **Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes.** *Bioinformatics* 2012, **28**:3042-3050.
25. Crappe J, Ndahe E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Crielinge W, Van Damme P *et al.*: **PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration.** *Nucleic Acids Res* 2014 <http://dx.doi.org/10.1093/nar/gku1283>.
26. Yang X, Lazar IM: **XMAN: a *Homo sapiens* mutated-peptide database for the MS analysis of cancerous cell states.** *J Proteome Res* 2014, **13**:5486-5495.
27. Jensen TH, Jacquier A, Libri D: **Dealing with pervasive transcription.** *Mol Cell* 2013, **52**:473-484.
28. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG: **Widespread RNA and DNA sequence differences in the human transcriptome.** *Science* 2011, **333**:53-58.
29. Piskol R, Peng Z, Wang J, Li JB: **Lack of evidence for existence of noncanonical RNA editing.** *Nat Biotechnol* 2013, **31**:19-20.
30. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Taihouarne GJS, Jackson SE, Wills MR, Weissman JS: **Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.** *Cell Rep* 2014, **8**:1365-1379.
31. Ingolia NT, Lareau LF, Weissman JS: **Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.** *Cell* 2011, **147**:789-802.
32. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: **Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.** *Cell* 2013, **154**:240-251.
33. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al.*: **A draft map of the human proteome.** *Nature* 2014, **509**:575-581.
One of the first two papers on the comprehensive characterization of the human proteome.
34. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al.*: **Mass-spectrometry-based draft of the human proteome.** *Nature* 2014, **509**:582-587.
One of the first two papers on the comprehensive characterization of the human proteome.
35. Civelek M, Lusis AJ: **Systems genetics approaches to understand complex traits.** *Nat Rev Genet* 2014, **15**:34-48.

36. Ott J, Wang J, Leal SM: **Genetic linkage analysis in the age of whole-genome sequencing.** *Nat Rev Genet* 2015, **16**:275-284.
 37. Li Q, Stram A, Chen C, Kar S, Gayther S, Pharoah P, Haiman C, Stranger B, Kraft P, Freedman ML: **Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types.** *Hum Mol Genet* 2014, **23**:5294-5302.
 38. Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, Tang H, Snyder M: **Variation and genetic control of protein abundance in humans.** *Nature* 2013, **499**:79-82.
 39. Ghazalpour A, Bennett BJ, Shih D, Che N, Orozco L, Pan C, Hagopian R, He A, Kayne P, Yang W *et al.*: **Genetic regulation of mouse liver metabolite levels.** *Mol Syst Biol* 2014, **10**:730.
 40. Wu Y, Williams EG, Dubuis S, Mottis A, Jovaisaite V, Houten SM, Argmann CA, Faridi P, Wolski W, Kutalik Z *et al.*: **Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population.** *Cell* 2014, **158**:1415-1430.
 41. Lundby A, Rossin EJ, Steffensen AB, Acha MR, Newton-Cheh C, Pfeufer A, Lynch SN, Olesen S-P, Brunak S, Ellinor PT *et al.*: **Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics.** *Nat Methods* 2014, **11**:868-874.
- The integration of interaction proteomics with GWAS to identify candidate genes for QT-interval variation, a quantitative trait.
42. Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.** *Nat Rev Genet* 2012, **13**:227-232.
 43. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control.** *Nature* 2011, **473**:337-342.
 44. Kristensen AR, Gsponer J, Foster LJ: **Protein synthesis rate is the predominant regulator of protein expression during differentiation.** *Mol Syst Biol* 2013, **9**:689.
 45. Li JJ, Bickel PJ, Biggin MD: **System wide analyses have underestimated protein abundances and the importance of transcription in mammals.** *PeerJ* 2014, **2**:e270.
 46. Li JJ, Biggin MD: **Statistics requantitates the central dogma.** *Science* 2015, **347**:1066-1067.
- This paper presents the latest perspectives of statisticians on the correlation of the abundance of transcripts and proteins, citing a number of latest and original papers within the field.
47. Kwon T, Huse HK, Vogel C, Whiteley M, Marcotte EM: **Protein-to-mRNA ratios are conserved between *Pseudomonas aeruginosa* strains.** *J Proteome Res* 2014, **13**:2370-2380.
 48. Deng S-P, Zhu L, Huang D-S: **Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks.** *BMC Genomics* 2015, **16**(Suppl. 3):S4.
 49. Wodak SJ, Vlasblom J, Turinsky AL, Pu S: **Protein-protein interaction networks: the puzzling riches.** *Curr Opin Struct Biol* 2013, **23**:941-953.
 50. Todeschini A-L, Georges A, Veitia RA: **Transcription factors: specific DNA binding and specific gene regulation.** *Trends Genet* 2014, **30**:211-219.
 51. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PWT, Bauer C, Münzel M, Wagner M, Müller M, Khan F *et al.*: **Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives.** *Cell* 2013, **152**:1146-1159.
 52. Butter F, Davison L, Viturawong T, Scheibe M, Vermeulen M, Todd JA, Mann M: **Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding.** *PLoS Genet* 2012, **8**:e1002982.
 53. Kwon SC, Yi H, Eichelbaum K, Föhr S, Fischer B, You KT, Castello A, Krijgsvelde J, Hentze MW, Kim VN: **The RNA-binding protein repertoire of embryonic stem cells.** *Nat Struct Mol Biol* 2013, **20**:1122-1130.
 54. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45**:1113-1120.
 55. Paik Y-K, Jeong S-K, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee H-J, Na K, Choi E-Y, Yan F *et al.*: **The chromosome-centric human proteome project for cataloging proteins encoded in the genome.** *Nat Biotechnol* 2012, **30**:221-223.
 56. Skipper M, Eccleston A, Gray N, Heemels T, Le Bot N, Marte B, Weiss U: **Presenting the epigenome roadmap.** *Nature* 2015, **518**:313.
 57. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D: **Methods of integrating data to uncover genotype-phenotype interactions.** *Nat Rev Genet* 2015, **16**:85-97.
- This review explores bioinformatics approaches for data integration: including meta-dimensional and multi-staged analyses to improve integrative understanding of the relationship between genomic variation and phenotypes.
58. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J *et al.*: **Defining functional DNA elements in the human genome.** *Proc Natl Acad Sci U S A* 2014, **111**:6131-6138.
 59. Hussein SMI, Puri MC, Tonge PD, Benevento M, Corso AJ, Clancy JL, Mosbergen R, Li M, Lee D-S, Cloonan N *et al.*: **Genome-wide characterization of the routes to pluripotency.** *Nature* 2014, **516**:198-206.
 60. Carroll D: **Genome engineering with targetable nucleases.** *Annu Rev Biochem* 2014, **83**:409-439.
 61. Brophy JAN, Voigt CA: **Principles of genetic circuit design.** *Nat Methods* 2014, **11**:508-520.
 62. Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, Dymond JS, Kuang Z, Scheifele LZ, Cooper EM *et al.*: **Total synthesis of a functional designer eukaryotic chromosome.** *Science* 2014, **344**:55-58.