Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

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 K Nomura**, **Alan Saghatelian** and **Eranthie Weerapana**

For a complete overview see the Issue and the Editorial

Available online 17th November 2015

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

Introduction

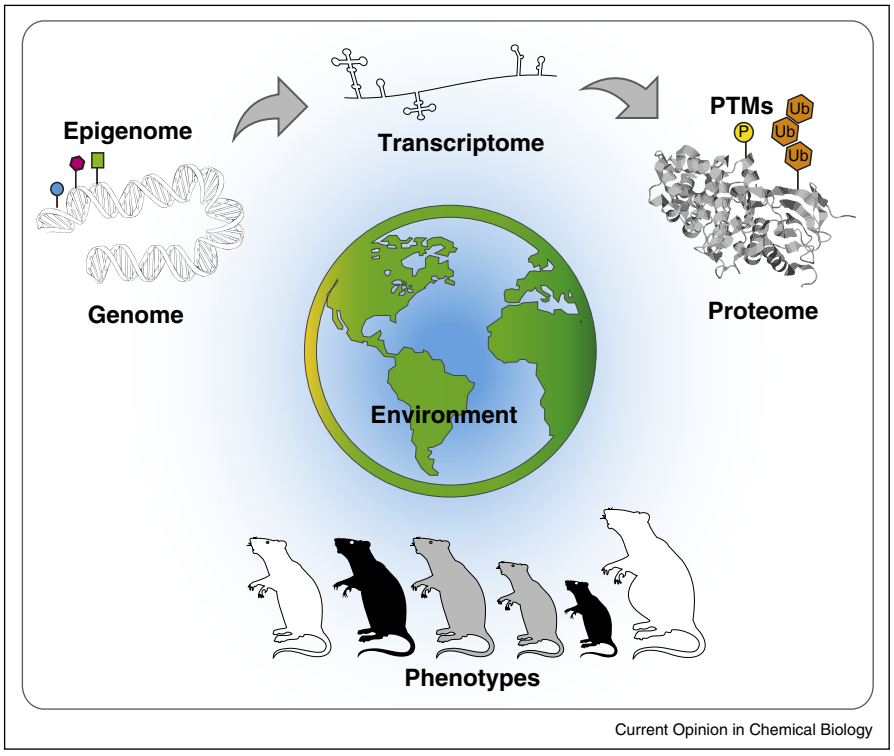
In essence, genomes form the blueprints of life. This genetic blueprint transforms *via* multiple layers of biomolecular conversions, involving RNA and proteins, being influenced by interactions with the environment, into specific phenotypes of cells, tissues and organisms (Figure 1). The phenotypes, 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 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.

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, trancriptomes, 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