Current Plant Biology 2 (2015) 21–24

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Review

Dual use of peptide mass spectra: Protein atlas and genome annotation

Justin W. Walley, Steven P. Briggs*

University of California San Diego, Section of Cell & Developmental Biology, La Jolla, CA 92093-0380, United States

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form 3 February 2015
Accepted 24 February 2015

Keywords:
Proteogenomics
Proteomics
Atlas
Annotation

ABSTRACT

One of the objectives of genome science is the discovery and accurate annotation of all protein-coding genes. Proteogenomics has emerged as a methodology that provides orthogonal information to traditional forms of evidence used for genome annotation. By this method, peptides that are identified via tandem mass spectrometry are used to refine protein-coding gene models. Namely, these peptides are used to confirm the translation of predicted protein-coding genes, as evidence of novel genes or for correction of current gene models. Proteogenomics requires deep and broad sampling of the proteome in order to generate sufficient numbers of unique peptides. Therefore, we propose that proteogenomic projects are designed so that the generated peptides can also be used to create a comprehensive protein atlas that quantitatively catalogues protein abundance changes during development and in response to environmental stimulus.

Contents

1. Introduction

The primary goal of genome annotation efforts is the discovery and accurate annotation of all protein-coding genes. A complete and accurately annotated proteome provides the building blocks for hypothesis-driven research seeking to enhance our understanding of biology. Genome annotation is a complex process involving multiple integrated tools, which have been described in detail [1–5] and are beyond the scope of this review. Briefly, traditional methods of genome annotation rely on combining various forms of evidence. This includes de novo gene prediction, which utilizes only patterns in the genomic sequence to infer gene structure. Additionally, transcript sequences from cDNA libraries can be leveraged to enhance gene prediction. Lastly, sequence conservation with related species can be incorporated into annotation pipelines. While DNA/RNA-based genome annotation approaches perform remarkably well, given the complexity of the challenge, they are currently unable to accurately predict all protein coding genes and their structure. Experimental evidence is required to determine if a transcript is translated and if the predicted protein sequence is correct.

The field of proteogenomics has emerged as a genome-wide method to improve genome annotations as well as to characterize the pattern of gene expression at the protein level. The concept of proteogenomics was introduced, by Jaffe and colleagues [6], as a method that utilizes peptides identified from their tandem mass spectra, for genome annotation (reviewed by [2,7–9]). Since its introduction, proteogenomics has successfully aided in the annotation of numerous prokaryotic and eukaryotic organisms. These studies have demonstrated that deep and broad sampling of the proteome is necessary, for proteogenomics, requiring the generation of hundreds of millions of mass spectra. Furthermore, protein

* Corresponding author.
E-mail address: sbriggs@ucsd.edu (S.P. Briggs).

http://dx.doi.org/10.1016/j.cpb.2015.02.001

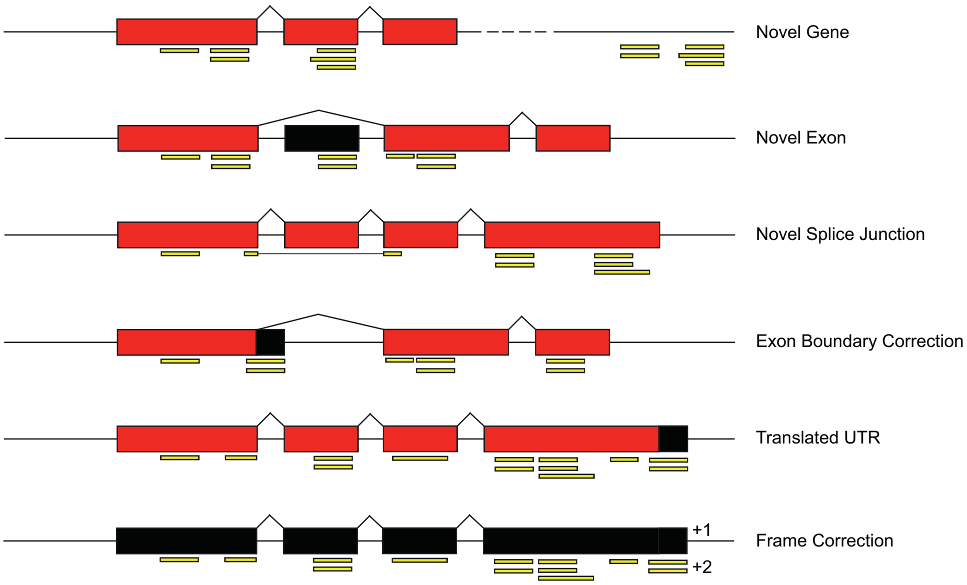

Fig. 1. Examples of gene model revision. Currently annotated exons are shown in red. Gene model revision suggested by novel peptides is depicted in black. Proteogenomically identified peptides are shown in yellow.

accumulation depends upon development and environmental conditions so spectra must be generated from a diverse set of samples to enable deep coverage of the proteome. Such broad sampling enables the additional use of the identified peptides for creation of a protein atlas that catalogs where, when, and how much of a given protein is present.

2. Proteogenomic enabled annotation

Proteogenomics provides a high-throughput method to incorporate protein level information into genome annotation. For this, tandem mass spectra are generated and then used to search genomic databases for peptide identification. The standard database utilized in proteogenomic pipelines is a six-frame translation of the genome [6]. Additionally, specialized types of databases such as an exon–splice graph, which is compact representation of predicted gene structures and splice junctions, have also been exploited [10]. The identified peptides fall into two categories. Namely, confirming peptides that match the current genome annotation and novel peptides, which do not (Fig. 1). It is important to emphasize that the confirming peptides represent critical events, as they directly confirm both the current structural annotation of a gene and demonstrate that the gene encodes a translated protein.

The novel peptides themselves can be further divided into two types of events. One category includes intergenic peptides, which map outside of known genes, and thus reveal the presence of novel genes. A second category is intragenic peptides that fall within a known locus, but do not match the currently annotated gene model. Intragenic peptides include those demonstrating the translation of 5′ or 3′ untranslated regions (UTR), alternative start/stop sites, proteins out of frame, incorrect exon boundaries, novel exons or novel splice sites. While one may assume that the identification these types of novel intergenic and intragenic peptides by proteogenomics to be rare, they are actually commonly found, even in well annotated model organisms (i.e. organisms that have been subjected to multiple rounds of genome annotation) (Table 1). This demonstrates that proteogenomics is a necessary addition to any comprehensive genome annotation effort.

3. Proteome sampling for proteogenomics

Deep and broad sampling of the proteome is necessary for comprehensive proteogenomic efforts. There are numerous strategies that have been developed for proteogenomic experiments to aid in maximizing the number of unique peptides identified by mass spectrometry [7,9,11]. Briefly, fractionation methods such as one-dimensional and two-dimensional gel electrophoresis, as well as gel-free chromatography based separations of proteins and peptides, aid in deep proteome sampling. Specialized sample preparations can also be used to sample subsets of the proteome such as phosphoproteins, basic proteins, small proteins, and N-terminal peptides [7,8,12–14]. Additionally, use of multiple proteases (examples include trypsin, chymotrypsin, Glu-C, and Lsy-C) helps to increase the percentage of sequence covered for a given protein. Another consideration is that the proteome composition depends on both developmental and environmental factors. Thus, analyzing a diverse array of samples is critical for achieving comprehensive proteome coverage [12,13].

Table 1
Proteogenomic publications in plants. (If Novel Genes and Model Revision were not clearly identified all values went into the Model Revision Column.)

Organism	Peptides	Proteins	Novel peptides	Novel genes	Model revision	Citation
Arabidopsis thaliana	86,456	13,029	261	22	35	[28]
Arabidopsis thaliana	144,079	12,769	18,024	778	695	[13]
Populus deltoides	4943			56		[34]
Chlamydomonas reinhardtii	9336		932	3	65	[35]
Oryza sativa	15,121	5034	166		40	[36]
Medicago truncatula	78,647	9843	1568	32	293	[37]
Zea mays	225,166	14,615	24,782	165	1904	[38]
Triticum aestivum	203		17	5	8	[39]