UNIVERSITY OF CALIFORNIA, SAN DIEGO

Proteogenomics: Applications of mass spectrometry at the interface of genomics and proteomics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Natalie Castellana

Committee in charge:

Professor Vineet Bafna, Chair
Professor Steven P. Briggs
Professor Pavel A. Pevzner
Professor Laurie G. Smith
Professor George Varghese

2012

The dissertation of Natalie Castellana is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

DEDICATION

To my parents, Philip and Susan, and my indefatigable husband, Ryan.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My graduate career has in large part been a series of happy coincidences, mainly due to the excellent people I've met along the way. First and foremost I must acknowledge my advisor, Vineet Bafna, without whom I would not have had the opportunity to work on such an exciting topic. I also acknowledge my IGERT co-advisor Steve Briggs, who expected no less than my best and pushed me to make the most of my data.

As a young graduate student, I had the pleasure of working with several mentors in the lab, particularly Sam Payne and Stephen Tanner. I must also recognize lab members and collaborators who helped me work less and take myself less seriously; Ali Bashir, Ari Frank, Sangtae Kim, Nitin Gupta, Nitin Udpa, Anand Patel, Adrian Guthals, Jocelyne Bruand, Sunghee Woo, Kyowon Jeong, Christina Boucher, Roy Ronen, and the entirety of the Bafna, Pevzner, Bandeira, and Briggs labs.

Chapter 2, in full, was published as "Proteogenomics to discover the full content of genomes: a computational perspective". NE Castellana and V Bafna. *Journal of Proteomics*, vol. 73, no. 11, pp. 2124-2135, 2010. The dissertation author was the primary author of this paper. The dissertation author reviewed the relevant papers in the field, and wrote the paper.

Chapter 3, in full, was published as "Discovery and revision of Arabidopsis genes by proteogenomics". NE Castellana, SH Payne, Z Shen, M Stanke, V Bafna, and SP Briggs. *Proceedings of the National Academy of Science*, vol. 105, no. 52, pp. 21034-21038, 2008. The dissertation author was one of three primary authors of this paper. The dissertation author implemented computational methods, ran the computational analysis, and wrote the paper.

Chapter 4, in full, is in preparation for publication as "Novel gene finding in *Zea mays* using an automated proteogenomics pipeline". NE Castellana, Z Shen, Y He, J Walley, LG Smith, SP Briggs, and V Bafna. in preparation The dissertation author is the primary author of this paper. The dissertation author developed the computational methods, ran the computational analysis, and wrote the paper.

Chapter 5, in part, was published as “Template Proteogenomics: sequencing whole proteins using an imperfect database”. NE Castellana, V Pham, D Arnott, JR Lill, and V Bafna. *Molecular & Cellular Proteomics*, vol. 9, no. 6, pp. 1260-1270, 2010. The dissertation author was the primary author of this paper. The dissertation author developed the algorithm, implemented the algorithm, ran the experiments, and wrote the paper.

VITA

2006	Bachelor of Science in Computer Science, Carnegie Mellon University, Pittsburgh
2009	Master of Science in Computer Science, University of California, San Diego
2012	Doctor of Philosophy in Computer Science, University of California, San Diego

PUBLICATIONS

Natalie E Castellana, Zhouxin Shen, Yupeng He, Justin Walley, Laurie G Smith, Steven P Briggs and V Bafna. Novel gene finding in *Zea mays* using an automated proteogenomics pipeline, in preparation

To-ju Huang, Claudiu Farcas, Jeremy Carver, Natalie Castellana, Ari Frank, Sangtae Kim, Jian Wang, Xiaowen Liu, Pavel A. Pevzner, Vineet Bafna, Ingolf Krüger, and Nuno Bandeira. ProteoSAFe: A Scalable, Accessible, and Flexible Software Environment for Proteomics Analysis, in preparation.

Natalie E Castellana, Krista McCutcheon, Victoria C Pham, Kristin Harden, Allen Nguyen, Judy Young, Camellia Adams, Kurt Schroeder, David Arnott, Vineet Bafna, Jane L Grogan, and Jennie R Lill. Resurrection of a clinical antibody: template proteogenomic de novo proteomic sequencing and reverse engineering of an anti-lymphotoxin-alpha antibody. Proteomics, 11, 395-405, 2011.

Natalie E Castellana and Vineet Bafna. Proteogenomics to discover the full content of genomes: a computational perspective. Journal of Proteomics, 73, 2124-2135, 2010.

Natalie E Castellana, Victoria Pham, David Arnott, Jennie R Lill, and Vineet Bafna. Template Proteogenomics: sequencing whole proteins using an imperfect database. Molecular & Cellular Proteomics, 9, 1260-1270,2010.

Natalie E Castellana, Samuel H Payne, Zhouxin Shen, Mario Stanke, Vineet Bafna, and Steven P Briggs. Discovery and revision of Arabidopsis genes by proteogenomics. Proceedings of the National Academy of Science, 105, 21034-21038,2008.

Natalie E Castellana, Kedar Dhamdhere, Srinath Sridhar, and Russell Schwartz. Relaxing Haplotype Block Models for Association Testing. Proceedings of the Pacific Symposium on Biocomputing, 454-466, 2006.

ABSTRACT OF THE DISSERTATION

Proteogenomics: Applications of mass spectrometry at the interface of genomics and proteomics

by

Natalie Castellana

Doctor of Philosophy in Computer Science

University of California, San Diego, 2012

Professor Vineet Bafna, Chair

Proteins are understood to be the main workhorses in the cell, participating in a wide variety of activities from cell structure to inter- and intra-cellular transport. Through improvements in sample preparation and instrumentation, mass spectrometry has become a popular, efficient, high throughput technology for studying protein expression.

The standard protocol for a mass spectrometry experiment includes digestion of the sample proteins into peptides that are subsequently analyzed by the mass spectrometer to produce tandem mass spectra. An important initial step in characterization of the sample is the identification of the peptide precursor of each spectrum. This routinely involves the comparison of the experimental spectrum to the theoretical spectrum associated with a peptide sequence contained in a protein sequence database. Publicly available protein sequence databases are believed to be complete for well understood, model organisms. However, in this thesis we demonstrate that even for organisms that receive extensive attention, the databases are missing a significant fraction of expressed proteins. We describe two situations in which a comprehensive protein sequence database is not available for peptide identification and propose methods for addressing the issue.

Determining the nucleotide sequence that comprises an organisms genome is only the first step to understanding the molecular basis for its phenotype. Genome annotation is required to determine the function of each nucleotide, including nucleotides that encode the blueprint for proteins. We present a semi-automated pipeline that accepts mass spectra and the sequenced genome, and addresses the dual goals of annotating the genome for protein-coding genes and identifying peptide sequences in the absence of a complete, curated protein sequence database.

The pipeline mentioned above for genome annotation, assumes that the genome is immutable. Immunoglobulins, proteins involved in our adaptive immune systems, require rearrangements in the genome resulting in a different immunoglobulin gene sequence in nearly every B-cell in the body. We build on the ideas of genome annotation to construct a database to represent the complement of possible immunoglobulin gene sequences in an organism. In addition, we move beyond the goal of peptide identification to sequence entire proteins.

Chapter 1

Introduction

Proteins play an astonishing variety of roles in the cell, from transporting vital molecules to the regulation of protein expression. While proteins have been studied using low-throughput methods for decades, mass spectrometry (MS) has emerged as a key technology for high-throughput analysis of protein expression [1]. Improvements in speed and sensitivity in the past two decades have led to large-scale profiling of the proteome in many organisms at a variety of developmental stages, environment conditions, and disease states. The omnipresent question of which proteins are present in a given sample has been addressed by numerous computational methods and received much attention from biologists, chemists, computer scientists, and statisticians [2]. Here, we explore the use of MS data to answer this question in a non-traditional context; when the complement of possible proteins is unknown.

A protein's primary structure is defined by a string on a 20 letter alphabet. Each letter represents an amino acid. There are two dominant methods for identifying a protein in a sample each with its own unique set of capabilities and limitations. In top-down MS, an intact protein is analyzed by a mass spectrometer resulting in a collection of tandem mass spectra, each one corresponding to a protein sequence. Though these mass spectra are very complex, great strides have been made in computational methods for analyzing them [3]. While a single top-down spectrum contains information about an entire protein, it is considerably lower throughput than other technologies, must be run on a mass spectrometer with high mass accuracy, and cannot be used to analyze insoluble proteins. In contrast, bottom-up mass spectrometry circumvents many of the physio-chemical limitations of top-down MS by first digesting the protein sample into short peptide sequences of between 10 and 20 amino acids. The peptides are then analyzed by the mass spectrometer. Analyzing the proteome in this bottom-up fashion, that is characterizing a protein's state through the collection peptides identified, also has several limitations. A particularly challenging problem is inference of proteins from the peptides. Due to redundancy in the proteome, many peptides may match to multiple proteins creating ambiguity in the inference. Second, in a standard mass spectrometry protocol that digests the proteins with a single enzyme, we only observe a fraction of a protein's amino acids as proteins. In these cases, the inference assumes that the protein sequence in the database is identical to the protein in the sample. When a protein's sequence is important for its characterization, this outcome is unacceptable.

Many computational methods for the bottom-up characterization of a sample rely on the identification of the peptide precursor of each spectrum. The most widely used algorithms [4, 5, 6, 7] for performing peptide identification compare the experimental spectrum to a theoretical spectrum associated with a peptide sequence contained in a protein sequence database. Publicly available protein databases are available for many organisms, and are often assumed to be complete. However, we found that even for well-studied organisms like *Arabidopsis thaliana*, the databases are missing a significant fraction of expressed proteins. In this thesis, we describe two situations in which a comprehensive protein sequence database is not available for peptide identification.

For many years, scientists believed that sequencing the entirety of the human genome would reveal the genetic basis for observed phenotypes. However, even after decades of analysis, the function of many genomic regions eludes us. Genome annotation has the goal of identifying the function of each nucleotide in the genome, including nucleotides that encode the blueprint for proteins [8]. Sophisticated annotation pipelines have emerged [9, 10], which aggregate many sources of evidence to arrive at a final prediction of genes and their protein products. These sources range in throughput, accuracy, and cost. *Ab initio* predictors [11, 12, 13], which rely solely on genomic signals, are by far the cheapest, fastest, and least accurate method of annotation. Expression-based assays, such as RNA-Seq [14] or expressed sequence tags (ESTs) [15], are also very popular. This evidence requires some laboratory work, but sequencing technologies have been improving faster than MS technologies and are capable of generating large datasets in a very short time. However, for protein-coding gene finding, transcript-based annotation introduces many false positives from unspliced mRNAs or non-sense transcripts. Another source of evidence is homology. Conservation is a strong indication of function, and when a related species's proteome has been characterized, the proteins can be mapped to the target genome to identify likely protein-coding regions.

Few annotation efforts to date have utilized MS data to improve the annotation of protein-coding genes. A crucial step for incorporating evidence from MS data into a genome annotation pipeline is creating a database for peptide identification. Since the goal is to identify proteins that are not already annotated, relying on a curated proteome is not applicable. Instead, we construct a database of putative peptide sequences generated directly from the genome; the six frame translation and a splice graph [16]. These databases are much larger than the curated proteome databases, presenting challenges in search speed and sensitivity. In addition, the genomic databases only contain fragments of proteins, causing the problem of protein inference becomes more challenging. In Chapter 2, we review the state-of-the-art in proteogenomics for genome annotation and emphasize the benefit of including MS data as an orthogonal source of evidence for gene finding. We identify key challenges which must be addressed to successfully incorporate proteomics into gene finding pipelines.

In Chapter 3, we describe our early efforts in improving the gene annotation of the model organisms *Arabidopsis thaliana*. In many ways, arabidopsis represented an easy first case for developing our pipeline; it has a genome of manageable size (200 million nucleotides) that contains little sequence redundancy. In addition, the number of tandem mass spectra included in the study could be analyzed in under a month on a computer cluster. Working against us was the fact that arabidopsis is a model organism for plant biologists and had undergone seven iterations of genome annotation. Despite the extensive efforts in gene finding, we were still able to generate 1,473 gene models that were either entirely new loci or revisions of annotated genes.

From lessons learned in improving the annotation in arabidopsis, including calculation of false-discovery rates at the level of a novel gene, we developed a semi-automated pipeline that accepts tandem mass spectra and a sequenced genome. In Chapter 4 we describe the pipeline applied to *Zea mays*. Compared to arabidopsis, maize presented a new set of challenges. First, the maize genome is an order of magnitude larger (2 billion nucleotides). The dramatically expanded genome is largly due to sequence duplications caused by transposable elements [17]. Our pipeline addresses the dual goals of annotating the genome and identifying peptide sequences in the absence of a complete, curated protein sequence database.

There are several natural extensions to the pipeline which unfortunately did not make it into this thesis. As genome sequencing becomes cheaper and faster, organisms are being sequenced at an incredible rate. One consequence of this trend is an increased reliance on computational tools, that (not to discredit my own field) tend to make many more errors than their slow but meticulous human counterparts. As a result, many genomes may contain errors in the form of incorrect nucleotides, inverted regions, or missing regions.

Another challenging class of organisms are those with dramatic genetic diversity. While the maize and arabidopsis plants used for our genome annotation efforts were believed to be clonal (exact copies of one another), many species like humans, exhibit nucleotide variability.

Both of the above described challenges boil down to sequence differences between the peptide sequence in the sample and the peptide sequence in the database. The pipeline described in Chapters 3 and 4 assumes that not only is the genome sequence complete, but it also is error-free. To solve both problems the peptide identification portion of the pipeline could be altered to allow the substitution of one amino acid for another. In the case of a missing genomic region, a homology sequence (which would contain some nucleotide differences from the target organism) could be used. Allowing sequence differences between the genomic databases and the tandem mass spectra greatly increases the rate of erroneous peptide identifications. We could potentially control the error by leveraging the information provided in transcriptomic experiments.

Previously, I mentioned the challenge of protein inference in bottom-up experiments. While peptides provide incomplete information about the proteins present in a sample, top-down mass spectrometry, which generates mass spectra from whole proteins, can unambiguously determine which proteins are expressed. By itself, top-down MS could not provide the proteome coverage necessary annotate a full genome. However, coupling bottom-up and top-down mass spectrometry could prove to be a useful technique for maximizing proteome coverage while at the same time improving protein inference. Since it is likely that the genomic databases do not contain all full length protein sequences, new algorithmic methods for searching the databases with top-down mass spectra would be required.

The genome annotation pipeline described in Chapters 3 and 4 , assumes a static genome that is identical in all cells of an organism. However, in organisms with adaptive immune systems such as humans, genome rearrangement plays a crucial role in defense against infection. Immunoglobulins, also referred to as antibodies, are proteins involved in recognizing pathogens and rely on protein sequence diversity to properly perform their functions. For that reason, the immunoglobulin gene in mature B-cells are different both from each other and from the germline sequence present in other cells. Each immunoglobulin gene is created through a genome rearrangement which brings together in a combinatorial fashion disparate regions of the genome. In Chapter 5, we build on the ideas of genome annotation to construct a database to represent the complement of possible immunoglobulin gene sequences encoded in an organism's genome. To introduce further sequence diversity, immunoglobulin genes also undergo hypermutation. For that reason, inference of a complete protein sequence from a few peptides is insufficient to properly characterize the immunoglobulin. Instead, the goal of this chapter is to identify the entire protein sequence.

Both applications, MS-based genome annotation and antibody sequencing, require a departure from conventional MS studies which rely heavily on a protein database. Instead, we propose solutions which identify peptide sequences using the genome as a template. Research of this type, at the junction of proteomics and genomics has been aptly named *proteogenomics*.

Chapter 2

Incorporating mass spectrometry into gene finding: a review

2.1 Why Proteogenomics? a primer on gene finding

Scientific progress is often associated with abstraction and compaction of available knowledge, so as to create a foundation on which future discoveries can be made. Our understanding of the gene has unfortunately bucked this trend. The recently concluded ENCODE project resulted in further ambiguity of the concept. The classical definition of the gene being a "unit of heredity" (from Mendel's work), has now evolved into "... a union of genomic sequences encoding a coherent set of potentially overlapping functional products" [18]. Other examples point to the plasticity of the genome itself, with large genomic rearrangements disrupting genes on the genome [19]. All of this has implications for proteomics.

Historically, the genomics and proteomics communities acted independently. It was the role of the genomics community to identify genes and the corresponding protein sequences. This was often done through large-scale annotation efforts, during and after the sequencing of the genomes (see below). The collection of derived proteins was considered to be a fixed set, although it was recognized that not all proteins are expressed in every cell. It was the role of the proteomics community to understand which proteins are expressed under specific conditions, or tissues, and to identify the various post-translational modifications, and other processing of the proteome. Proteogenomics challenges this perspective: if the definition of the gene itself is not clear, the proteomics (mass spectrometry) and genomics communities should work together from the beginning, to clarify gene structures. Therefore, a good place to start is to look at gene-finding.

2.1.1 Gene Structure

The central dogma of molecular biology suggests a flow of information from DNA to protein. First, the genic region of the DNA is 'transcribed' (copied) into mRNA (also called a transcript, or message). Special 'promoter', 'enhancer', and 'repressor' regions proximal to the gene help switch the transcription on and off, thereby regulating the production of protein. Next, the ribosomal machinery reads the message, and 'translates' it into proteins. Note that the beginning and end of the transcript are not translated and are referred to as the 5' and 3' untranslated regions (UTRs), respectively. While the process of protein production is common to both prokaryotic and eukaryotic organisms, the structure and organization of genes is quite different between the two, and will be discussed separately.

Prokaryotic Genes In prokaryotes, related genes may be clustered into operons (Figure 2.1A). All genes in an operon share the same promoter region, and are transcribed as a single mRNA. While the transcript produced by an operon contains mRNA from all the genes, regulation at a higher granularity occurs at translation. Even with this simple structure, there is genic diversity. Programmed frame-shifts can produce alternate or truncated proteins [20], but are nearly impossible to predict from genomic data.

Eukaryotic Genes In eukaryotes, the coding regions of the gene are often present in discontinuous regions called *exons*. Multiple exons are separated by *introns*: regions that are transcribed, but not translated (Figure 2.1B). Introns

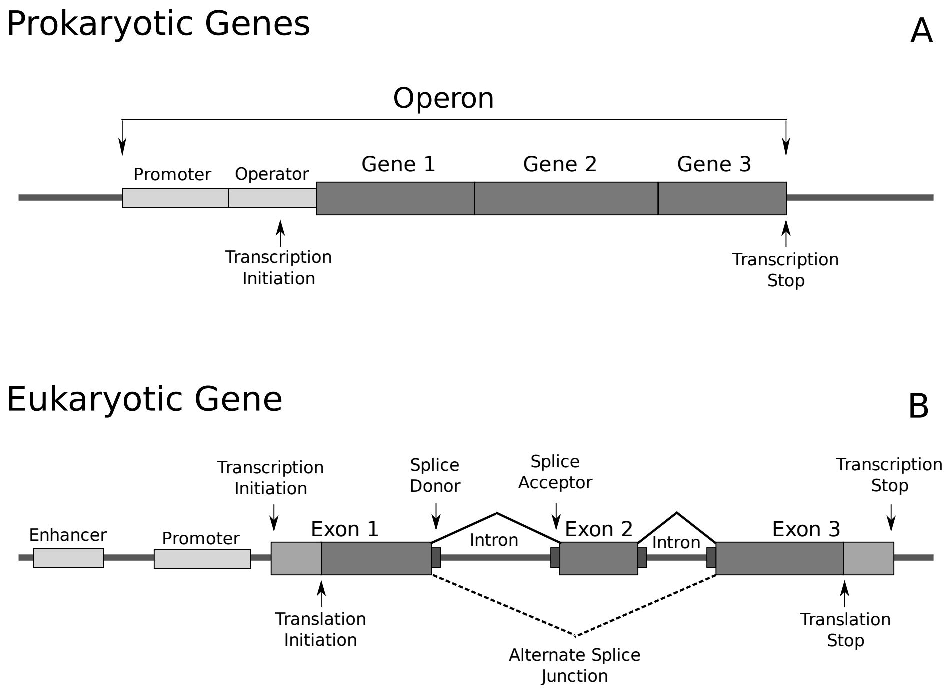


Figure 2.1: A: Prokaryotic genes may be arranged in an operon, sharing the same promoter. B: A eukaryotic gene contains protein-coding regions called exons, separated by non-protein-coding regions called introns. Once transcribed, the introns are spliced out. An alternate splice junction is shown using a dotted line.

are spliced out of the mRNA prior to translation by an RNA-protein complex called the spliceosome, producing the mature mRNA. For a given transcript, there may be alternate splice patterns each of which produces a different mature mRNA and may cause the protein sequence to be altered.

2.1.2 Gene Annotation

The goal of gene finding can be roughly stated as the process of identifying the genomic coordinates of exons, and the splicing patterns. Here we focus only on the coding exons. Established methods of gene annotation today combine evidence from multiple orthogonal sources [21]. One form of evidence is from *ab initio* gene predictors that analyze genomic signals for coding exons and splice junctions. In addition, large-scale transcript sequencing projects (often in the form of expressed sequence tags, or ESTs [22]) yield cDNA sequences that can be mapped against the genome to identify coordinates. Finally, evolutionary conservation with related species is often suggestive of genes, and other functional elements [23]. Even so, gene finding is challenging. The recent release of the Arabidopsis Information resource (TAIR8 to TAIR9) modified over 1,000 genes and added 282 new protein-coding loci [24]. Even with the well-studied human genome, a manual investigation by the ENCODE Consortium [25] resulted in the assignment of additional exons to 80% of studied genes.

Predicting the beginning of translation is a major challenge [8] for current annotation pipelines. Translation start is usually marked by one of a handful of canonical start codons, the most common of which codes for the amino acid methionine. Studies have shown that this is not a hard rule, with many non-standard start sites observed in prokaryotes [26]. Eukaryotic gene annotation is further complicated by the prevalence of alternatively spliced genes, which play a key role in generating proteome diversity. The reliable detection of splice-junctions is difficult, and most *ab initio* gene finding algorithms only predict a single transcript at a locus, ignoring completely alternate splice isoforms. Several tools have been developed to identify alternative splice variants using ESTs [27, 28], but accurate annotation remains a challenge due to intrinsic problems in EST sequencing including incomplete gene coverage, low sequencing accuracy, and chimerism. The issue of determining whether the alternative transcripts differ in protein-coding regions or UTRs also remains.

While gene annotation efforts for model organisms rely heavily on transcript sequencing, recent studies suggest that evidence of transcription might not be evidence of translation. Clamp *et al.* [29] suggest that approximately 4,000 genes in human do not code for protein despite cDNA evidence, citing their lack of conservation with primates. Genomic signals, which are the primary component of *ab initio* gene predictors, may be equally misleading. For example, the codons ‘TGA’, ‘TAA’, and ‘TAG’ are very strong indicators of translation stop. However, in order to accurately determine translation stop, the frame for the terminal exon must be correctly predicted. Coding signals, based on hexamer compositions, are not sufficient to determine frame in short exons [30], and sometimes cause annotators to miss the exon completely. Similar challenges make it hard to identify short genes (< 100 amino acids), which constitute a significant portion of protein-coding genes [31]. Even for longer genes, differing GC composition change coding signals to the point that the tools have to be retrained for each new genome [32].

2.2 The promise and challenge of proteogenomics

The idea of searching un-interpreted mass spectra against a translated nucleotide database is hardly new. We see an early account in the paper by Yates *et al.* [33]. However, the true power of the approach comes from a holistic use of these peptides in gene finding. See Figure 2.2. A proteogenomically identified peptide provides unique information for gene annotation by (a) confirming translation and separating pseudogenes(see below) from coding genes [34]; (b) establishing that a protein is not targeted for degradation (c) automatically determining frame, even multiple overlapping frames; (d) constraining the location of the translation start and end sites, as well as sites of post-translational processing (e.g. signal cleavage); (e) identifying exact splicing boundaries and alternative splice-forms, if the peptide is split across exons; and, finally (f) predicting a completely novel gene, by mapping to an uncharacterized genomic location. One may argue that next-generation sequencing of transcripts is a more promising technology for sampling the translated genome for the purpose of gene annotation. However, recent studies suggest that many transcripts are targeted for nonsense-mediated decay [34], or upon translation are unable to form stable, functional proteins[29]. Indeed, the optimist in us would claim that proteogenomics is the panacea for the ailments that plague gene annotation. Proteomic analysis also carries beneficial side-effects like peptide abundance calculations, and the identification of post-translational modifications.

However, proteogenomic studies are not without substantial challenges. First, peptide identification is arguably more error-prone than matching cDNA. Incomplete fragmentation, noise, and ‘isometric’ peptides can all lead to erro-

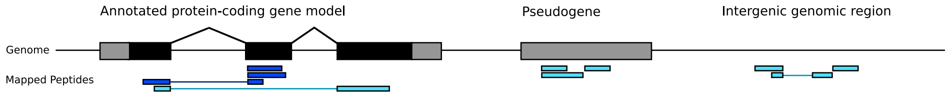

Annotated protein-coding gene model
Pseudogene
Intergenic genomic region
Genome
Mapped Peptides

Figure 2.2: Several peptides fall within an annotated gene locus, validating translation of two exons and an intron. One peptide indicates a novel splice isoform which skips the internal exon. Additional peptides fall within an annotated pseudogene giving strong indication for its translation in the cell. Peptides which fall within the intergenic region may indicate novel protein-coding loci. Cyan colored peptides would likely not be identified using a standard proteomic database.

neous identification. The problem is compounded for proteogenomics as genomic databases are much larger than existing protein databases. A 6-frame translation of the human genome has 6 billion residues in it; contrast that with 180Mb needed for the UniProt database [35] consisting of proteins from over 250 organisms. The number of spectra for single proteogenomic studies is also large, often on the order of tens of millions of spectra.

Second, sampling and dynamic range is a concern in nearly all mass spectrometry studies. Current techniques can reliably sample proteins over 3-4 orders of magnitude [1], which is smaller than the estimated true abundance range (~ 6 orders) [36]. Detectability of peptides is a concern as not all peptides show up in mass spectrometric scans due to chemical attributes unfavorable to ionization or fragmentation.

Peptides which span splice junctions contain a wealth of information useful to gene structure prediction. In the ENSEMBL [37] database of human proteins (assembly GRCh37, release 57), approximately 26% of fully tryptic peptides of length 7 or greater span a splice boundary. These peptides are invaluable, as reliable prediction of splice-junctions is a major challenge for gene finding. However, identifying split peptides using proteogenomics seems to be equally challenging, if not more.

Finally, the output of proteogenomics is peptides, and peptides are not complete genes. Determining the gene structure from incomplete coverage by peptides is a difficult task. While these are all valid criticisms, we assert below that recent developments in technologies and computation are tipping the balance.

2.3 The proteogenomics solution (eukaryotes)

In the wake of technological advances in DNA sequencing, the number of eukaryotic genomes sequenced has increased dramatically in the past 20 years, with full genomes available for *Saccharomyces cerevsiae* [38], *Caenorhabditis elegans* [39], *Arabidopsis thaliana* [40], *Drosophila melanogaster* [41], *Homo sapiens* [42, 43], *Anopheles gambiae* [44], and recently, *Zea mays* [17]. As the genome sequences of many model organisms become available, so too are large-scale proteogenomic studies beginning to fill a much needed niche in gene annotation. In the past ten years, proteogenomic studies have confirmed expression of 25% of ORFs in Yeast [31], 73 transcripts in A. gambiae [45], 9,124 transcripts in D. melanogaster [46], 224 hypothetical proteins in Human [16], and over 13,000 transcripts in Arabidopsis [47, 48]. Peptides identified in these studies provide validation of putative genes. These successes are due in part to recent developments described below.

Sampling the proteome Achieving broad coverage of the proteome is critical to constructing a complete and accurate catalog of genes. A distinct set of proteins is expressed by cells of different tissues or conditions, and sampling each reveals a unique cross-section of the proteome. By acquiring multiple biological replicates of samples from different organs [47, 46, 16, 48] and developmental stages [46] a wider range of proteins can be identified. While broadly sampling the proteome increases the number of proteins detected, absence of peptides from a protein cannot be used as an indicator for absence of the protein in the sample. As Figure 2.3 shows, the number of unique peptides identified in Arabidopsis nearly doubled with a broader sampling strategy [48]. Coupled with technological advances in the form of high-resolution mass spectrometers [36], spectra can be collected from peptides with a wider dynamic range, while providing accurate spectral information for downstream analysis. Improved protein separation techniques [49] have also enabled the identification of more peptides in a single mass spectrometer run. In addition, fractionation methods can be used to isolate underrepresented subsets of the proteome such as small proteins [31], phosphoproteins [48], and basic proteins [46]. Brunner *et al.* achieved coverage of 63% of the *Drosophila melanogaster* proteome by utilizing these techniques as part of an 'analysis-driven experimentation feedback loop'. They used the analysis of previous data sets to determine categories of the proteome where their sampling was deficient.

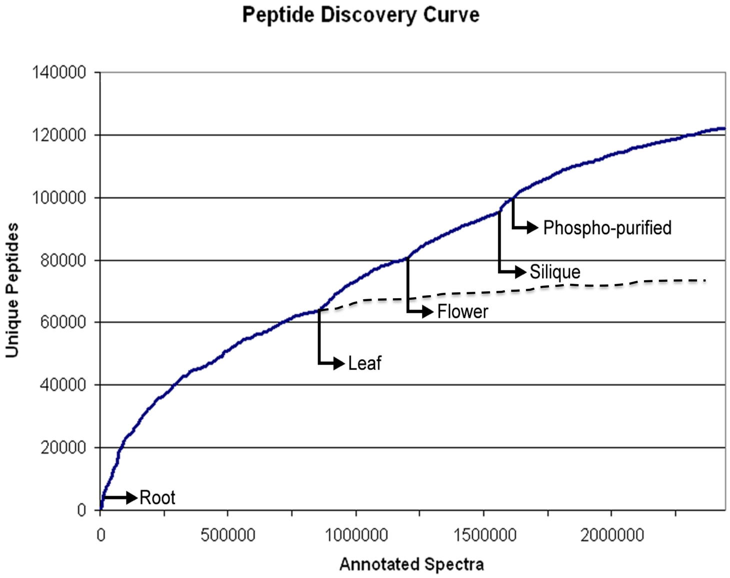


Figure 2.3: A graph showing the discovery curve for Arabidopsis peptides [48]. The x-axis in the figure is the number of spectra considered, while the y-axis is the number of distinct peptides recovered from the spectra. As spectra were added to the experiment, the rate of distinct peptide sequences identified slows. The figure suggests that including more spectra from root tissue will not substantially increase the number of peptides identified. However, including spectra from a different tissue type provided additional distinct peptides. Extrapolation suggests that the number of distinct peptides identified is nearly doubled by including additional tissues and additional fractionation

.

Error rates The problem of erroneous identifications is common to all proteomics projects, but is magnified for proteogenomics. Searching large spectral data sets ($\sim 10M$) against large databases ($\sim 1B$ AA) translates into a large number of erroneous identifications even at a low error rate. At the same time, the evidence for a new gene (usually low-abundance) might only come from a small number of peptides.

Algorithmically, the identification problem is broken up into two parts: *scoring/ranking* of each candidate peptide for a spectrum, so that the correct identification gets the top score, and is well separated from the others; and, *validation*, which provides confidence that the top-scoring peptide is the right identification. Scoring has seen great improvement, based on probabilistic models [50, 51, 52, 53, 54, 55, 56] for peptide fragmentation. Large data sets of annotated peptides allow for a systematic data-mining of fragmentation patterns, which are then encoded into scoring models.

To understand why a secondary validation step is needed, consider the case when the correct peptide is not in the database, and never considered. Even if the ranking of candidates is perfect, the validation part is needed to reject the top scoring peptide. To assess the likelihood of the top-scoring peptide being the correct identification, parametric and non-parametric approaches have been tried. In the model-based approach, it is suggested that the correct and incorrect peptides follow a distinct distribution of scores. By modeling the two distributions, once can use a statistical test to identify the correct peptides [57]. Gygi and colleagues suggest a model-free approach based on constructing a decoy database [58]. The spectra are simultaneously searched against the standard and the decoy database (typically a scrambled version of the standard database). Peptides identified using the decoy are all spurious and can be used to estimate the false discovery rate (FDR).

An issue with FDR is that all peptides that exceed the score cutoff are treated equally (have the same FDR). However, we know that higher scoring peptides are more likely to be correct. Additionally, our confidence in peptide identification also depends upon its size, charge, and physico-chemical characteristics. One direction to improve FDR is to bin peptides that are similar (by score, size, charge, etc.), and compute FDR separately for each bin [48]. This *local* FDR

(ℓ-FDR) [59] computation is possible in proteogenomics, where the large number of peptides allow each bin to be populated. Second, the goal of proteogenomics is to find protein-coding regions, not just peptides. If 2-3 peptides support the same genic locus, or gene model refinement event, then the error occurs only if all of the identifications are wrong. In this case, the ℓ-FDR values of these peptides (under the assumption that the identifications are independent) can be multiplied to give an event level FDR (eFDR). The generic approach is as follows: a list of proteogenomic events is created, such as 'spliced exons', 'translated ORF'. Each event is supported by a collection of peptides and their associated ℓ-FDR values. A Bayesian approach is used to compute eFDR values for the event being incorrect [48]. For example, a spliced peptide may have a high probability of being correct, but makes a small contribution to a splicing event because of a small overlap with the second exon. On the other hand, a small collection of peptides that hit two exons, with a few spliced-peptides is strong evidence for a splicing event, even when each of the peptides has a poor ℓ-FDR score. In the set of 591 gene models submitted by Castellana *et al.* to TAIR, a lower eFDR correlated with manual curation and acceptance into TAIR9 resource.

The decoy database approach, and its variations, have been widely adopted. However, critics point out that including the decoy database doubles the search time and, more importantly, the FDR values greatly depend on the size of the database and the distribution of peptides in it. The most obvious difficulty is in the construction of a decoy database. One desirable attribute of the decoy database is that it does not share peptide sequences with the target database. This becomes a difficult task when the target database exceeds 6 billion amino acids.

Another difficulty stems from the construction of the target database. In proteogenomic studies, a major goal in constructing a database from genomic data is to include as many putative protein sequences as possible. This often is done by performing a translation of the genome in all six-frames. In addition to containing all putative proteins sequences, the resulting database also contains spurious sequences, often at a much higher rate than standard proteomic databases. This implicit addition of decoy peptides in the database results in diminished sensitivity at the same FDR. To combat this, an additional goal of constructing proteogenomic databases is compactness. For example, the six frame translation can be restricted to sequences exceeding the average size of an exon in the organism of interest, or to regions which receive high scores by *ab initio* gene predictors.

One might argue that instead of attempting to construct a database of putative protein sequences, interpreting the peptide sequence *de novo* will guarantee that any possible peptide sequence is considered. Several groups have proposed database-independent p-value computation methods [60, 61]. These methods rate peptide-spectrum matches using spectrum-specific score distributions, but make the assumption that all peptides are equally likely *a priori*.

This argument illustrates a philosophical difference regarding the importance of prior information (the database) in peptide identification, and is analogous to the debate between Bayesian and frequentist inferences. *De novo* approaches to peptide identification seek to distinguish the correct peptide among *all* possible peptides, and consequently are highly error-prone. In a database search, the space of candidate peptides is greatly reduced. This automatically increases the confidence in identification, but only if we agree that the database is complete. In proteogenomic studies, the search is on ever larger genomic databases, and the differences between *de novo* and database searches diminishes, particularly when modifications and mutations are permitted. If we consider only the set of peptides of length 9, *de novo* algorithms must consider $20^9 < 10^{12}$ candidate peptides. The six-frame translation of the human genome contains about 10^9 peptides. However, if we allow a single mutation per peptide (which would have no effect on *de novo* algorithms), the size of the genomic database increases 200 times, to 10^{11}. For large databases typical of proteogenomics, the boundary between *de novo* and database search is blurred. Confident assessment of a genomic region being translated must depend upon the discovery of multiple, large peptides with accurate fragmentation patterns.

Another aspect of large genomic databases is the non-random distribution of nucleotides. In fact, segmental duplications, and retrotransposon mediated elements, often create multiple copies of the same gene [43, 42], leading to identical peptides at multiple locations in the genome (shared peptides). Sometimes, only one of the gene copies is active. The inactive genes (possibly transcribed into RNA, but not translated), are called pseudogenes, and will cause problems for proteogenomic identification. Identifying pseudogenes is one of the major challenges for gene annotation.

One approach to handling genomic redundancy is to consider all locations of the shared peptide [45]. However, this may lead to the false reporting of proteins. A stricter approach is to ignore the shared peptides [62, 47], significantly reducing the number of protein identifications. Grobei *et al.*, developed a classification method of peptides based on their occurrence in the database. Peptides which could uniquely identify a protein sequence were labeled Class 1. Peptides which mapped to multiple locations were classified depending on whether their matches were to isoforms of the same gene (Class 2), members of the same gene family (Class 3a), or from multiple gene families (Class 3b). In the study by Grobei *et al* [63], all Class 3b peptides were discarded. Other groups [48] have used peptide locality to decide whether to keep or discard shared peptides. If a shared peptide appears in close proximity to a uniquely-located peptide, the additional identification boosts the confidence in the shared peptide location.

Spliced peptides In humans, approximately one quarter of peptides cross a splice junction. These peptides are especially informative for gene annotation, giving boundary and frame information for two exons and a splice junction. The identification of these spliced peptides is a challenge unique to eukaryotic proteogenomics.

Historically, peptides identified against predicted or known proteins were mapped back to the genome to validate splicing events [64, 65, 16]. The detection of new splice-junctions, however, calls for a special database that encodes putative splice-forms. Such a database can be constructed using ESTs which are produced from mature mRNA and have the introns already spliced out, thus enabling the identification of peptides which span the intron boundaries [62, 66]. However, ESTs are error-prone, and highly redundant. Edwards proposed a compression scheme for reducing EST database size, based on a de Bruijn graph representation of cDNA fragments [66, 67], while retaining all potential peptide sequences.

Even with large sampling efforts, ESTs do not adequately cover all splice-junctions since many ESTs are sequenced from the $3'$ end which provides unique tags for identification, but only limited representation of the coding sequence. A second source of putative spliced sequences is *ab initio* gene prediction tools, such as GeneID [68], Fgenesh [11], Augustus [69], and GeneMark [13]. Kuster *et al.* [70] used a two-pass system to identify spliced peptides by first identifying likely novel coding regions using unspliced peptides, then predicting a new model for that region and searching the spectra against the new model.

Tanner *et al.* [16] proposed a spliced-exon graph to compactly represent all gene structures and splice-junctions generated by gene prediction tools and EST mappings. In the graph, each exon is a node and each edge between exons represents a putative splice junction. While the graph provides a compact encoding of all splice-forms, the MS2 identification tools need to be modified to search the specialized database. In recent studies, this approach confirmed over $15,000$ spliced peptides in human, including over 40 instances of alternative splicing, where peptides confirm the splicing of one exon with multiple partner exons [16]. In arabidopsis, $4,018$ novel spliced peptides were identified when compared to the TAIR7 annotations [48]. By structuring proteogenomic databases as spliced-exon graphs and de Bruijn graphs, the sequence redundancy that is inherently present in the proteomes of higher organisms is reduced. This is of particular importance in proteogenomics where database size has significant impact on error rates and search time.

The general issue of identifying ‘discontinuous’ peptides (of which spliced-peptides are a special case) is likely to persist. The genomes of individual humans are now being sequenced [71], and show remarkable plasticity, with large rearrangements leading to gene disruptions, fusions, and trans-splicing [19]. Additionally, the rearrangements often characterize the transition of a normal genome to a tumor genome [72, 73]. Identification of discontinuous peptides confirming fusion events is likely to expand the role of proteogenomics in cancer detection and therapy. It is important to note that in the case of diseases which result from genome rearrangements, such as cancer, that the reference genome should not be limited to the wild type individual. Proteogenomic studies, coupled with deep genomic and transcript sequencing, can provide valuable information on aberrant protein expression.

Search speed In 2001, Choudhary *et al.* [74] constructed the 6-frame translations of the human draft genome sequence. On a single processor, searching 169 spectra required 10 hours of compute time. Since then search algorithms and computing resources have improved greatly, while the framework of proteogenomic studies has remained largely unchanged. Filtering spectra for quality [75, 76], or clustering them to increase the signal to noise ratio of each spectrum [77] are techniques employed to improve both the quality of identifications, as well as the speed of the search. A second advance is in 'database filtering', where a two-pass search is employed for MS2 identifications. The goal of the first search (the filter step) is simply to discard most of the database, while retaining the correct peptides, using minimal computation. The more expensive scoring is relegated to a second stage, and is fast because only the filtered peptides are scored. Novel strategies for filtering are under active development, including tag-based filtering, and peak-based filtering resulting in two orders of magnitude speedup, with little loss of sensitivity [6, 7, 53]. Today, many database search engines can be run on multiple cores, or in parallel on large compute clusters. A typical search of 1M spectra against the 6-frame translation of the maize genome containing over 1B amino acids takes on the order of days using a compute cluster of 100 nodes, while also identifying difficult peptides with unexpected modifications and mutations [7, 78].

While technological advances in the form of multi-core hardware and improved database search engines certainly have been enabling factors in proteogenomics, there are biological and experimental tradeoffs which can be made to improve speedup. For example, the database size can be reduced by filtering out unlikely coding regions, such as repeat regions or open reading frames of insufficient length to contain an exon.

2.3.1 Improving gene annotation

The proteogenomic identification of a peptide might come from a region of the genome not previously known to code for protein. We refer to these peptides as 'novel'. Novel peptides might be *intragenic* (fall within the locus of a known gene structure), or *intergenic* (fall outside the locus of a known gene model), and suggest different categories of genome annotation. A set of possible events with supporting peptides is shown in Figure 2.4.

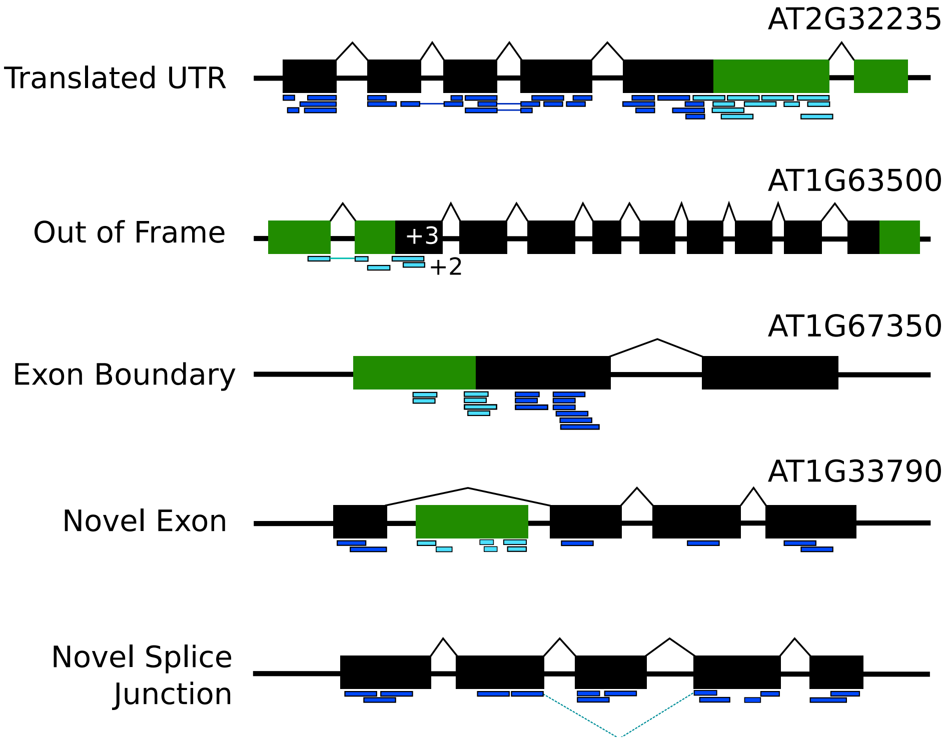


Figure 2.4: Five different refinement events which may be suggested by intragenic peptides. Exons are shown in black boxes while novel coding regions suggested by the peptides are shown in green. Peptides (dark blue and cyan) are shown aligned to the gene models. Examples drawn of refined genes in TAIR7 from Castellana *et al.*[48] are annotated with the updated gene model. Several of these genes have been updated in subsequent gene annotation releases to include the peptide evidence.

Refining gene models For intragenic novel peptides, it is difficult to distinguish if the gene structure needs to be corrected, or if it can be explained by a novel splice-form. The sampling of the proteome is not dense enough to observe peptides from multiple isoforms. Therefore, extrinsic data, such transcript sequences, or homology to genomic regions, is used to distinguish the two cases [47, 48].

Reconstructing gene models using mapped peptides is non-trivial, mostly because the peptide information is not sufficient to completely determine the structure due to limited coverage. While spliced-peptides provide information on which exons might splice together, they are not informative about distal events (isoforms with multiple alternative splicing patterns). Top-down proteomics, in which intact proteins are analyzed, might help in this case, but has not been used for proteogenomics due to the complexity of the samples.

The peptides can be used to increase the likelihood of a gene model being correct. New gene finding tools such as Augustus [69], are able to combine *ab initio* signals with external hints, including homology with related species, ESTs, annotated gene models, and now, proteogenomic peptides. Recent proteogenomic studies have proposed automated prediction of the updated gene model including the peptides as hints [70, 16, 48]. A total of 339 arabidopsis gene models predicted in this way were incorporated into the most recent gene annotation release for Arabidopsis, TAIR9 [24].

Gene discovery Intergenic peptides which are not proximal to a known gene may indicate a novel coding region. To reduce errors, eFDR or ProteinProphet can be used to combine the evidence from multiple peptides in support of the novel gene [79]. Validation of the novel genes remains a challenge, but supporting evidence is obtained from expressed transcript sequence, RT-PCR validation [80, 81] or homology searches of newly predicted gene models [48, 82]. The homology searches can also be used for functional annotation of the corresponding protein sequence [64, 83].

2.4 Proteogenomics in prokaryotes

Bacterial genomes are being sequenced at an astonishing rate, and as a consequence that gene annotations are primarily computational predictions. Prokaryotic genomes tend to be smaller and less genetically complex than eukaryotes. As prokaryotic genes do not undergo splicing, all proteins can be captured by translating the genome in all six frames.

Several studies on prokaryotic genomes have shown that *ab initio* tools alone are insufficient, particularly for identifying gene boundaries, and for short ORFs [84, 64, 26, 82]. Proteogenomic validation of predictions is a pragmatic compromise between computational prediction, and full-experimental validation. Jaffe *et al.* [64] validated 81% of predicted ORFs in *Mycoplasma pneumoniae* and Gupta *et al.* [26] validated 40% of genes in *Shewanella oneidensis.* Wang *et al.* [83] constructed a database of gene predictions, and validated 901 proteins in *Mycobacterium smegmatis.*

The first study to search tandem mass spectra against the 6-frame translation of a fully-sequenced bacterium (*Haemophilus influenzae*) identified 263 proteins and 2 genomic loci which were not previously believed to be translated. Since then, several high-throughput studies have identified novel translated loci in *Mycoplasma pneumoniae* (16 ORFs) [64], *Rhodopseudomonas palustris* (85 ORFs) [80], *Shewanella oneidensis* (8 ORFs) [26], and *Deinococcus deserti* (15 ORFs) [85].

Peptides which map in close proximity to annotated genes may suggest changes to the gene model, rather than separate novel loci. While determining the translation end site is simply the location of the first in-frame, down-stream stop codon, determining translation start is much trickier. Proteogenomic mapping of peptides to regions proximal to the N-terminus of the annotation gene may correct these errors. In *Shewanella oneidensis*, 30 genes appeared to have incorrect 5' boundaries based on peptides mapping upstream of the annotated start site as well as alignment with proteins in related species [26]. Peptides which are mapped near an annotated gene, but are in a different frame, may indicate the rare event of programmed frame shift [64], which is nearly impossible to predict by other automated methods. Baudet *et al* [82] derived protein N-termini using a labeling reagent, TMPP, to correct the translation start sites of 60 genes in *Deinococcus radiodurans.*

With the dramatic increase of sequenced genomes of related prokaryotic organisms, proteogenomics is now being performed on multiple sequences in tandem. Gallien *et al.* [86] combined comparative genomics and N-terminal protein labeling to correct 19% of translation start sites in *M. smegmatis* and 601 start sites in 16 other *Mycobacterium* species. As an extension to previous work in *Shewanella oneidensis* [26], Gupta *et al.* [87] sampled the proteomes of three *Shewanella* species to simultaneous annotate their genomes. Due to the high level of sequence similarity between the species, 2,590 orthologous ORFs were defined as 'shared genes'. By allowing peptides identified on an orthologous protein to contribute evidence for expression of a protein, Gupta *et al.* are able to rescue over 140 proteins which would have been excluded from a proteomic experiment using the 'two peptide per protein' inference rule. While using comparative proteogenomics represents a new frontier for annotating genomes, methods for determining statistical significance of these inferences have yet to be developed.

In addition to gene annotation, a study in the bacteria *Shewanella oneidensis* discovered over 10,000 sites of chemical modification [26]. The diversity of modifications identified is beyond what can be specified by popular database search tools, underscoring one of the main challenges to proteomics and proteogenomics. Gupta *et al.* also discovered non-chemical protein modifications which reveal the dynamic nature of the proteome. By considering the positions of the most N-terminal peptides observed with relation to the predicted translation start site (Figure 2.5A), Gupta *et al.* [26] observed possible instances of signal peptides and N-terminal methionine cleavages. The study was able to distinguish potential signal peptides from post-source decay by identifying non-tryptic peptides contained in tryptic peptides. The peptide compositions also allowed them to determine motifs for signal peptide cleavage sites that closely agree with motifs used by computational predictors (Figure 2.5B). A comparative analysis reveals a functional role for N-terminal methionine excision [90]. Jaffe *et al.* [64] showed post-processing of a gene by identifying two halves of the resulting protein appearing separately in the same mass spectrometry run.

2.5 New directions for proteogenomics

The discussion above assumes that the peptide encoded by the spectrum can be found in the genomic database. This may not always be the case. However, the peptide may be inferred by comparing the spectrum against a related genomic template. We refer to this as comparative proteogenomics. An exciting, if somewhat controversial, recent example is the sequencing of T. rex and mastodon peptides [91, 92, 93, 94].

MS-Blast [95] is often cited as an early tool for comparative proteogenomics. It relies on a *de novo* analysis to establish tags. A collection of tags is then searched using Blast to identify homologous sequences. Likewise the search for mutated and modified peptides also implies an imperfect genomic template. Tools such as MSAlignment [78], Mod^i [96], SPIDER [97], and TagRecon [98] perform a search of a homologous database with an unrestricted set of modifications or mutations.

When the genomic templates are very different (a different species), a new set of tools are required. Comparative shotgun protein sequencing [99] uses clustering, spectrum alignment, and de novo sequencing techniques to create sequence contigs of the target protein. *Champs* [100] identifies the most similar protein to the target protein in the database, and uses SPIDER to correct de novo sequenced peptides against the protein. *GenoMS* [101] resembles both de novo and database search techniques. It first identifies one or more *templates* from a database of homologous proteins or a related genome. Mutated or missing portions of the target proteins or proteins are then sequenced using model-based spectral alignment and de novo sequencing.

The gene annotation for an organism is not a once and done enterprise, but relies on a feedback loop involving the genomic and proteomic communities. Proteogenomics has developed beyond the proof-of-principle level, and is becoming an integral part of the annotation pipeline for model organisms. The realization of the method, in studies to date, has only been as a downstream analysis tool, for improvement of a first pass annotation. However, the high-throughput nature and the ability to directly ascertain the elements of the genome which are translated, highly recommend proteogenomics as a method to be used on the front-line of gene annotation.

The need for proteogenomics is highlighted by the exponential rate of growth of genomic databases, not only across species, but of individuals within species. For eukaryotes, the notion of gene is evolving to diverse trans-splicing and rearrangement induced splicing events. For prokaryotes, a vast majority of the genomes will never be sampled due to the difficulty in culturing. Instead, metagenomic studies sample genomic sequence from a community of genetically diverse organisms, which makes even species identification difficult. The development of sequencing technologies is allowing for the sequencing of genomes and meta-genomes at an unprecedented rate [102, 103]. At the same time, advances in instrumentation, MS2 identification algorithms, specialized database construction, and comparative tools suggest that the future is bright for proteogenomics.

2.6 Acknowledgements

Chapter 2, in full, was published as “Proteogenomics to discover the full content of genomes: a computational perspective”. NE Castellana and V Bafna. *Journal of Proteomics*, vol. 73, no. 11, pp. 2124-2135, 2010. The dissertation author was the primary author of this paper. The dissertation author reviewed the relevant papers in the field, and wrote the paper.

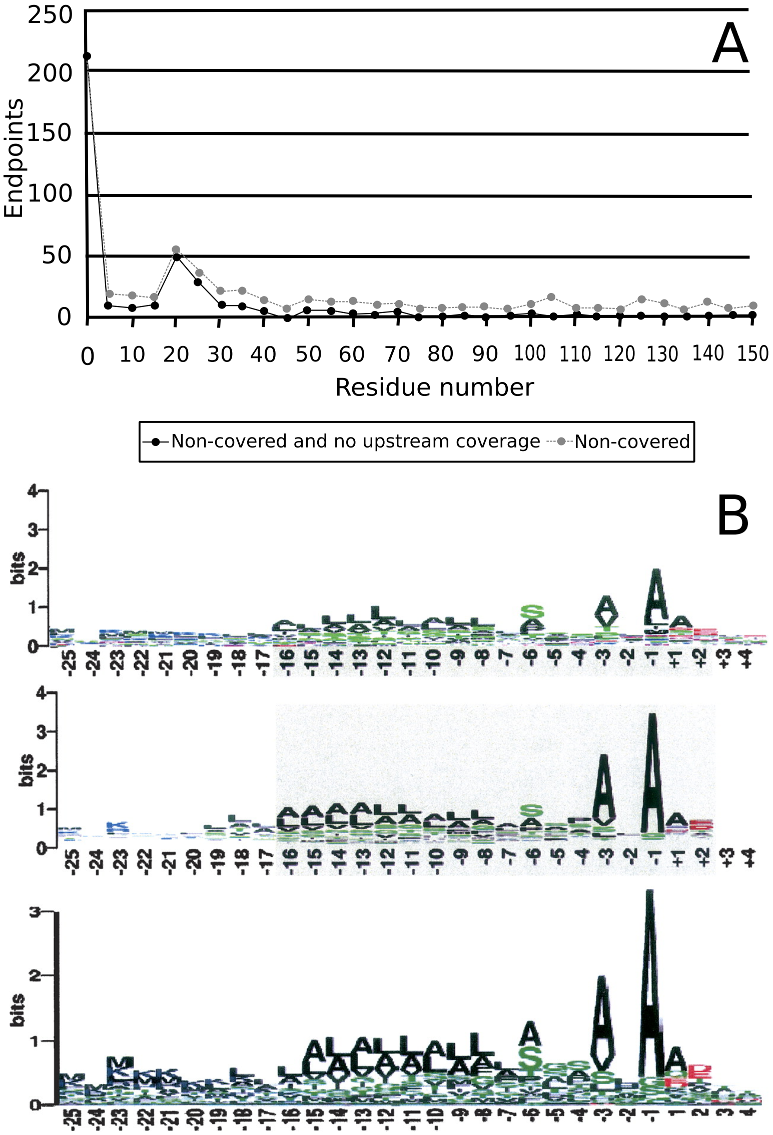


Figure 2.5: Proteomic annotation results courtesy of Gupta *et al.* [26]. A: A histogram of the position of non-tryptic N-terminal peptides reveals two protein processing events; cleavage of N-terminal methionine and cleavage of signal peptides. B: The signal peptide motif recovered by MS/MS analysis compared to the same motif determined by two computational predictors [88, 89].

Chapter 3

Discovery and revision of Arabidopsis thaliana genes

3.1 Introduction

A fundamental goal of genome projects is to generate a protein-coding catalog. Much of modern biological research depends on a complete and accurate proteome. Extensive proteomic catalogs have been developed through the integration of gene prediction algorithms, cDNA sequences, and comparative genomics [104, 105]. As emerging research is incorporated into annotation pipelines and manual curation efforts, gene models continue to improve. High throughput gene annotation pipelines use a variety of information sources, and benefit most significantly when new data contains information that is orthogonal to the kinds currently available [8].

Recent advances in chemistry and algorithms for peptide mass spectrometry have enabled the production of large proteomics datasets with broad coverage of the proteome [47, 46, 16]. Proteo-genomics (using proteomic information to annotate the genome) complements nucleotide-based annotation in that it unambiguously determines reading frame, translation start and stop sites, splice boundaries, and the validity of short ORFs. By combining DNA-based annotation with proteogenomics, an accurate and more complete protein-coding catalog can be obtained [16, 80, 26, 62, 65]. With its clear potential for improving genome annotation, proteogenomics could be integrated with genome projects.

A recent publication by Baerenfaller et al. [47] demonstrated the ability of extensive resampling to provide good coverage of the Arabidopsis proteome. From 1,354 LC runs the authors identified 86,456 distinct peptides covering 13,029 proteins. In addition to providing an organ specific proteome catalog, they demonstrated the ability of proteomics to refine plant genome annotation by presenting evidence for 57 new gene models, including 7 from intergenic regions not suspected to contain genes.

We reported a proteogenomic study of humans that described an exon splice graph that enabled efficient searches of potential coding sequences, including peptides that span splice junctions [16]. We reasoned that we could extend the observations of Baerenfaller et al. deeper into the unmapped proteome by building an exon splice graph of Arabidopsis and obtaining a novel set of peptides. We used two strategies to obtain novel peptides. First we used a nested 3D LC strategy to obtain much greater peptide separation permitting a deeper sampling of the proteome. This is reflected by our yield of 144,079 distinct peptides from only 45 LC runs, with a false-discovery rate $< 1\%$. Second we used TiO2 to enrich for phosphopeptides. Phosphorylated proteins are less abundant and are mostly missing from profiles of whole proteomes. Considering only cases in which we observed 2 or more previously non-annotated peptides mapping within 1 kb of each other, we discovered 1,473 new or revised genes; a model was generated for each using the gene finder AUGUSTUS [12]. Two hundred eighty genes were previously unrecognized, 498 were previously annotated as pseudogenes, and 695 were revisions of known genes that were annotated in the wrong reading frame, with missing exons, or with incomplete exons. Extrapolating from our sample we estimate that 13% of Arabidopsis protein-coding genes were either not yet identified or they contained significant errors in their exon definitions. We have remedied $\approx 40\%$ of these deficiencies.

3.2 Methods for gene finding in Arabidopsis

3.2.1 Sample Chemistry

In total, 21,170,989 MS/MS spectra were collected from 45 LC-MS/MS experiments. For Arabidopsis organ samples (leaf, root, flower, and silique), 2 g of fresh organs were cut from wild type Arabidopsis (Col-0) using a sharp razor blade and transferred into a 50 mL conical tube filled with liquid nitrogen immediately. Frozen organs were ground in a ceramic mortar and pestle with liquid nitrogen for 15 minutes to fine powders, and then transferred to a 50 mL conical tube. 50 mL cold (20°C) methanol containing 0.2 mM Na3VO4 was added to the conical tube. Samples were incubated at 20°C for 15 minutes and then spun down in a refrigerated centrifuge at 4,000g for 5 minutes. Supernatant was discarded. Two more methanol washes were performed, and followed by three acetone washes using the same procedure. After final acetone wash, sample pellets were dried in an Eppendorf Vacufuge Concentrator at 4°C. Proteins were extracted by adding 1 mL of 0.2% RapiGest (Waters) with 0.2 mM Na3VO4 to the dry pellet and incubated on ice for 15 minutes. Samples were spun down at 16,000 g in a refrigerated centrifuge for 15 minutes. Pellets were discarded and the supernatants were ready for protein digestion. For MM2d cells, cell pellets (100 L pellet volume) were washed by 1 mL Hepes saline buffer (10mM Hepes, 150 mM NaCl) three times. A total of 250 L of protein extraction buffer (2% RapiGest from Waters plus 0.2 mM Na3VO4) was added to the cell pellet. Samples were sonicated in a Branson Sonifier 450 sonicator equipped with a high intensity cup horn (Branson Part No. 101-147-046) at 40% output power for 2 minutes with circulating cooling water. Cell lysates were centrifuged at 16,100g, 4°C, for 15 minutes. Pellets were discarded and the supernatants were diluted 10 times in 50 mM Hepes buffer (pH 7.2). Cysteines were reduced and alkylated using 1 mM Tris(2- carboxyethyl)phosphine (Fisher, AC36383) at 95°C for 5 minutes then 2.5 mM iodoacetamide (Fisher, AC12227) at 37°C in dark for 15 minutes. Proteins were digested with trypsin (Roche, 03 708 969 001) overnight then 1% TFA (pH 1.4) was added to precipitate RapiGest.

Samples were incubated at 4 C overnight and then centrifuged at 16,100 g for 15 minutes. Supernatant was collected and centrifuged through a 0.22 uM filter. An Agilent 1100 HPLC system (Agilent Technologies) delivered a flow rate of 300 nL min to a 3-phase capillary chromatography column through a splitter. Using a custom pressure cell, 5 m Zorbax SB-C18 (Agilent) was packed into fused silica capillary tubing (200mID, 360mOD, 20 cm long) to form the first dimension reverse phase column (RP1). A 5 cm long strong cation exchange (SCX) column packed with 5 m PolySulfoethyl (PolyLC) was connected to RP1 using a zero dead volume 1m filter (Upchurch, M548) attached to the exit of the RP1 column. A fused silica capillary (100 m ID, 360 m OD, 20 cm long) packed with 5 m Zorbax SB-C18 (Agilent) was connected to SCX as the analytical column (RP2). The electrospray tip of the fused silica tubing was pulled to a sharp tip with the inner diameter smaller than 1 m using a laser puller (Sutter P-2000). The peptide mixtures were loaded onto the RP1 column using the custom pressure cell. Columns were not re-used. Peptides were first eluted from the RP1 column to the SCX column using a 0 to 80% acetonitrile gradient for 150 minutes. The peptides were fractionated by the SCX column using a series of salt gradients (from 10 mM to 1 M ammonium acetate for 20 minutes), followed by high resolution reverse phase separation using an acetonitrile gradient of 0 to 80% for 120 minutes. Spectra were acquired on LTQ linear ion trap tandem mass spectrometers (Thermo Electron) employing automated, data dependent acquisition. The mass spectrometer was operated in positive ion mode with a source temperature of 150°C. As a final fractionation step, gas phase separation in the ion trap was employed to separate the peptides into 3 mass classes prior to scanning; the full MS scan range was divided into 3 smaller scan ranges (300800, 8001,100, and 1,1002,000 Da) to improve dynamic range. Each MS scan was followed by 4 MS/MS scans of the most intense ions from the parent MS scan. A dynamic exclusion of 1 minute was used to improve the duty cycle. Final totals for spectrum count were: 6,336,450 spectra from roots, 1,415,293 spectra from M. incognita infected roots, 2,660,544 from leaves, 1,284,713 from flowers, 1,206,222 from siliques, and 8,267,767 from phospho-peptide enriched MM2D cell lysates. The data associated with this manuscript may be downloaded from Tranche (http://tranche.proteomecommons. org) using the following hash: eTyqbeZEgF7KOZNqcE0OAbFGAmrIzV1xKx4OCC0CJN9A1MwZmuP2drhEsT7XohMx8FM8wtckHv7mqSnWHLhVuGmrsYAAAAAAASfeg. The Tranche hash can also be used to verify that files have not changed since publication.

3.2.2 Database Construction and Use

Proper database construction is crucial for novel peptide recovery. For gene model confirmation, we used the TAIR7 release of the Arabidopsis proteome (www.arabidopsis. org). For proteomic discovery, we constructed two greatly expanded databases. The first database was the six frame translation of the Arabidopsis genome, containing 210 M amino acids. The second database was a spliced-exon graph containing ab initio gene predictions from the AUGUSTUS software [16, 12, 106] AUGUSTUS reported multiple transcripts per locus with sampling parameter 100. Additionally, we edited the exon length distribution to make short exons (100 base pairs) 3 more likely. All resulting exon and intron predictions were incorporated into the graph where each node is a putative exon and each directed edge indicates a putative splice junction. The resulting graph contained 16 M amino acids. For the MS/MS searches, all three databases were combined with decoy sequences formed by shuffling each target sequence. To ensure minimal overlap between target and decoy sequences, any 8-mer appearing in the decoy sequences which also appears in the target database was re-shuffled.

3.2.3 Mass Spectrometry, Peptide Identification and Location

All spectra above were converted from the vendor formatted RAW files to mzXML using the ReAdW software in centroid mode (Nov 1, 2006 version). Spectra were searched against the three databases with the Inspect software, release 2007.09.05. All datasets, excepting the phospho-peptide enriched samples, were searched without allowing any post-translational modifications (PTMs). Parameters for this search were: PM tolerance 3.0 Da, 0.5 Da fragment ion tolerance, 25 tags/spectrum, 57 Da fixed modification on cysteine. For phospho-enriched samples, we allowed a variable modification of 80 on STY, max of 2 PTMs/peptide and searched with InsPecTs phosphopeptide specific scoring function [107]. All results are filtered to 1% spectrum-level false discovery rate using the decoy database strategy [58]. In this strategy, a scrambled database of the same size is concurrently searched with the target sequences against the spectra. A score cutoff is chosen such that no more than 1% of the spectra are annotated with a peptide from decoy sequences. To count proteins validated by our TAIR7 database search, we map peptides back to their protein(s). We report proteins with two or more peptides, and at least 1 uniquely mapped peptide. For proteins groups which have exactly identical coding sequences we report the group of proteins, as they share all peptides and do not have any uniquely mapped peptides. As we require multiple peptides per protein identification, our 1% spectrum level FDR translated to an empirical 0.6% protein-level FDR. The source code for Inspect is available at our laboratory web site, http://peptide.ucsd.edu

3.2.4 Clustering and Homology Search

Novel peptides (all of which have a genomic location) were clustered. Peptides within 1,000 nucleotides were linked; clusters were aggregated by single linkage. We find that the vast majority of peptides within current genes fit this clustering (98%). Any fixed width cluster has the potential to misgroup peptides from multiple genes into a single group. This is overcome by the gene finding algorithm which creates the best gene model given the evidence, including splitting clusters. Clusters were classified as intragenic or intergenic depending on whether they overlapped a TAIR7 protein coding gene model. We extracted the DNA sequence of each cluster with 500 nucleotides abutting the first and last amino acid of the predicted peptides and searched versus the NCBI nonredundant protein database (NCBI nr) using blast with default parameters.

3.2.5 Frame Correction

We found evidence of many novel peptides out of frame with the current gene models. From these we picked a subset to highlight. From the list of all novel peptides that overlap a known gene locus, we generated a list of peptides that overlap the coding region of the locus but in a different frame. Our reported results require at least two out-of-frame novel peptides. To increase our confidence in the assertion that the gene is (at least partially) mispredicted, we also tabulated several features for these novel peptides. As splicing sometimes results in only a portion of the peptide being out of frame with the reference annotation, we filtered out peptides that had 3 aa out of frame. In these cases we do not doubt the accuracy of the MS/MS annotation. However, with only one or two amino acid(s), there are likely several close genomic regions with an appropriate nucleotide sequence. Additionally, we determined whether the novel peptides conflicted with observed MS/MS peptides that support the current gene model and frame. On a few occasions there was peptide and homology support for both the annotated frame and a new frame, possibly suggesting alternative splicing. (There are instances within the current annotation of the same DNA sequence being translated in multiple frames.) Seventy proteins had novel peptides that met these three requirements: multiple peptides out of frame, sequence out of frame is at least 3 aa, and no conflicting TAIR peptides. There were also instances of novel peptides present in the 5' and 3' untranslated regions of genes. In some instances these are likely to merely be expansion of the current sequence. However, some of these also appear to be frame mis-predictions

3.2.6 False Discovery Rate Calculations

Local false discovery rate (lFDR) The most commonly reported statistic for false-discovery rate is the cumulative false-discovery rate, cFDR, or the fraction of false positive spectra with a score greater than t. This number is meant as an estimate of error for a data set and is often misinterpreted as a confidence in a single spectrum annotation. For example, consider a data set with 1,000 spectra annotated at a 1% cFDR at score t_0. At this cutoff, 10 of the 1,000 spectra are estimated to be false-positives. As the score cutoff is relaxed and more spectra are accumulated, we set the next cutoff t_1 for a 5% cFDR and note 1200 spectra are annotated. 60 of the 1200 are estimated false positives. cFDR assignments to these new spectra would be between 1% and 5%. However, this is misleading. The entire data set has a 5% false discovery rate, but for the 200 newly included spectra, the false discovery rate is much higher. Of the 60 total false-positive spectra, 50 came from these 200 new annotations, or 25% false discovery. Thus a cumulative FDR calculation should not be used to estimate confidence in single annotations. Unfortunately, this point has not been previously addressed in proteomics studies. When considering that we annotate over 2.7 million spectra at a cFDR of 1%, a more lax 5% cFDR could have included a significant number of spectra that in reality have an unacceptable false-discovery rate. To more accurately measure the quality of our assignments, we defined a local false discovery rate [59]. For score$t$, and bin-size δ, define *local*-FDR (lFDR$_\delta$(t)), as the fraction of incorrect identifications with score in [t,t+δ).

$$lFDR_\delta(t) = f_0(t)/f_1(t)$$

where $f_0(t)$ is the number of false annotations with score in [t,t+δ) and $f_1(t)$ is the number of true annotations with score in [t,t+δ). Although local FDR is a continuous function, we empirically measure it over a discretized range. Unlike microarray experiments where the number of false data points must be estimated, by using the decoy database search strategy, we can directly count this value; f_0 is simply the distribution of matches to the decoy database and f_1f is the distribution of matches to the true database. We compute a local FDR for each spectrum using 0.1. Our dataset is large, therefore, a significant number of spectrum-peptide matches fall in each bin and we can achieve a more accurate local FDR. For higher score regions with fewer spectrum-peptide matches, bins were expanded to include at least 1,000 annotations. As spectra of different charge states have distinct score distributions (data not shown), the FDR should be separately calculated. Inspect identifies spectra of charge 3, and we compute lFDR separately for charge

3 spectra. A change in FDR for peptides of different lengths has also been reported. As Inspect explicitly takes peptide-length into account while scoring, this bias is not observed in our identifications (data not shown). The minimum peptide length for Inspect is 7 aa; 0.8% of all reported spectral identifications are of length 7.

Spectral FDR versus peptide FDR (pFDR) The redundancy introduced by repeated observation (multiple spectra) of peptides changes the false discovery rate for peptides. If 100 spectra identified the same peptide, the peptide identification is incorrect only if all spectral identifications are incorrect. At the same time, spectra identifying the same peptide cannot be treated as independent. A systematic error might lead to similar spectra to all be mis-annotated. Therefore, we conservatively assign the FDR for a peptide to be the minimum local FDR of all spectra identifying that peptide.

Event-level FDR (eFDR) The identification of distinct peptides can be reasonably assumed to be independent. Even peptides that overlap in sequence have completely different spectra, as prefix/ suffix masses are all changed by the distinct terminal residues. In identifying an event, (e.g. a novel exon), we estimate the FDR of the event as the product of the local FDR of the peptides supporting that event, and use an eFDR cut-off of 5%.

Estimation of level of alternative splicing To estimate the extent of alternative splicing, we start with a few simplifying assumptions. Each gene has a common splice-isoform that has the highest expression level. Alternative splice isoforms are characterized by relative expression e_{alt}, $(0 \leq e_{alt} \leq 1)$, denoting the expression of the alternative isoform relative to the common one ($e_{common} = 1$). Each alternative splicing event can be characterized by a branching intron that splices together exons A and B which are not spliced together in the common form. We would like to estimate N, the number of branching introns as a proxy for the number of alternative splice-forms. We sample an intron connecting exons A,B, if we identify a uniquely located peptide that spans the splice junctions of A,B. In the proteogenomic search of TAIR reference sequences, we sampled 12,616 of the 115,040 introns ($f = 0.11$). Assuming that each intron is sampled independently with probability f,

Pr[a branching intron (with relative expression e_{alt}) is sampled] $\approx e_{alt}f$.

Pr[an intron (A,C) and a branching intron (A,B) are sample] $\approx e_{common}e_{alt}f^2$ $= e_{alt}f^2$.

Therefore,

E[common and branching introns sampled] $\approx e_{alt}f^2N$.

A sampling of the graph revealed a total of 47 instances of multiple branch-points. According to recent studies, it is not unreasonable to assume one multiple branch-point event occurs between each pair of isoforms [108]. At $e_{alt} = 0.5$, the model tells us that in order for our observations to reside within one standard deviation of the mean, $6,718 \leq N \leq 8,983$. However, the TAIR7 protein database contains only 3,799 alternatively spliced genes, where $N = 3,141$. Therefore, in order for our branch-point observation to be consistent with the observed N, minor isoforms must be expressed at no less than 88% of the dominant isoform. This indicates that alternatively spliced protein isoforms are only detectible when they have roughly equal expression levels, or that the predicted number of alternatively spliced genes in arabidopsis is considerably lower than what is actually present.

3.3 Results in Arabidopsis

3.3.1 Proteome Coverage

To achieve broad coverage of the proteome, we acquired 21 million mass spectra from protein extracts of 4 Arabidopsis organs (leaf, root, flower, and silique) and a cell culture (MM2d). In addition, phosphopeptides were enriched from MM2d proteins. Inspect was used to search spectra against 3 reference databases: TAIR7; a 6 frame translation of the genome; and an exon splice-graph that compactly encodes putative splicing events (6) (See Figure 3.1 for an overview of the method). The data were filtered to a 1% cumulative false-discovery rate (FDR) at the spectrum level. We required at least two peptides per protein for identification, so our 1% spectrum-level FDR provided an empirical, protein-level FDR of 0.6%.

A total of 144,079 distinct peptides were mapped to at least 1 of our 3 Arabidopsis protein databases. Most (126,055 peptides) resided in TAIR7 gene models (12,769 proteins confirmed). We mapped 18,024 peptides not present in the TAIR7 annotation including 4,018 peptides (22%) that were derived from mRNA splicing. Of these, 16,348 peptides mapped to single loci (i.e., uniquely-located) in the genome, whereas the rest were shared between 2 or more related proteins. The 6-frame translation and the spliced-exon databases contributed equally to the discovery of novel peptides and their contributions had little overlap; only 5% were found in both databases. This indicates that both types of database should be used for proteogenomic studies because they provide complementary novelty. Every reported peptide can be uploaded as a track in TAIR8. These files are available at http://peptide.ucsd.edu. The AUGUSTUS model building was restricted to nuclear genes and they encompass 2,873 novel peptides. These models can be accessed from http://peptide.ucsd.edu.

3.3.2 NovelGenes

Using the protein identification standard of 2 peptides per protein, we focused on 1,765 novel peptide clusters containing 5,426 novel peptides, 4,575 of which are uniquely located. An additional 6,361 novel peptides were observed outside of clusters but with a unique genomic location and a local FDR < 0.05. These were not analyzed in detail. We classified novel peptide clusters according to their position relative to annotated protein coding models. We defined intragenic clusters as those falling within the boundaries of a known protein coding gene and intergenic clusters as those falling in the intergenic space (i.e., these indicate novel genes). Some of the novel clusters overlapped loci that had been annotated as non-coding pseudogenes (31% of the peptides or 1,420 peptides derived from 561 clusters) or genes that had not been recognized at all by gene finding programs or annotators (20% of the peptides or 905 peptides derived from 331 clusters).

With our novel intergenic peptides, we defined 778 new genes consisting of 930 transcripts using the gene finder AUGUSTUS. Evidence from peptides plus EST alignments, and genomic conservation with rice, poplar, and Medicago, were given as hints to AUGUSTUS, which derives gene models that are in agreement with the hints and that have high likelihood in an ab initio probabilistic gene structure model. Resulting gene models include alternative splice variants, if suggested by the evidence. Of the 778 novel genes, 55 have EST and homology support, in addition to peptides; 455 genes have support by the peptides and ESTs; and 70 genes are supported by the peptides and homology only. The remaining 198 genes have no other support than the peptides. As an independent validation of our discoveries, 52 of the 778 loci have now been incorporated in the newest Arabidopsis genome release (TAIR8).

To discover homology with the novel genes, we excised the surrounding nucleotide sequence and searched against the nonredundant database of proteins (National Center for Biotechnology Information nr version 03/26/08). For 539 of the loci, the underlying sequence revealed a close homolog (e value $<$ 1E-10), providing additional validation, and functional assignments for the new genes. Although many of the novel genes we discover are homologous to genes of unknown function, we highlight a novel gene involved in photosynthesis. Our predicted protein, supported by 13 novel and uniquely located peptides, aligns with proteins targeted to the chloroplast thylakoid lumen (e value = 1E-75). It also contains the PsbP pfam domain characteristic of photosystem II (See Figure 3.2). A second novel locus containing 4 uniquely located peptides on chromosome 4 shows strong similarity (e value = 1E-85) with a heat-shock protein (AT4G12770).

We also note several interesting structural features of the intergenic clusters. First, a significant fraction (64%) of intergenic clusters overlap annotated pseudogenes or transposons. An example of a translated pseudogene is at locus AT2G15040, ATRLP18: Receptor like protein 18, which has high homology to disease resistance proteins in both Arabidopsis and other plants. We identify 5 peptides, 3 of which are uniquely located at this locus, confirming translation. It is presumed that pseudogenes do not produce proteins, but transposons (which like pseudogenes are not typically included in the proteome) can contain active protein-coding genes. We find evidence in transposons of translated proteins that are unrelated to transposon activity. For example, we identified 3 peptides within the locus AT4G07947 (See Figure 3.3A). Although annotated as a pseudogene in TAIR7 it has been reclassified as a transposable element gene in TAIR8. The genomic region containing these peptides has high similarity to the ubiquitin-like protease (Ulp1) family in Arabidopsis (See Figure 3.3B), suggesting this may be a gene traveling as cargo with the transposable element [109].

Since the release of TAIR7, there have been several community annotation efforts, including publication of short ORFs [110]. We compared the novel peptides for overlap with this set. Hanada et al. [110] reported that 7,442 non-annotated small ORFs in Arabidopsis are transcribed. Our peptides confirm the translation of 155 of these predicted ORFs. An additional 85 ORFs overlap at least 1 of our novel peptides, but the peptides indicate that the frame of the ORF may be incorrect.

3.3.3 Refined Gene Models

In addition to the novel genes, we discovered peptide clusters overlapping annotated gene models, suggesting refinement of the existing annotation, e.g., a new exon, exon boundary change, exon skipping, or modified translation boundaries. The refinement events can be classified according to their type, location, and the transcript being modified. A majority (521) of the events are novel exons, of which 314 are located within introns and 207 are in untranslated regions (UTR) of TAIR7 gene models. Of the 314 instances of novel coding sequence predicted between 2 exons of a gene, 26 are observed in the same frame as both adjacent exons and may indicate a single exon, a portion of which may be spliced out in some isoforms. Exon boundary changes were also prevalent, with typical instances including 5' extension of the first exon and alternative donor/acceptor splice sites. We find evidence for 180 instances of exon extension, and 191 instances of exon shortening. In 5 transcripts, peptide evidence supports an exon skipping event. Some intragenic loci indicate gene extension beyond the borders of the annotated gene model; 323 of these gene extension events were discovered. Using AUGUSTUS to refine existing models with the new peptide evidence, we predict 964 new or altered transcripts in 695 genes.

It is difficult to determine whether a new transcript predicted by AUGUSTUS is a refinement of a gene model or an additional isoform of an alternatively spliced gene. To better distinguish between these two cases, we compared each new transcript and the TAIR7 transcripts to all available cDNA evidence. For 122 genes, EST evidence, in addition to the peptide evidence, was found in support of the new transcript; no ESTs were found in support of the TAIR transcript. For an additional 130 genes, EST evidence was found to support both the new transcript and the TAIR7 gene models, suggesting that the peptides are produced by a newly discovered isoform.

To provide additional support to our gene refinements, we excised the predicted amino acid sequence surrounding a novel cluster and searched for homology to the nonredundant database of proteins (National Center for Biotechnology Information nr version 03/26/08). For 348 loci, we found a close homolog (e value $<$ 1E-10). Several genes that have been extensively studied are included among the refined gene models. For example, we found an additional 200-aa exon in the 5' UTR of MAPK phosphatase (AT3G55270). Also, we identified 8 peptides corresponding to 4 missing or mispredicted exons at locus AT1G79920 (heat shock protein 70). The new sequence completes the canonical HSP70 pFam domain. A final example is the gene PMI1 (AT1G42550), which, when mutated, results in impaired plastid movement and localization [111]. We found 6 peptides upstream of the annotated start codon, providing at least 130 aa of additional sequence.

We identified 70 cases in which the annotated reading frame is different from the observed peptides. Assignment of reading frame is particularly difficult for nucleotide-based genome annotation (e.g., cDNA). However, proteomic evidence unambiguously defines the frame of translation. The 70 frame corrections are supported by multiple peptides and extensive homology to other proteins. We will use two proteins to illustrate: first, a whole gene frame correction; and second, a partial gene correction. Locus AT3G22240 is a 51-aa protein with no discernible homologs. Four of our peptides indicate translation in different frame than has been annotated. Translation in the new reading frame yields a protein with high sequence identity to PCC1, pathogen and clock controlled protein. The second example is AT1G63500, a protein kinase, which has 4 novel peptides in the annotated 5' UTR. These peptides point to a large expansion of the gene and a misprediction of the current first exon (Figure 3.4).

In addition to the peptides that are described above, 3,534 uniquely located singleton peptides with high confidence (lFDR < 0.05) overlap genes and indicate refinement events. These peptides likely indicate corrections to gene models and are a starting point for further investigation.

Similarly, 2,827 singleton peptides (also uniquely located and with lFDR ¡0.05) are found in intergenic regions. Some of the peptides may be mis-annotations, however, subsequent work has indicated that many are correct: 665 peptides are contained in ORFs with strong sequence similarity to known proteins (BLAST e value $<$ 1E-10). Spectral counts are also an indication of strength of an annotation; 291 peptides have higher spectral counts than 50% of all peptides identified in this study. The intergenic peptides indicate novel coding regions that may have produced a single detectible peptide for several reasons including protein composition or protein length.

3.3.4 Validated Gene Models

In addition to discovering new protein-coding loci, we identified 126,055 distinct peptides (1.72 million amino acids) that confirmed annotated gene models for 12,769 proteins (40% of the TAIR7 genes). Our claims of coverage are conservative. We count only proteins covered by at least two peptides, one of which must uniquely map to the designated locus. A total of 11,801 peptides were lone supporters of proteins or shared peptides, and therefore were not counted toward the confirmed proteins. Of the sequenced peptides, 87% map to a unique genomic location, unambiguously identifying 10,692 proteins. In addition, we observed proteins from highly homologous gene groups that could not be attributed to a single locus (see Methods). The Arabidopsis genome has high rates of tandem and segmental duplication and many loci contain multiple gene predictions that differ only in the non-translated regions [112]. We observed peptides from 913 groups of indistinguishable proteins (2,077 proteins), bringing the total confirmed gene models to 12,769.

3.3.5 Splicing

It is difficult to estimate the true extent of alternative splicing, given that the alternative splice forms are often not as highly expressed, and might not be sampled. However, our deep proteogenomic sampling revealed a total of 47 genes in which multiple splice forms were observed. We estimate that with high probability, the number of genes with alternative splice forms is between 6,718 and 8,983. This is considerably higher than the number of alternatively spliced genes in TAIR7 (3,799) and the number recently predicted by cDNA and ESTs (4,707 at the transcript level) [108].

3.4 Discussion

In tandem mass spectrometry, a peptide (from an enzymatic digestion of a protein mixture) is fragmented, usually through collisions. While the physics of the fragmentation is incompletely understood, the fragmentation pattern is consistent, and the collection of fragments (the spectrum) can be used to fingerprint the peptide. Recent advances in mass resolution and the availability of software tools to analyze spectra make mass spectrometry the tool of choice for proteomics. Nevertheless, technological limitations create many challenges for the approach.

First, the sampled peptides are biased toward the more abundant proteins in the cell. To comprehensively sample the proteome, a diversity of samples must be assayed. Second, incomplete fragmentation patterns and spectral noise smudge the fingerprint and introduce errors in peptide identification. Additionally, identification is typically based upon looking up a database of known peptides to pick the most likely candidate. If the true peptide is not in the database, it will not be identified. Finally, post-translational modifications change the mass and pattern of the fragments, making identification harder. Our study addresses each of these issues. Broad sampling of the proteome was achieved through assaying multiple plant organs and phosphopeptide enriched peptides. We address identification error rates through the introduction of a local false discovery rate. The genome is explicitly and thoroughly queried for potential protein coding sequences. Finally, we use a phosphopeptide spectra specific algorithm for sensitive and efficient annotation of phosphorylated peptides.

The database search tool we used, Inspect, contributed significantly to our ability to extensively annotate spectra. Inspect's Bayesian scoring function is more sensitive than that of SEQUEST, annotating more spectra at a given false-discovery rate. The exon splice-graph database allowed us to detect peptides that span splice boundaries. Our experimental techniques enabled a sampling of the phosphoproteome, which typically contains low abundance proteins. We used 3D LC, which provides much greater resolution and renders unnecessary the extensive resampling that is typical of LC ESI MS/MS experiments. To illustrate, we identified 67% more total peptides using only 3% as many LC runs compared with a study based on resampling (4). We used our novel peptides and an automated gene prediction pipeline to derive 1,473 new and revised gene models. The technical advances reported here dramatically reduce the time and cost required to obtain deep proteome coverage.

Historically, the proteomic and genomic communities have operated independently, with the genomic community in charge of annotation efforts. The predicted proteome is then passed over to the proteomics community for validation, and identification of post-translational events. We assert that much is to be gained by joining forces, and incorporating proteomic evidence upfront into the genomics pipelines. Proteogenomics provides an orthogonal data source to predict gene models, with levels of sensitivity that are complementary to cDNA sequencing. By investing in proteogenomics to complement more traditional cDNA and EST data at the onset of genome annotation, a more complete and accurate proteome can be achieved even in the early releases. Here, we provide proteomic evidence for 778 new genes and refine 695 current gene models, using the reference annotation from TAIR7. Recently, TAIR released the next revision of the genome/proteome, TAIR8. Only a small number of our novel peptides (3%) appear in the TAIR8 release indicating that the proteogenomic approach is complementary to computer-based annotation.

3.5 Acknowledgements

Chapter 3, in full, was published as “Discovery and revision of Arabidopsis genes by proteogenomics”. NE Castellana, SH Payne, Z Shen, M Stanke, V Bafna, and SP Briggs. *Proceedings of the National Academy of Science*, vol. 105, no. 52, pp. 21034-21038, 2008. The dissertation author was one of three primary authors of this paper. The dissertation author implemented computational methods, ran the computational analysis, and wrote the paper.

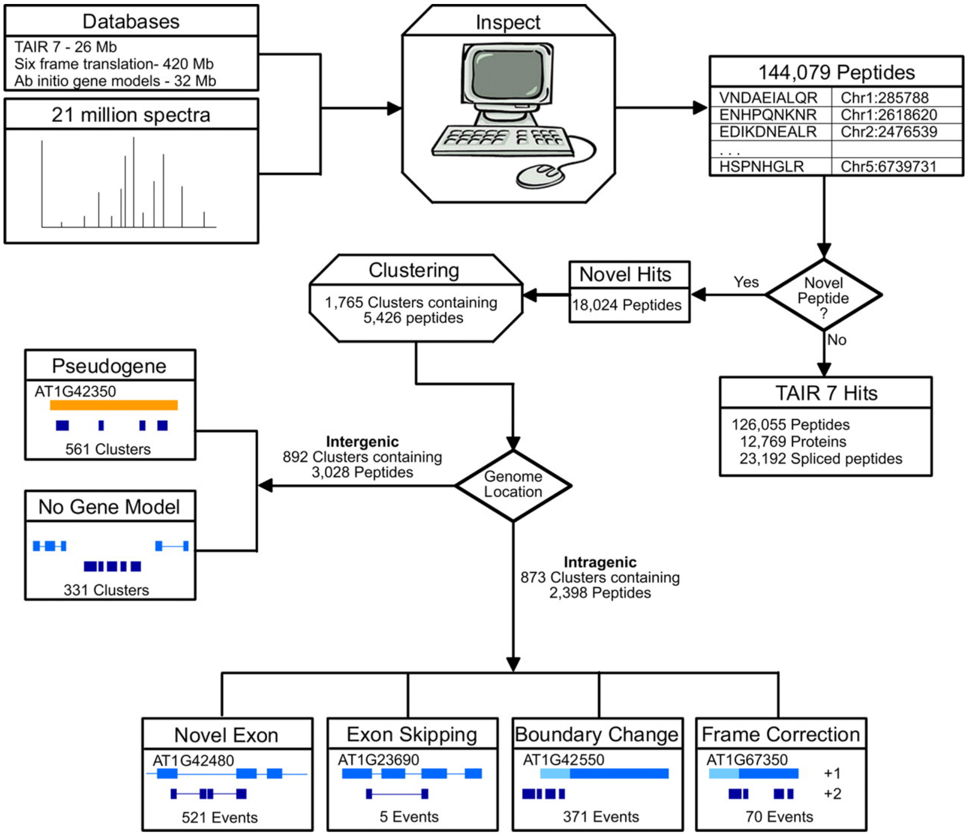
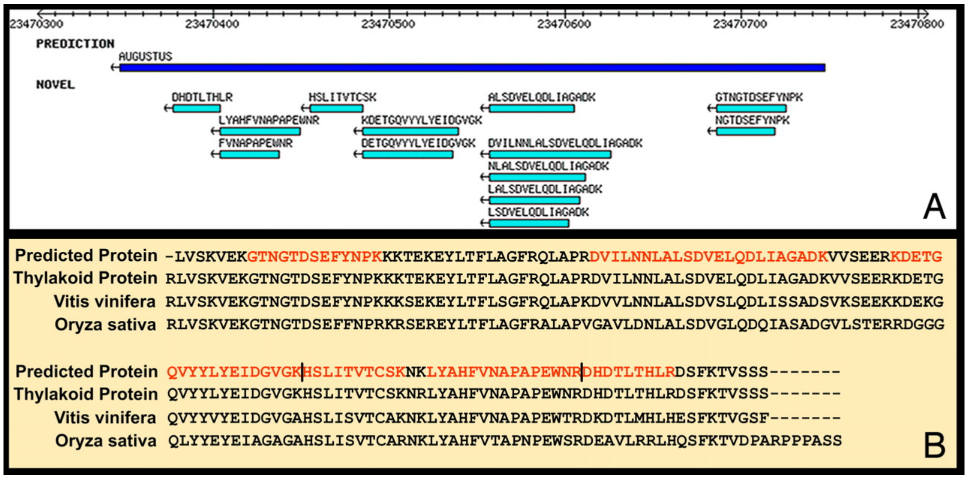
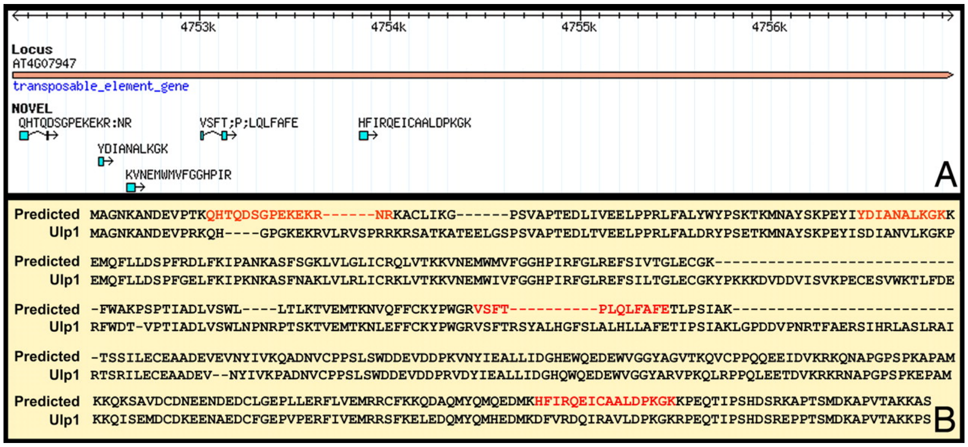


Figure 3.1: All mass spectra are compared with three databases using Inspect. Spectra are filtered to a 1% false discovery rate and grouped into peptides. Novel peptides are separated from those that appear in TAIR7 and clustered. It is important to note that only a subset of the novel peptides appear in a peptide cluster. Novel peptide clusters are then segregated based on genome location. Those that overlap a current gene model (intragenic) are further classified by how they refine the model. Peptides that do not overlap a gene model (intergenic) are classified by whether they overlap a pseudogene. The peptide clusters, along with evidence from cDNA and current gene annotations, are given to the gene predictor AUGUSTUS to produce new gene models. Not all peptides in the peptide clusters are included in the final AUGUSTUS models.



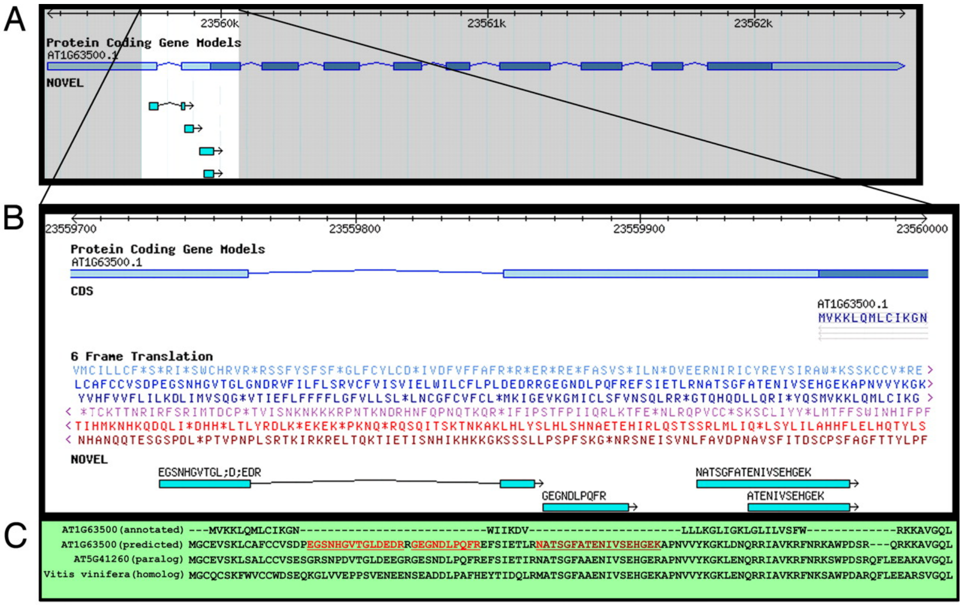
```
Predicted Protein -LVSKVEKGTNGTDSEFYNPKKKTEKEYLTFLAGFRQLAPRDVILNNLALSDVELQDLIAGADKVVSEERKDETG
Thylakoid Protein RLVSKVEKGTNGTDSEFYNPKKKTEKEYLTFLAGFRQLAPRDVILNNLALSDVELQDLIAGADKVVSEERKDETG
   Vitis vinifera RLVSKVEKGTNGTDSEFYNPKKKSEKEYLTFLSGFRQLAPKDVVLNNLALSDVSLQDLISSADSVKSEEKKDEKG
     Oryza sativa RLVSKVEKGTNGTDSEFFNPRKRSEREYLTFLAGFRALAPVGAVLDNLALSDVGLQDQIASADGVLSTERRDGGG

Predicted Protein QVYYLYEIDGVGKHSLITVTCSKNKLYAHFVNAPAPEWNRDHDTLTHLRDSFKTVSSS-------
Thylakoid Protein QVYYLYEIDGVGKHSLITVTCSKNRLYAHFVNAPAPEWNRDHDTLTHLRDSFKTVSSS-------
   Vitis vinifera QVYYVYEIDGVGAHSLISVTCAKNKLYAHFVNAPAPEWTRDKDTLMHLHESFKTVGSF-------
     Oryza sativa QLYYEYEIAGAGAHSLISVTCARNKLYAHFVTAPNPEWSRDEAVLRRLHQSFKTVDPARPPPASS
```

Figure 3.2: (A) A cluster of 13 uniquely located peptides that do not overlap a current gene model (Chr3). The prediction track shows the single exon gene model produced by AUGUSTUS. (B) The predicted sequence shows strong homology to a Thylakoid lumen family protein (sp||P82658||TL19_ARATH). It also shows strong similarity to proteins in both grapevine (emblCAO40861.1, a hypothetical gene) and rice (Os08g0504500, a cDNA derived gene).



```
Predicted MAGNKANDEVPTKQHTQDSGPEKEKR------NRKACLIKG------PSVAPTEDLIVEELPPRLFALYWYPSKTKMNAYSKPEYIYDIANALKGKK
     Ulp1 MAGNKANDEVPRKQH----GPGKEKRVLRVSPRRKRSATKATEELGSPSVAPTEDLTVEELPPRLFALDRYPSETKMNAYSKPEYISDIANVLKGKP

Predicted EMQFLLDSPFRDLFKIPANKASFSGKLVLGLICRQLVTKKVNEMWMVFGGHPIRFGLREFSIVTGLECGK--------------------------
     Ulp1 EMQFLLDSPFGELFKIPKNKASFNAKLVLRLICRKLVTKKVNEMWIVFGGHPIRFGLREFSILTGLECGKYPKKKDVDDVISVKPECESVWKTLFDE

Predicted -FWAKPSPTIADLVSWL----LTLKTVEMTKNVQFFCKYPWGRVSFT----------PLQLFAFETLPSIAK------------------------
     Ulp1 RFWDT-VPTIADLVSWLNPNRPTSKTVEMTKNLEFFCKYPWGRVSFTRSYALHGFSLALHLLAFETIPSIAKLGPDDVPNRTFAERSIHRLASLRAI

Predicted -TSSILECEAADEVEVNYIVKQADNVCPPSLSWDDEVDDPKVNYIEALLIDGHEWQEDEWVGGYAGVTKQVCPPQQEEIDVKRKQNAPGPSPKAPAM
     Ulp1 RTSRILECEAADEV--NYIVKPADNVCPPSLSWDDEVDDPRVDYIEALLIDGHQWQEDEWVGGYARVPKQLRPPQLEETDVKRKRNAPGPSPKEPAM

Predicted KKQKSAVDCDNEENDEDCLGEPLLERFLVEMRRCFKKQDAQMYQMQEDMKHFIRQEICAALDPKGKKPEQTIPSHDSRKAPTSMDKAPVTAKKAS
     Ulp1 KKQISEMDCDKEENAEDCFGEPVPERFIVEMRRSFKELEDQMYQMHEDMKDFVRDQIRAVLDPKGKRPEQTIPSHDSREPPTSMDKAPVTAKKPS
```

Figure 3.3: (A) 5 peptides, 4 unique, overlap locus AT4G07947, which is annotated as a transposable element gene. (B) Sequence alignment to an Arabidopsis Ulp1 (ubiquitin like protease) showing strong conservation (56% identity, e value 0.0). Observed peptides are highlighted.



```
 VMCILLCF*S*RI*SWCHRVR*RSSFYSFSF*GLFCYLCD*IVDFVFFAFR*R*ER*RE*FASVS*ILN*DVEERNIRICYREYSIRAW*KSSKCCV*RE >
 LCAFCCVSDPEGSNHGVTGLGNDRVFILFLSRVCFVISVIELWILCFLPLDEDRRGEGNDLPQFREFSIETLRNATSGFATENIVSEHGEKAPNVVYKGK>
  YVHFVVFLILKDLIMVSQG*VTIEFLFFFFLGFVLLSL*LNCGFCVFCL*MKIGEVKGMICLSFVNSQLRR*GTQHQDLLQRI*YQSMVKKLQMLCIKG >
< *TCKTTNRIRFSRIMTDCP*TVISNKNKKKRPNTKNDRHNFQPNQTKQR*IFIPSTFPIIQRLKTFE*NLRQPVCC*SKSCLIYY*LMTFFSWINHIFPF
<TIHMKNHKQDQLI*DHH*LTLYRDLK*EKEK*PKNQ*RQSQITSKTNKAKLHLYSLHLSHNAETEHIRLQSTSSRLMLIQ*LSYLILAHHFLELHQTYLS
< NHANQQTESGSPDL*PTVPNPLSRTKIRKRELTQKTIETISNHIKHKKGKSSSLLPSPFSKG*NRSNEISVNLFAVDPNAVSFITDSCPSFAGFTTYLPF
```

```
    AT1G63500(annotated) ---MVKKLQMLCIKGN---------------------------WIIKDV----------------------LLLKGLIGKLGLILVSFW-------------RKKAVGQL
    AT1G63500(predicted) MGCEVSKLCAFCCVSDPEGSNHGVTGLDEDRRGEGNDLPQFREFSIETLRNATSGFATENIVSEHGEKAPNVVYKGKLDNQRRIAVKRFNRKAWPDSR---QRKKAVGQL
      AT5G41260(paralog) MGCEVSKLSALCCVSESGRSNPDVTGLDEEGRGESNDLPQFREFSIETIRNATSGFAAENIVSEHGERAPNVVYKGKLENQRRIAVKRFNRKSWPDSRQFLEEAKAVGQL
Vitis vinifera(homolog) MGCQCSKFWVCCWDSEQKGLVVEPPSVENEENSEADDLPAFHEYTIDQLRMATSGFAAENIVSEHGEKAPNVVYKGKLENQRRIAVKRFNKSAWPDARQFLEEARSVGQL
```

Figure 3.4: TAIR locus AT1G63500 encodes a protein kinase. (A) Four novel peptides map within the 5' UTR and the first exon. (B) Zoom of the region shows that the current first exon (frame 3) is out of frame with the peptides (frame 2). (C) Sequence alignment with Arabidopsis and grapevine proteins supports translation in the frame supported by peptides (observed peptides highlighted in alignment).

Chapter 4

Novel genes discovered in *Zea mays* using an automated gene finding pipeline

4.1 Introduction

Automated annotation of genome sequences with the location of protein coding genes remains critically important. The wide availability of inexpensive next-generation sequencing technologies ensures that model organisms from all branches of the tree of life will continue to be sequenced at an ever increasing pace. However, the annotation of these genomes remains challenging.

Much recent focus on computational gene finding is on incorporating transcript evidence. As with genomic sequencing, availability of high-throughput technologies for transcript sequencing such as RNAseq [14] has dramatically changed the genome annotation landscape. While RNA-Seq provides valuable evidence for genome annotation [13, 69, 68], it does not provide a comprehensive solution either. Increasing evidence suggests that a discrepancy exists between protein isoforms that are transcribed versus translated [113]. We observed a large range in abundance (6 orders of magnitude) of transcripts suggesting that some genes are not observed even with deep sampling. Indeed in our own observation, we find evidence for genes in sampling proteins that are not visible at the transcript level. Moreover, the transcript evidence is confounded by pre-spliced messages, non-targeted expression noise, ncRNA, and lack of strand and frame information. All of these remain challenges for gene finding.

Tandem mass spectrometry is a key technology for assaying the expressed proteome. In typical bottom-up workflows, enzymatically digested peptides are isolated via chromatography and mass spectrometry, then fragmented. The collection of masses of peptide fragments (tandem mass spectrum) are used as a fingerprint for identification of expressed peptides. Historically, the genomics community has provided the annotations (aa sequences) and the proteomics community has focused on identifying peptides and proteins from this annotated list to assay for expression of proteins in specific contexts. However, rapidly improving mass spectrometry (MS) instrumentation and advances in sample preparation have enabled the field of proteogenomics, relying relying on direct interpretation from the genome [114]. In this context, the evidence of peptide expression is used to annotate the genome and reconstruction of gene structures in model organisms [48, 47, 115, 46], multiple organisms in parallel [87, 86, 116], and difficult to annotate genomes such as those with high GC content [117].

However, significant concerns remain in the development of this technology. The tandem mass spectra are noisy, and large-scale MS-based proteomic studies are confounded by false positives resulting from the intrinsic testing of millions of hypothesis. Further, many of the peptides cross splice-junctions, and would not be identified by searching against a 6-frame translation. At the same time identification of spliced-peptides is key to reconstructing gene structures. Finally, the annotated peptides must be reconciled into complete gene models, possibly with alternatively spliced isoforms.

Here we present a semi-automated proteogenomic method and apply it to the annotation of the maize (*Zea mays*) genome. Our method extends our previous proteogenomics efforts on human and arabidopsis [16, 48] using special spliced-graph databases to search spectra for spliced-peptides. In addition, we automate the refinement process to predict complete gene models, and automatically refine existing structures. We develop a framework for evaluating the quality of mass spectrometry-based discovery of gene refinement 'events' to control the false discovery rate. The maize genome is particularly challenging due to its large size (2 billion nucleotides), and abundance of mobile elements that re-insert themselves into the genome creating many repetitive regions. Instead of discarding peptides that appear in multiple locations, we use them to identify paralagous sets of genes, one or more of which is likely expressed.

We analyzed over 109 million tandem mass spectra generated from *Zea mays* seed tissues. Comparison of the spectra against our putative protein database containing nearly 2 billion amino acids required extensive computing power ($\sim$ 7.5 million CPU-hours). Our analysis revealed a revised genome annotation with updated gene models for 1,989 loci and the addition of 1,102 novel protein coding loci. This study represents possibly the largest proteogenomic effort undertaken on a single organism.

4.2 Results

The core of our method is the identification of peptides that are discordant with the annotated proteome. To identify the novel peptides, we generate a database of putative translated sequences directly from the genome. The database contains both the six frame translation and an exon-splice graph [16] that is informed by mapped mRNA sequences and *ab initio* predictions. The splice graph compactly represents many possible protein products for each genomic locus, collapsing shared sequence between multiple isoforms into single nodes in the graph and incorporating splice junctions by adding edges in the graph.

We identified 225,166 distinct peptide sequences by searching the MS/MS data against a database of both annotated and putative protein sequences. The peptide sequences either confirm annotated protein sequences in the 5a.59 maize proteome release or map to a genomic location that was not previously believed to be protein-coding. Based on the genomic-mapped peptide sequences, we codified eight distinct refinement 'events'; novel genes, novel exons, frame change, exon extension, gene extension, translated 'untranslated region' (UTR), novel splice junction, and overlapping antisense translation.

To control false-positive predictions, we applied a filter, the eventProb, which scores events based on the spectral evidence supporting them. Using the peptides from the filtered novel events, combined with over 2 million *Zea mays* ESTs, 875 million RNA-Seq reads, and the mappings of proteins from rice and sorghum, we generated putative gene structures using the *ab initio* tool, Augustus [69]. The gene predictions and proteogenomic evidence, together with the Maize Protein Atlas, can be found at http://maizeproteome.ucsd.edu/

4.2.1 Validation of Zea mays Genes

The maize proteome has two sets of annotated proteins. The filtered gene set (version 5b.60) contains 39,656 trusted gene predictions encoding 63,540 proteins. The working gene set (version 5a.59) contains an expanded, hypothetical set of genes and is a superset of the filtered set, containing a total of 136,770 proteins in 110,028 genes.

We identify 200,384 peptides sequences matching to proteins in the filtered set or working set. These peptides confirm the expression of 14,615 genes. The majority of these genes are from the filtered set (13,811 genes, 94%), suggesting that the 5b.60 annotation contains most maize genes expressed in seed tissue. The majority of maize genes produce two or more protein isoforms. The protein isoforms often share exons in the gene model and therefore peptides matching one isoform often match another isoform. However, we are able to determine that 10,604 specific protein isoforms from 10,507 genes are expressed, each with at least one uniquely mapping peptide. Again, the vast majority of the proteins identified are from the filtered protein set (9,874 proteins). We identify unique peptides in 730 working set proteins, suggesting that these proteins should be promoted to the trusted protein set for future proteome releases.

The specific set of promoted proteins have several characteristics which distinguish them from other proteins in the working set. Promoted working set proteins tend to be longer (247 amino acids) than the average working set protein (130 amino acids). Over 88% of promoted working set proteins begin with the canonical start site encoding Methionine. In contrast, only 67% of all working set proteins have this starting amino acid. We propose that Methionine at the translation start site is important for correct expression. The filtered set is predominantly comprised of Methionine-initiated proteins (98% of proteins in the filtered set). Our set of promoted working set proteins indicates that alternate start sites may play a larger role in plant proteomes than evidenced by the current genome annotation. We observe another distinction between working set proteins and filtered set proteins regarding the number of independent sources of evidence supporting the models. Sources of evidence reported from the gene sets include cDNA, ESTs, mRNA, *ab initio* predictions, and protein mappings from both Maize and other species. We found that 65% of filtered set proteins had 2 or more types of evidence, while only 23% of working set proteins had as much evidence. The lack of evidence for most working set proteins is likely a contributing factor in the classification. Of the working set proteins we believe should be promoted, 45% had at least 2 other types of evidence. It appears that peptide mass spectrometry provides an orthogonal source of information that improves the identification of a trusted proteome.

Deep Versus Broad Sampling

In a previous study [48], we demonstrated that broad sampling of a diverse set of tissues improves coverage of the proteome by MS/MS data. In this study, however, we evaluate the benefit of deep sampling of a small collection of tissues. In this study, we used a more conservative scoring scheme compared to the previous study. To create a fair comparison, we re-scored the Arabidopsis peptides using an identical procedure to maize (see Methods) resulting in the identification of 128,432 distinct peptides. Compared to the arabidopsis study, the maize study analyzed 5 times more spectra, but identified only 50% more peptides. In part, the diminished return in maize is due to the much larger maize genome, which resulted in a protein sequence database 10 times larger than the arabidopsis database. Larger search databases are known to reduce sensitivity. However, the reduction in the number of peptides is also due to repeated sampling of the same peptide species. We observe on average 12 clustered spectra (24 raw spectra) supporting each peptide in our maize study, compared to 19 raw spectra per peptide in Arabidopsis.

While broad sampling helps increase the diversity expressed peptides, deep sampling allows us a robust, label-free quantification of each peptide. A Maize Protein Atlas describing the comparative abundance will be presented in a companion study [118]. In addition to deeper sampling of each peptide, we also observe greater coverage for the identified proteins. Among all maize proteins that have at least one identified peptide (unique or shared), we achieve 27% amino acid coverage on average (13 peptides per protein). In contrast, in arabidopsis we only achieved 20% coverage on average (9 peptides per protein). Multiple peptides identified for a protein can dramatically improve the spectral-count based quantification of proteins, and help in improved gene structure annotation.

About 25% of the peptides (49,778 peptides) do not map uniquely to a location, likely due to the presence of duplicated gene copies in maize. We group sets of indistinguishable proteins into protein groups (see Methods). These ‘paralogous locus groups’ of proteins are discussed below.

4.2.2 Discovery of Novel Events

Peptide sequences which do not match to either the filtered gene set or the working gene set are considered novel peptides. We identify 24,782 novel peptides matching to 91,059 locations in the genome. Many of the novel peptides match to a single location in our genomic databases (16,659, 67.2%). Upon clustering the peptides, we identify 6,384 novel annotation refinement events. We require that each event have at least one uniquely-located peptide, but only a single peptide is required to constitute an event. The identified events by type and the number of affected genes are shown in Table 4.1. Using the peptides from the filtered novel events, combined with over 2 million *Zea mays* ESTs, 875 million RNA-Seq reads, and the mappings of proteins from rice and sorghum, we generated putative gene structures using the *ab initio* tool, Augustus [69]. The gene predictions and proteogenomic evidence, together with the Maize Protein Atlas, can be found at

Table 4.1: Each event is defined by a collection of peptides and a protein. Therefore, it is possible for a single collection of peptides to count towards multiple events on different proteins.

Event Type	Events Identified	Genes Affected
Antisense Translation	1,019	812
Frame Change	678	507
Exon Extension	520	371
Novel Exon	624	440
Novel Splice	69	48
Translated UTR	706	504
Gene Extension	1066	799

http://maizeproteome.ucsd.edu/

Our stringent *eventProb* cutoffs and the requirement for at least one uniquely located peptide to be present in each event, results in 7,572 novel peptides (30.6% of the 24,782 novel peptides identified) contributing to accepted events. Nearly half of the peptides not contributing to events map to two or more genomic locations (45.22%), a much higher fraction than shared peptides among all novel peptides (34.78%). We believe that this prevalence of shared peptides is a result of the high level of sequence redundancy in the genome. While we cannot localize the protein expression to a single locus, we find multiple loci which share most of their peptides (See results for Paralagous Locus Groups).

4.2.3 Revision of Annotated Genes

Using the filtered novel events as additional evidence, we constructed revised gene models with peptide support for 733 working set genes and 1,256 filtered set genes. We used augustus to predict updated gene models, using novel peptides, ESTs, RNA-Seq, homology with rice and sorghum, and the current gene models as hints. For 1,989 genes, we predicted at least one novel transcript supported by the novel peptides.

A key challenge in genome annotation is distinguishing pseudogenes from protein coding genes. Pseudogenes are believed to be genes which have lost their capacity to produce proteins, and therefore are considered non-functional. We find annotated pseudogenes that are incorrectly labeled. It is likely that many of

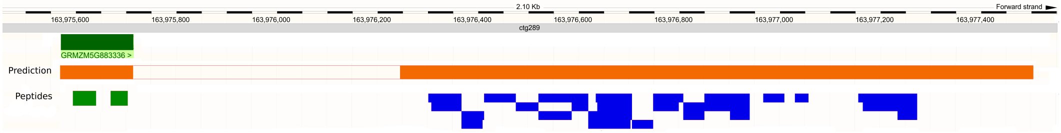
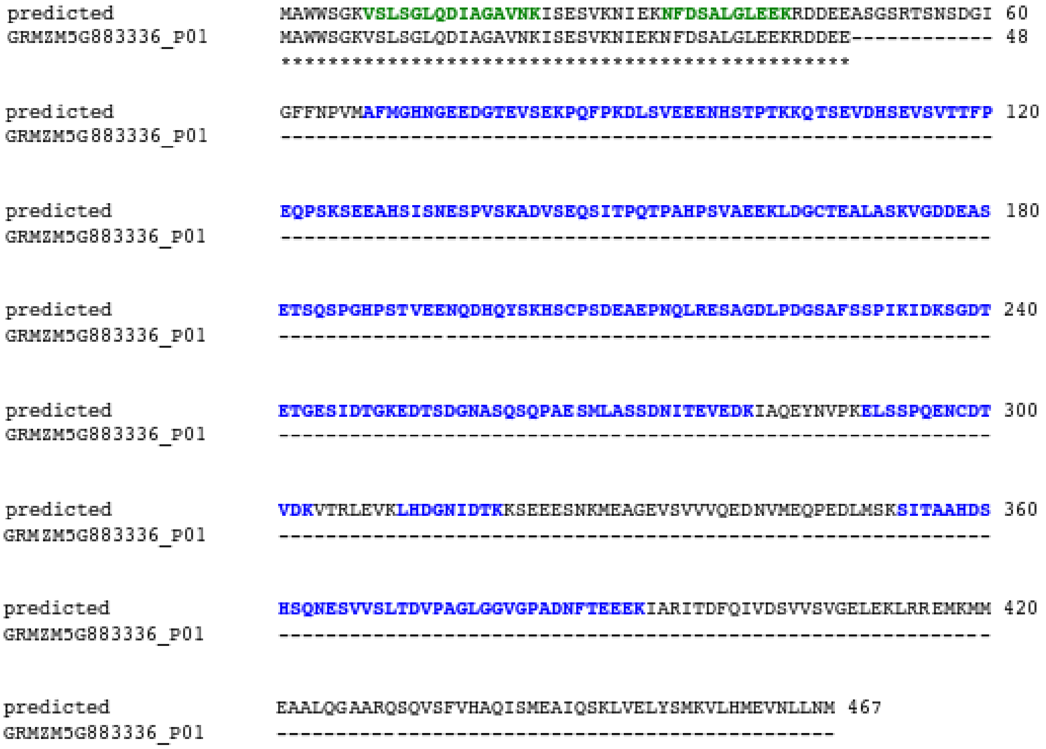


Figure 4.1: We observe 2 peptides matching to the annotated pseudogene, GRMZM5G883336, as well as 25 novel peptide sequences downstream of the gene. The predicted sequence extends the annotated gene by over 400 amino acids. Our peptides confirm that this locus is translated and should be reclassified as protein-coding.

these loci were labeled as pseudogenes because they are short or show little homology with other proteins. As an example, gene GRMZM5G883336 is annotated as a pseudogene in the 5a.59 gene annotation. We propose that this gene is in fact translated, as we identify two peptides that match the protein produced at this locus, one of which matches uniquely. We also observe 25 distinct, uniquely located novel peptides downstream of the gene (Figure 4.1). We predict that GRMZM5G883336 should be extended in the 3' direction by 419 amino acids. The alignment of the annotated gene sequence and the predicted gene sequence is shown in Figure 4.2.

Determining whether a translated region is a single gene or two proximal genes is a difficult problem for gene annotation. We observe cases where two or more genes were annotated, but given the evidence we predict a single gene. For example, two peptides are identified in the 5' UTR of GRMZM5G881353 giving evidence for an extension of the protein coding region towards GRMZM5G831724. Given these peptides, we predict an updated gene model that merges GRMZM5G83 1724 and GRMZM5G881353. In addition, we observe peptides downstream of the genes, suggesting a gene extension. Figure 4.3 shows the two genes as well as the novel prediction. Both GRMZM5G831724 and GRMZM5G881353 are annotated as pseudogenes, however our peptide evidence suggests that these loci are translated.

The predicted protein sequence has significant sequence similarity (evalue = 2E-173) to annotated maize protein GRMZM5G376743_P01. In Figure 4.4 the alignment between the predicted sequence and GRMZM2G376743. The sequences



```
predicted             MAWWSGKVSLSGLQDIAGAVNKISESVKNIEKNFDSALGLEEKRDDEEASGSRTSNSDGI 60
GRMZM5G883336_P01     MAWWSGKVSLSGLQDIAGAVNKISESVKNIEKNFDSALGLEEKRDDEE------------ 48
                      ************************************************

predicted             GFFNPVMAFMGHNGEEDGTEVSEKPQFPKDLSVEEENHSTPTKKQTSEVDHSEVSVTTFP 120
GRMZM5G883336_P01     ------------------------------------------------------------

predicted             EQPSKSEEAHSISNESPVSKADVSEQSITPQTPAHPSVAEEKLDGCTEALASKVGDDEAS 180
GRMZM5G883336_P01     ------------------------------------------------------------

predicted             ETSQSPGHPSTVEENQDHQYSKHSCPSDEAEPNQLRESAGDLPDGSAFSSPIKIDKSGDT 240
GRMZM5G883336_P01     ------------------------------------------------------------

predicted             ETGESIDTGKEDTSDGNASQSQPAESMLASSDNITEVEDKIAQEYNVPKELSSPQENCDT 300
GRMZM5G883336_P01     ------------------------------------------------------------

predicted             VDKVTRLEVKLHDGNIDTKKSEEESNKMEAGEVSVVVQEDNVMEQPEDLMSKSITAAHDS 360
GRMZM5G883336_P01     ------------------------------------------------------------

predicted             HSQNESVVSLTDVPAGLGGVGPADNFTEEEKIARITDFQIVDSVVSVGELEKLRREMKMM 420
GRMZM5G883336_P01     ------------------------------------------------------------

predicted             EAALQGAARQSQVSFVHAQISMEAIQSKLVELYSMKVLHMEVNLLNM 467
GRMZM5G883336_P01     -----------------------------------------------
```

Figure 4.2: The predicted sequence contains the annotated sequence, and adds a 419 amino acid extension to the C terminus of the protein.

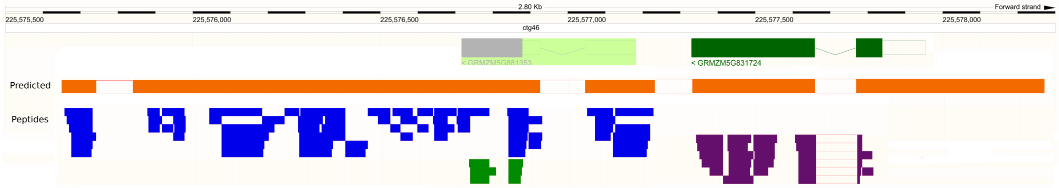


Figure 4.3: We observe 44 peptides (green and purple) uniquely matching to two annotated genes, GRMZM5G881353 and GRMZM5G831724. In addition, we observe 27 uniquely located novel peptides (blue) in the UTR of GRMZM5G881353 as well as downstream of both genes. We predict a single gene at this location which is consistent with all peptides.

are sufficiently distinct that we can uniquely map the peptides to the refined locus. Using blastp against the non-rendundant protein sequence database (nr) at NCBI, we found that the single gene model is conserved in Brachypodium distachyon.

As next generation sequencing of RNA molecules becomes a popular and inexpensive assay for expression, we will doubtless see an increase in genome annotation studies incorporating it. For protein coding genes, however, determining the frame of translation for an exon is a difficult problem, one which is not unambiguously solved by transcriptomics. Instead, proteomics can be used to address this need. We identify novel peptides at 507 genes suggesting that the annotated frame is incorrect.

For example, we identify 12 uniquely located peptides matching GRMZM2G 121186_P01, a nucleosome/chromatin assembly factor group A (nfa102). In addition, we find 41 novel peptides overlapping the annotation but in a different frame. The proposed revision is consistent with the N terminus of the annotated protein, but the final two exons, consisting of 78 altered amino acids, are updated slightly. The alignment between the revised protein and the annotated protein is shown in Figure 4.5.

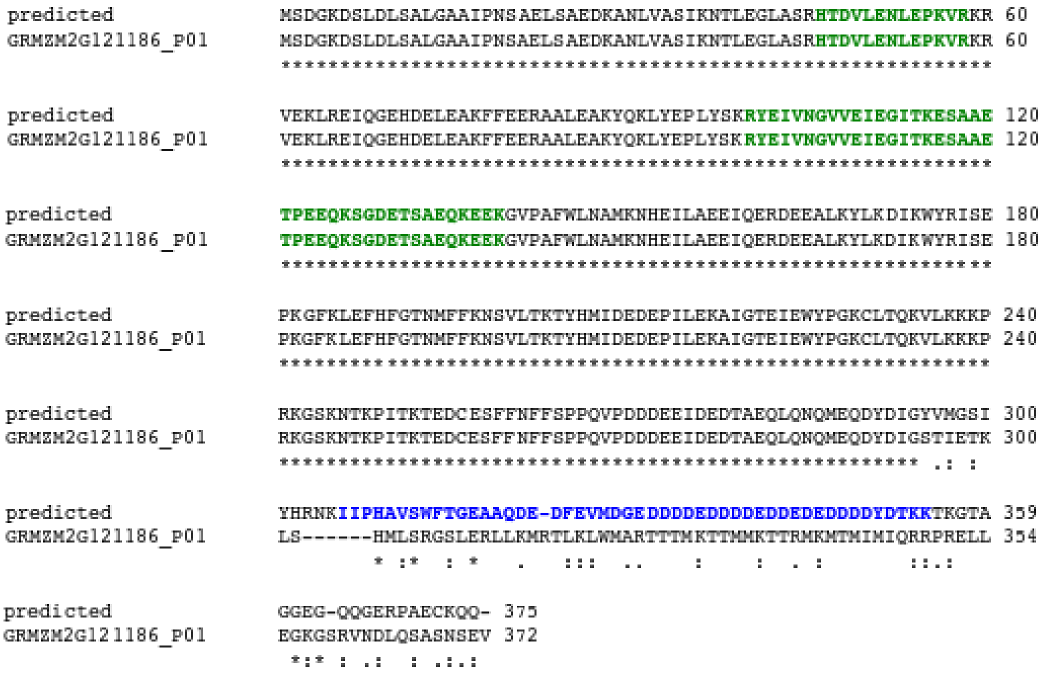
The coordinates of the predicted sequence do not differ significantly from the annotation. In fact, the prediction contains a late splice donor site, extending exon 9 by 16 nucleotides (See Figure 4.6). Exon 10 in the prediction is identical in coordinates to exon 10 in the annotation, however, the difference in frame results in a dramatically different C terminus.

Discovery of Novel Genes

We identified 1,702 novel gene events. Occasionally, Augustus chooses to split some novel gene events into two or more distinct gene model predictions based on the available evidence. In total, augustus predicted 3,713 novel protein sequences in 3,676 genes. 1,113 (29%) of the novel proteins in 1,102 genes were fully consistent with at least one of the uniquely-located peptides at the locus. Of these proteins, 785 were only supported by peptides and would not have been identified in a transcriptomic-based annotation pipeline. We used blastp to identify paralogs

```
predicted                 MPTSRSTEAGKGSTRHGGVACPPPACPCPARPPPDGPRDHEISSSVIIPFQSSSRFSLPK 60
GRMZM2G376743_P01         ------------------------------------------------------------


predicted                 PNRSTAVSAPATRPPALPAINHVCVEVALASPARSYYILLIVEHPSTHAPTLAIASQARY 120
GRMZM2G376743_P01         ------------------------------------------------------------


predicted                 KQPDIQDHRSRSQKQRLCVATSTAMDAPAMFARKQVPDDFGLRNIGEAEDEGAGAQAQAQ 180
GRMZM2G376743_P01         ------------------------MDAPAMFTRKHVPEDIGLRNIGVGEAEDDGAGGQAQ 36
                                                  *******:**:**:*:****** .* *. ** .***

predicted                 LHRDEKQHKPAVLKKVKEKVKKIKNTLAGHGHGHDGEHGGGERMG--DADAGSDSSEEAE 238
GRMZM2G376743_P01         LHRDEKQHKP-VLKKVKEKVKKIKSTLAGHGHGHDGEHGGDERMGEEDAYAGSDSSEEGE 95
                          ********** *************.***************.****  ** ********.*

predicted                 EDAAEREAALEKGGYMEDVEDKPLVME--PEPNPELHGAPMYESARAPAVQDLIAEYDQP 296
GRMZM2G376743_P01         EDAVEREAALEKDGYMEDIEDKPVVMESDPDPDPDVHGAPMYESARAPAVQDLVAEYDQP 155
                          ***.********.*****:****:***  *:*:*::*****************:******

predicted                 AWTPAVQEVEGDGSAPRVRLGDIGGPVVEDPAAPRSTTPTAREGEDIGTTPVVRQFETMS 356
GRMZM2G376743_P01         AWTPAVREVEGDGVSPRVRLGDVGGPVVEDPAAPRSKTPAAREGEDIGTTPVVQQFETMS 215
                          ******:****** :*******:*************.**:*************:******

predicted                 LSDDPTHVGAGKKGAKAEEWKDKAADTVRGADGGGGGGASYTDTLKNAAAGTTEYGKKLA 416
GRMZM2G376743_P01         LSDGPAHVGAAKEGAKAEEWEGNAADMVGGVAGG----ASYTDRLKNAAAGTTEYGRKLA 271
                          ***.*:****.*:*******:.:*** * *. **    ***** ************:***

predicted                 STVYEKVAGVTT----AVGVGKRGDERTEATPASNTRTEERGAAPEATDATRGAGYTDKI 472
GRMZM2G376743_P01         STVYEKVAGAGTGTGTAVGVGKRDDERTEAVPASNTRT----------DAARGAGYTDNI 321
                          *********. *    *******.******.*******          **:*******:*

predicted                 KSAAAGTTGYGRQLASTVYEKVAGVGTAVAPNLRPQEGSAKAEGAHSEATPVSDTGVEEW 532
GRMZM2G376743_P01         KSAAAGTTGYGKQLASTVYDKVAGVGTAVAPSRRQQEGSAKAEGAHSEAMPVSDTGAEEW 381
                          ***********:*******:***********. * ************** ******.***

predicted                 KGAPAATDAANGASGPAGYTHKVKSAAAGTTEYGKQLASTVYEKVAGVGTAVAGKVQQAT 592
GRMZM2G376743_P01         KDAPATVEEANTESRP-GYTDKIRSAAAGTTEYGKQLASTVYEKVASVGTAVAGKVQQAT 440
                          *.***:.: **  * * ***.*::**********************.*************

predicted                 QSPSTATPGADAQQDTDAAATATPGAGGQSKGTTVTGYIAEKLRPGDEDRALSEAISGAV 652
GRMZM2G376743_P01         QSAGTATPGVGAQRDTGAAATATPGAGEQDKGVTVTGYIAEKLRPGDEDRALSEAISGAV 500
                          **..*****..**:**.********** *.**.***************************

predicted                 QRRKEDVGGTVAQRVPAPGQVVTKAREAVASLTGGKRVSETVQPTTATGKDVKEVYAAEA 712
GRMZM2G376743_P01         QRRKDDAGDTVVQRVPAPG-------QAVASLTGGNRVSETVQPTTATGEDVKEGYAAEA 553
                          ****:*.*.**.*******       :********:*************:**** *****

predicted                 PVIHGEEIGGPKLNTNTM 730
GRMZM2G376743_P01         PVIRGEEIGGAKLNTNAI 571
                          ***:******.*****::
```

Figure 4.4: The predicted sequence is aligned to paralogous protein, GRMZM2G376743_P01. Peptides that match to the annotated protein GRMZM831724 as well as the newly predicted protein are shown in purple. Peptides matching both GRMZM5G881353 and the predicted sequence are shown in green. Novel peptides are highlighted in blue. For each peptide, amino acid differences prevent the peptide from matching GRMZM2G376743_P01.



```
predicted            MSDGKDSLDLSALGAAIPNSAELSAEDKANLVASIKNTLEGLASRHTDVLENLEPKVRKR 60
GRMZM2G121186_P01    MSDGKDSLDLSALGAAIPNSAELSAEDKANLVASIKNTLEGLASRHTDVLENLEPKVRKR 60
                     ************************************************************

predicted            VEKLREIQGEHDELEAKFFEERAALEAKYQKLYEPLYSKRYEIVNGVVEIEGITKESAAE 120
GRMZM2G121186_P01    VEKLREIQGEHDELEAKFFEERAALEAKYQKLYEPLYSKRYEIVNGVVEIEGITKESAAE 120
                     ************************************************************

predicted            TPEEQKSGDETSAEQKEEKGVPAFWLNAMKNHEILAEEIQERDEEALKYLKDIKWYRISE 180
GRMZM2G121186_P01    TPEEQKSGDETSAEQKEEKGVPAFWLNAMKNHEILAEEIQERDEEALKYLKDIKWYRISE 180
                     ************************************************************

predicted            PKGFKLEFHFGTNMFFKNSVLTKTYHMIDEDEPILEKAIGTEIEWYPGKCLTQKVLKKKP 240
GRMZM2G121186_P01    PKGFKLEFHFGTNMFFKNSVLTKTYHMIDEDEPILEKAIGTEIEWYPGKCLTQKVLKKKP 240
                     ************************************************************

predicted            RKGSKNTKPITKTEDCESFFNFFSPPQVPDDDEEIDEDTAEQLQNQMEQDYDIGYVMGSI 300
GRMZM2G121186_P01    RKGSKNTKPITKTEDCESFFNFFSPPQVPDDDEEIDEDTAEQLQNQMEQDYDIGSTIETK 300
                     ****************************************************** .: : 

predicted            YHRNKIIPHAVSWFTGEAAQDE-DFEVMDGEDDDDEDDDDEDDEDEDDDDYDTKKTKGTA 359
GRMZM2G121186_P01    LS------HMLSRGSLERLLKMRTLKLWMARTTTMKTTMMKTTRMKMTMIMIQRRPRELL 354
                             * :*  : *   .   :::  ..    :    :  . :       ::.:

predicted            GGEG-QQGERPAECKQQ- 375
GRMZM2G121186_P01    EGKGSRVNDLQSASNSEV 372
                      *:* : .:  : .:.:
```

Figure 4.5: The predicted sequence is aligned to the annotated sequence at gene GRMZM2G121186. The N terminal portion of the prediction is consistent with the annotation, and we observe 12 peptides matching this region. We also identify 41 distinct novel peptide sequences matching the C terminus of the predicted sequence, which is in a different frame from the annotation.

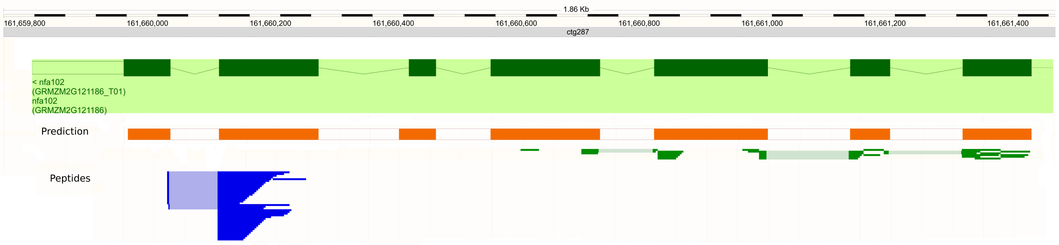
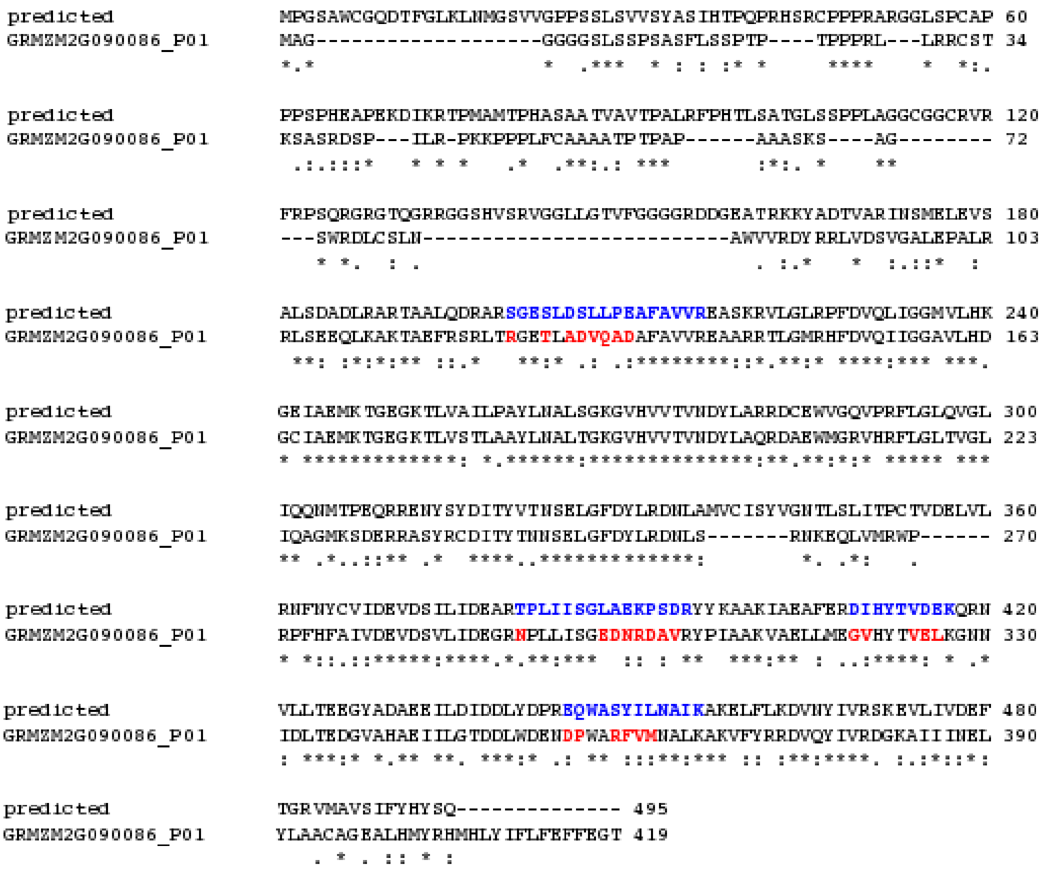


Figure 4.6: The locus GRMZM2G121186 with both the annotated gene model and the predicted gene model shown. The predicted gene model is only slightly different in coordinates from the annotated model, but even the small different in coordinates causes a large change in over 70 amino acids.

of the novel protein sequences in the annotated maize proteome. We found 302 of the 1,113 peptide-supported novel proteins that have significant sequence similarity to a known protein in maize (evalue < 1E-10), suggesting that these new proteins are extensions of gene families. Blastp was also used to identify conserved domains and protein homologs between predicted protein sequences and nr.

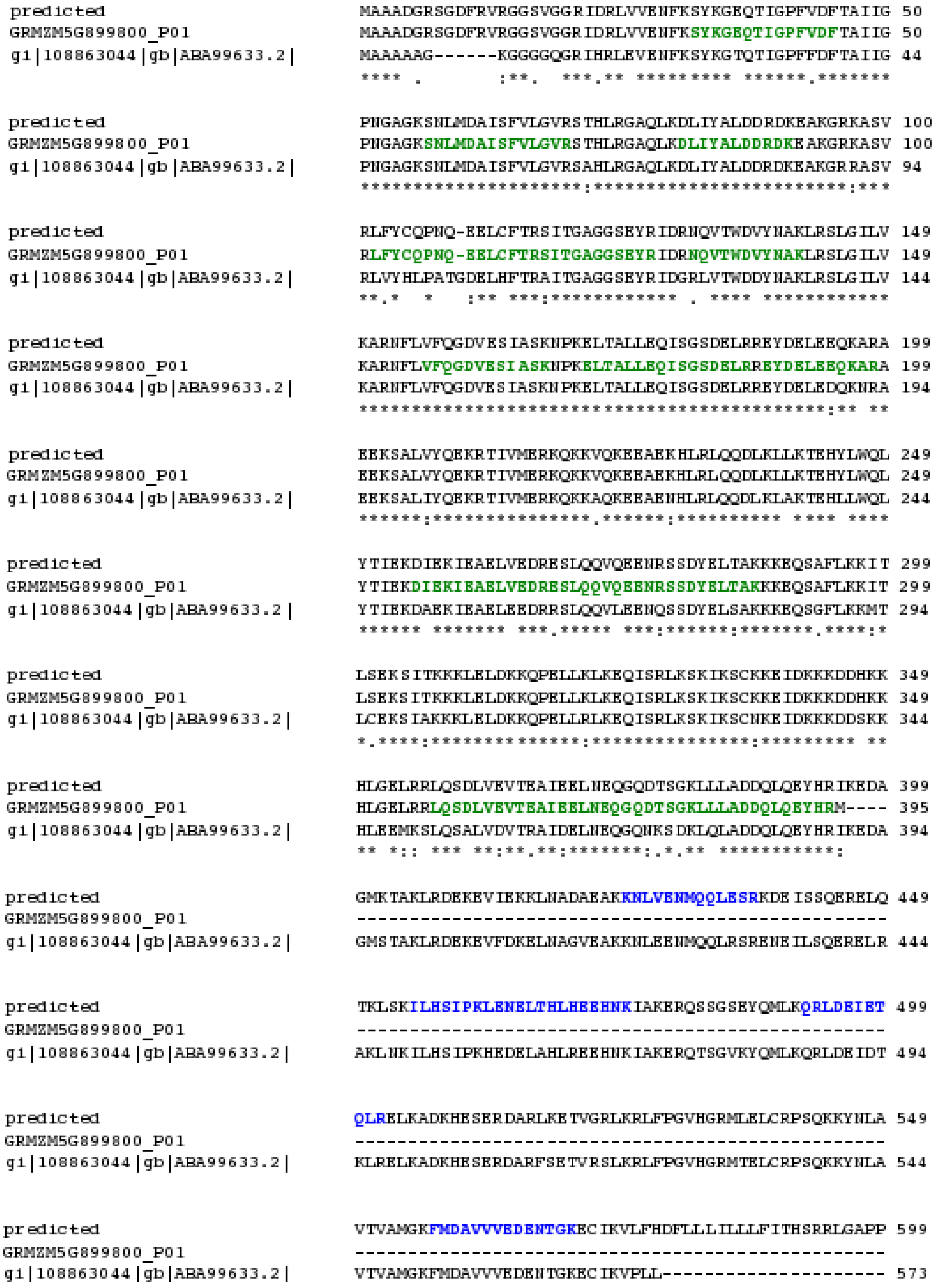
On chromosome 3, we identify a novel gene with strong sequence similarity to maize protein GRMZM2G090086_P01 (evalue = 4E-88). While the annotated protein is not functionally annotated at MaizeSequence.org, we believe this protein is translocase subunit SecA from its similarity to the protein by that name (Gene ID: 100841935) in Brachypodium distachyon (evalue = 0.0). Figure 4.7 shows the alignment between our predicted novel protein and the closest maize protein, GRMZM2G090086_P01. The four identified novel peptides, bolded and highlighted in blue, are unique to this genomic location. The amino acid differences between the two proteins support the assignment of these peptides to the novel protein. The amino acid differences between the two proteins at the identified peptides are bolded and highlighted in red on GRMZM2G090086_P01. We note that a *Zea mays* protein in nr (Gene ID: 542347 tha1) has a near perfect match to our predicted protein, but appears in a truncated form. Our predicted protein contains 495 amino acids, while the protein in nr has only 291 amino acids.

We identify a locus on chromosome 1 that bears sequence similarity to annotated maize protein GRMZM5G899800_P01 (evalue = 0.0). Again, this protein does not have a functional annotation, but similarity to a protein in Oryza sativa (Gene ID: 108863044, evalue = 0.0) suggests that these proteins are Structural Maintenance of Chromosomes (SMC) domain containing proteins. In fact, the rice protein suggests that the annotated maize protein, GRMZM5G899800_P01 may be truncated. The alignment of our predicted protein, GRMZM5G899800_P01, and the SMC N terminal domain containing protein is shown in Figure 4.8. The seven identified peptides, bolded and highlighted in blue, are unique to this genomic location. We also observe 21 peptides uniquely matching GRMZM2G899800_P01, although if our predicted sequence was promoted to a gene, these peptides would be shared across both loci.



```
predicted           MPGSAWCGQDTFGLKLNMGSVVGPPSSLSVVSYASIHTPQPRHSRCPPPRARGGLSPCAP 60
GRMZM2G090086_P01   MAG-------------------GGGGSLSSPSASFLSSPTP----TPPPRL---LRRCST 34
                    *.*                   *  .***  * : : :* *     ****    *  *:.

predicted           PPSPHEAPEKDIKRTPMAMTPHASAATVAVTPALRFPHTLSATGLSSPPLAGGCGGCRVR 120
GRMZM2G090086_P01   KSASRDSP---ILR-PKKPPPLFCAAAATPTPAP------AAASKS----AG-------- 72
                     .:.:::*   * * *   .*  .**:.: ***       :*:. *    **

predicted           FRPSQRGRGTQGRRGGSHVSRVGGLLGTVFGGGGRDDGEATRKKYADTVARINSMELEVS 180
GRMZM2G090086_P01   ---SWRDLCSLN--------------------------AWVVRDYRRLVDSVGALEPALR 103
                       * *.  : .                             . :.*   *  :.::*  :

predicted           ALSDADLRARTAALQDRARSGESLDSLLPEAFAVVREASKRVLGLRPFDVQLIGGMVLHK 240
GRMZM2G090086_P01   RLSEEQLKAKTAEFRSRLTRGETLADVQADAFAVVREAARRTLGMRHFDVQIIGGAVLHD 163
                     **: :*:*:** ::.*   **:* .: .:********::*.**:* ****:*** ***.

predicted           GEIAEMKTGEGKTLVAILPAYLNALSGKGVHVVTVNDYLARRDCEWVGQVPRFLGLQVGL 300
GRMZM2G090086_P01   GCIAEMKTGEGKTLVSTLAAYLNALTGKGVHVVTVNDYLAQRDAEWMGRVHRFLGLTVGL 223
                    * *************: *.******:**************:**.**:*:* ***** ***

predicted           IQQNMTPEQRRENYSYDITYVTNSELGFDYLRDNLAMVCISYVGNTLSLITPCTVDELVL 360
GRMZM2G090086_P01   IQAGMKSDERRASYRCDITYTNNSELGFDYLRDNLS-------RNKEQLVMRWP------ 270
                    ** .*..::** .*  ****..*************:        *. .*:   .

predicted           RNFNYCVIDEVDSILIDEARTPLIISGLAEKPSDRYYKAAKIAEAFERDIHYTVDEKQRN 420
GRMZM2G090086_P01   RPFHFAIVDEVDSVLIDEGRNPLLISGEDNRDAVRYPIAAKVAELLMEGVHYTVELKGNN 330
                    * *::.::*****:****.*.**:***  :: : **  ***:** : ..:****: * .*

predicted           VLLTEEGYADAEEILDIDDLYDPREQWASYILNAIKAKELFLKDVNYIVRSKEVLIVDEF 480
GRMZM2G090086_P01   IDLTEDGVAHAEIILGTDDLWDENDPWARFVMNALKAKVFYRRDVQYIVRDGKAIIINEL 390
                    : ***:* *.** **. ***:* .: ** :::**:*** :: :**:****. :.:*::*:

predicted           TGRVMAVSIFYHYSQ-------------- 495
GRMZM2G090086_P01   YLAACAGEALHMYRHMHLYIFLFEFFEGT 419
                       . * . :: * :
```

Figure 4.7: The alignment of a novel protein to the closest maize protein, GRMZM2G090086_P01.



```
predicted                      MAAADGRSGDFRVRGGSVGGRIDRLVVENFKSYKGEQTIGPFVDFTAIIG 50
GRMZM5G899800_P01              MAAADGRSGDFRVRGGSVGGRIDRLVVENFKSYKGEQTIGPFVDFTAIIG 50
gi|108863044|gb|ABA99633.2|    MAAAAAG------KGGGGQGRIHRLEVENFKSYKGTQTIGPFFDFTAIIG 44
                               **** .       :**.  ***.** ********* ******.*******

predicted                      PNGAGKSNLMDAISFVLGVRSTHLRGAQLKDLIYALDDRDKEAKGRKASV 100
GRMZM5G899800_P01              PNGAGKSNLMDAISFVLGVRSTHLRGAQLKDLIYALDDRDKEAKGRKASV 100
gi|108863044|gb|ABA99633.2|    PNGAGKSNLMDAISFVLGVRSAHLRGAQLKDLIYALDDRDKEAKGRRASV 94
                               *********************:************************:***

predicted                      RLFYCQPNQ-EELCFTRSITGAGGSEYRIDRNQVTWDVYNAKLRSLGILV 149
GRMZM5G899800_P01              RLFYCQPNQ-EELCFTRSITGAGGSEYRIDRNQVTWDVYNAKLRSLGILV 149
gi|108863044|gb|ABA99633.2|    RLVYHLPATGDELHFTRAITGAGGSEYRIDGRLVTWDDYNAKLRSLGILV 144
                               **.*  *   :** ***:************ . **** ************

predicted                      KARNFLVFQGDVESIASKNPKELTALLEQISGSDELRREYDELEEQKARA 199
GRMZM5G899800_P01              KARNFLVFQGDVESIASKNPKELTALLEQISGSDELRREYDELEEQKARA 199
gi|108863044|gb|ABA99633.2|    KARNFLVFQGDVESIASKNPKELTALLEQISGSDELRREYDELEDQKNRA 194
                               ********************************************:** **

predicted                      EEKSALVYQEKRTIVMERKQKKVQKEEAEKHLRLQQDLKLLKTEHYLWQL 249
GRMZM5G899800_P01              EEKSALVYQEKRTIVMERKQKKVQKEEAEKHLRLQQDLKLLKTEHYLWQL 249
gi|108863044|gb|ABA99633.2|    EEKSALIYQEKRTIVMERKQKKAQKEEAENHLRLQQDLKLAKTEHLLWQL 244
                               ******:***************.******:********** **** ****

predicted                      YTIEKDIEKIEAELVEDRESLQQVQEENRSSDYELTAKKKEQSAFLKKIT 299
GRMZM5G899800_P01              YTIEKDIEKIEAELVEDRESLQQVQEENRSSDYELTAKKKEQSAFLKKIT 299
gi|108863044|gb|ABA99633.2|    YTIEKDAEKIEAELEEDRRSLQQVLEENQSSDYELSAKKKEQSGFLKKMT 294
                               ****** ******* ***.***** ***:******:*******.****:*

predicted                      LSEKSITKKKLELDKKQPELLKLKEQISRLKSKIKSCKKEIDKKKDDHKK 349
GRMZM5G899800_P01              LSEKSITKKKLELDKKQPELLKLKEQISRLKSKIKSCKKEIDKKKDDHKK 349
gi|108863044|gb|ABA99633.2|    LCEKSIAKKKLELDKKQPELLRLKEQISRLKSKIKSCNKEIDKKKDDSKK 344
                               *.****:**************:***************:********* **

predicted                      HLGELRRLQSDLVEVTEAIEELNEQGQDTSGKLLLADDQLQEYHRIKEDA 399
GRMZM5G899800_P01              HLGELRRLQSDLVEVTEAIEELNEQGQDTSGKLLLADDQLQEYHRM---- 395
gi|108863044|gb|ABA99633.2|    HLEEMKSLQSALVDVTRAIDELNEQGQNKSDKLQLADDQLQEYHRIKEDA 394
                               ** *:: *** **:**.**:*******:.*.** ***********:

predicted                      GMKTAKLRDEKEVIEKKLNADAEAKKNLVENMQQLESRKDEISSQERELQ 449
GRMZM5G899800_P01              --------------------------------------------------
gi|108863044|gb|ABA99633.2|    GMSTAKLRDEKEVFDKELNAGVEAKKNLEENMQQLRSRENEILSQERELR 444


predicted                      TKLSKILHSIPKLENELTHLHEEHNKIAKERQSSGSEYQMLKQRLDEIET 499
GRMZM5G899800_P01              --------------------------------------------------
gi|108863044|gb|ABA99633.2|    AKLNKILHSIPKHEDELAHLREEHNKIAKERQTSGVKYQMLKQRLDEIDT 494


predicted                      QLRELKADKHESERDARLKETVGRLKRLFPGVHGRMLELCRPSQKKYNLA 549
GRMZM5G899800_P01              --------------------------------------------------
gi|108863044|gb|ABA99633.2|    KLRELKADKHESERDARFSETVRSLKRLFPGVHGRMTELCRPSQKKYNLA 544


predicted                      VTVAMGKFMDAVVVEDENTGKECIKVLFHDFLLLILLLFITHSRRLGAPP 599
GRMZM5G899800_P01              --------------------------------------------------
gi|108863044|gb|ABA99633.2|    VTVAMGKFMDAVVVEDENTGKECIKVPLL--------------------- 573
```

Figure 4.8: The alignment of a novel protein to the closest maize protein, GRMZM2G899800_P01, and a functionally annotated rice protein.

Antisense Overlapping Translation

Overlapping natural antisense transcripts (NATs), once believed to predominantly exist in bacterial and viral genomes [119], have recently been observed in many eukaryotes [120] including humans [121] and *Arabidopsis thaliana* [83]. While most overlapping NATs are identified by transcriptional evidence [83, 122] and homology [123], there are few documented cases of overlapping NATs in which both transcripts are translated. There are may hypotheses attempting to explain how overlapping genes are regulated, most of which suggest that the simultaneous expression of both genes is unlikely. One model suggests that the mRNAs produced at overlapping antisense loci are complementary and therefore likely to produce double stranded RNA, preventing translation of either locus [124]. Another hypothesis points to the transcriptional complexes physically blocking one another when transcription on both strands is attempted [125].

In the maize 5a.59 annotation, there are 10,724 overlapping anti-sense gene pairs in the working gene set. Some large genes overlap multiple genes on the opposite strand, resulting a total of only 20,159 genes overlapping another gene on the opposite strand. We observe 3,470 overlapping pairs in the 5a.59 annotation where one of the genes is expressed in our data. We also observe 230 overlapping pairs where both genes are expressed.

Upon examining our novel peptides we observe 812 genes with novel peptides within 3 kilobases of the transcribed region, but on the opposite strand. Among these cases, we observe 243 genes with peptides also matching the gene on the annotated strand.

We find an unusual case where the 5a.59 genome annotation indicates overlapping anti-sense protein coding genes. Our novel peptides indicate that not only are both loci translated in seed tissues, but one of the annotations is incorrect. On chromosome 1, gene GRMZM2G149751 is annotated on the forward strand, while gene GRMZM2G449274 is annotated on the reverse strand. The transcribed regions of these two genes overlap. We observe 282 distinct peptide sequences matching to GRMZM2G149751 covering 85% of the one of the six annotated isoforms with the longest coding region (GRMZM2G2G149751_P05). We observe no

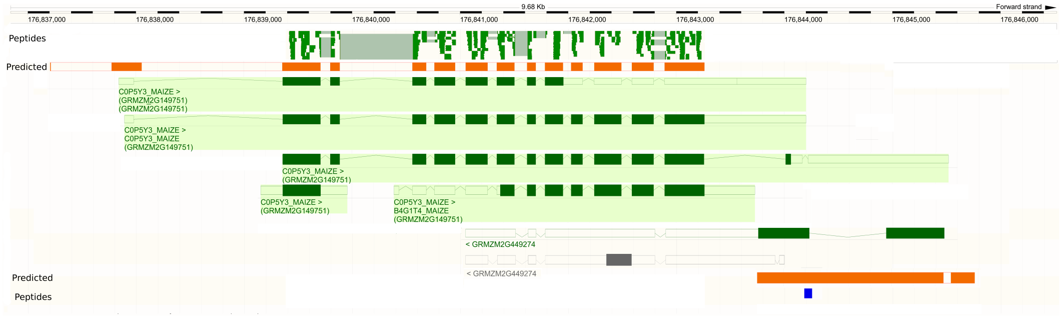
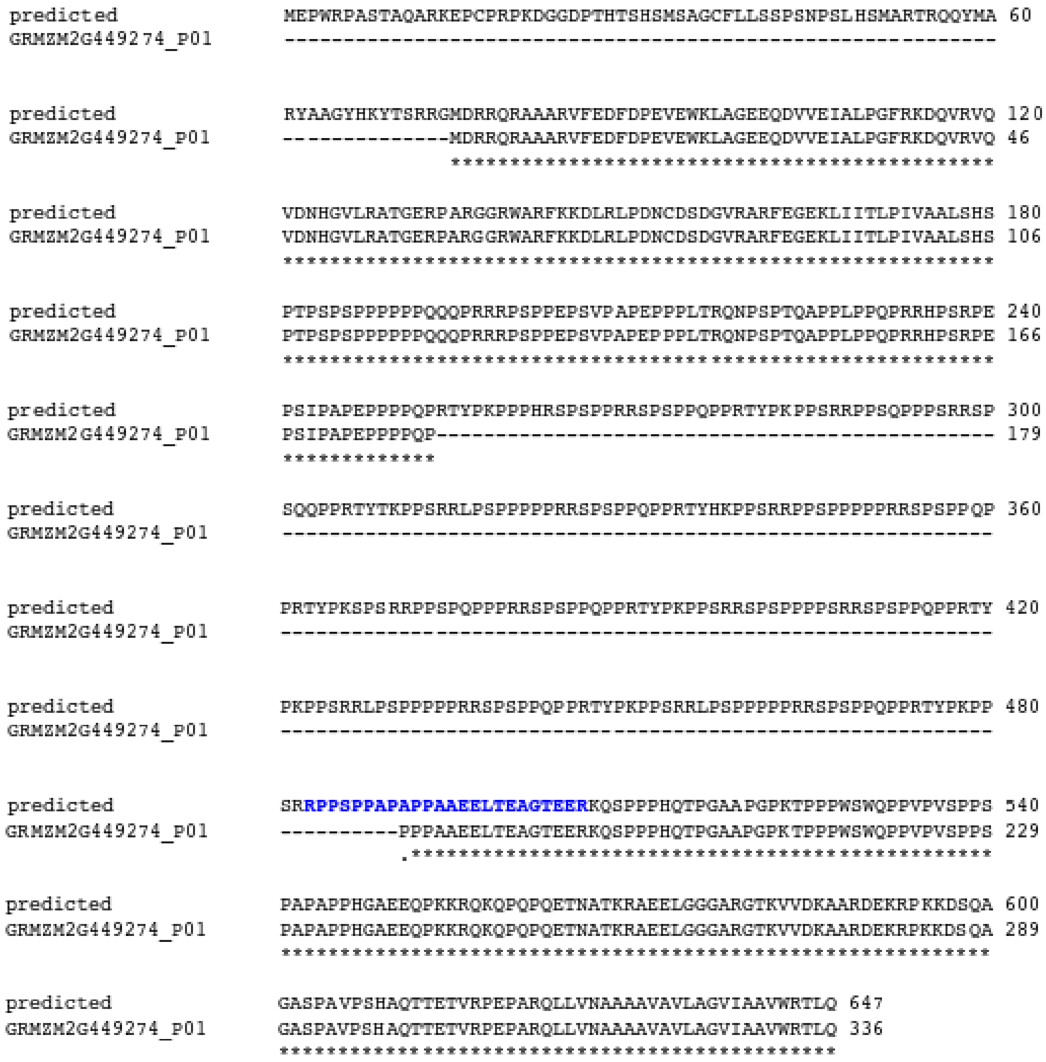


Figure 4.9: The genome browser view of a revised gene locus. GRMZM2G149751, on the forward strand, is confirmed by 282 distinct peptide sequences matching only this locus (highlighted in green and labeled ‘Peptides’). The novel peptide is shown on the reverse strand. We predict a novel protein coding gene on the reverse strand that is consistent with this peptide, but distinct from the annotated gene GRMZM2G449274.

peptides that can be localized to GRMZM2G449274.

We observe a novel peptide on the reverse strand, overlapping the intron from GRMZM2G449274. We predict a novel gene that is consistent with the novel peptide (Figure 4.9). The predicted protein indicates that the gene model at GRMZM2G449274 has a protein isoform consisting of both exons, but also including a read through of the annotated intron. The alignment of the predicted protein to the single protein product of GRMZM2G449274 is shown in Figure 4.10

While the prediction is based on a single peptide sequence, our data contains strong support for the peptide identification. The peptide is identified by spectra at two charge states. Figures 4.11 and 4.12 show the annotated spectra. The spectra in the images represent the clustered spectra produced by MSCluster [77] (see Methods), each clustered spectrum represents a single raw spectrum. The small number of identified spectra could indicate that the predicted protein is expressed at a low level, and only in the germ kernel. The average length of fully tryptic peptides in the predicted protein is only 7 amino acids, which is on the small side of the species identifiable by mass spectrometry. In contrast, there is abundant spectral evidence for the forward strand annotation, GRMZM2G149751 in endosperm, embryo, germ kernel, and pericarp/aleurone.



```
predicted             MEPWRPASTAQARKEPCPRPKDGGDPTHTSHSMSAGCFLLSSPSNPSLHSMARTRQQYMA 60
GRMZM2G449274_P01     ------------------------------------------------------------

predicted             RYAAGYHKYTSRRGMDRRQRAAARVFEDFDPEVEWKLAGEEQDVVEIALPGFRKDQVRVQ 120
GRMZM2G449274_P01     --------------MDRRQRAAARVFEDFDPEVEWKLAGEEQDVVEIALPGFRKDQVRVQ 46
                                    **********************************************

predicted             VDNHGVLRATGERPARGGRWARFKKDLRLPDNCDSDGVRARFEGEKLIITLPIVAALSHS 180
GRMZM2G449274_P01     VDNHGVLRATGERPARGGRWARFKKDLRLPDNCDSDGVRARFEGEKLIITLPIVAALSHS 106
                      ************************************************************

predicted             PTPSPSPPPPPPQQQPRRRPSPPEPSVPAPEPPPLTRQNPSPTQAPPLPPQPRRHPSRPE 240
GRMZM2G449274_P01     PTPSPSPPPPPPQQQPRRRPSPPEPSVPAPEPPPLTRQNPSPTQAPPLPPQPRRHPSRPE 166
                      ************************************************************

predicted             PSIPAPEPPPPQPRTYPKPPPHRSPSPPRRSPSPPQPPRTYPKPPSRRPPSQPPPSRRSP 300
GRMZM2G449274_P01     PSIPAPEPPPPQP----------------------------------------------- 179
                      *************

predicted             SQQPPRTYTKPPSRRLPSPPPPPRRSPSPPQPPRTYHKPPSRRPPSPPPPPRRSPSPPQP 360
GRMZM2G449274_P01     ------------------------------------------------------------

predicted             PRTYPKSPSRRPPSPQPPPRRSPSPPQPPRTYPKPPSRRSPSPPPPSRRSPSPPQPPRTY 420
GRMZM2G449274_P01     ------------------------------------------------------------

predicted             PKPPSRRLPSPPPPPRRSPSPPQPPRTYPKPPSRRLPSPPPPPRRSPSPPQPPRTYPKPP 480
GRMZM2G449274_P01     ------------------------------------------------------------

predicted             SRRPPSPPAPAPPAAEELTEAGTEERKQSPPPHQTPGAAPGPKTPPPWSWQPPVPVSPPS 540
GRMZM2G449274_P01     ----------PPPAAEELTEAGTEERKQSPPPHQTPGAAPGPKTPPPWSWQPPVPVSPPS 229
                                .*************************************************

predicted             PAPAPPHGAEEQPKKRQKQPQPQETNATKRAEELGGGARGTKVVDKAARDEKRPKKDSQA 600
GRMZM2G449274_P01     PAPAPPHGAEEQPKKRQKQPQPQETNATKRAEELGGGARGTKVVDKAARDEKRPKKDSQA 289
                      ************************************************************

predicted             GASPAVPSHAQTTETVRPEPARQLLVNAAAAVAVLAGVIAAVWRTLQ 647
GRMZM2G449274_P01     GASPAVPSHAQTTETVRPEPARQLLVNAAAAVAVLAGVIAAVWRTLQ 336
                      ***********************************************
```

Figure 4.10: The alignment of the predicted sequence to the protein product of GRMZM2G449274. The single novel peptide is highlighted in blue. The missing portion of the sequence of GRMZM2G449274 is aligned to the translated intron in the predicted sequence.

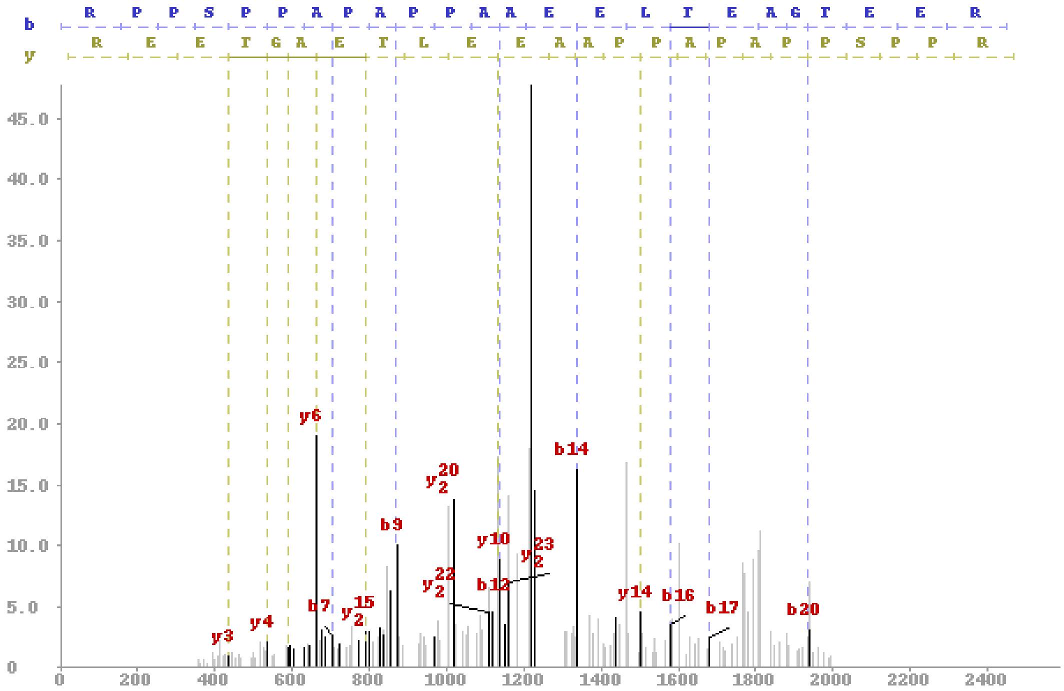


Figure 4.11: The annotated cluster spectrum from maize germ kernel (file 104, scan number 6176). This spectrum is annotated with the peptide in charge state 2.

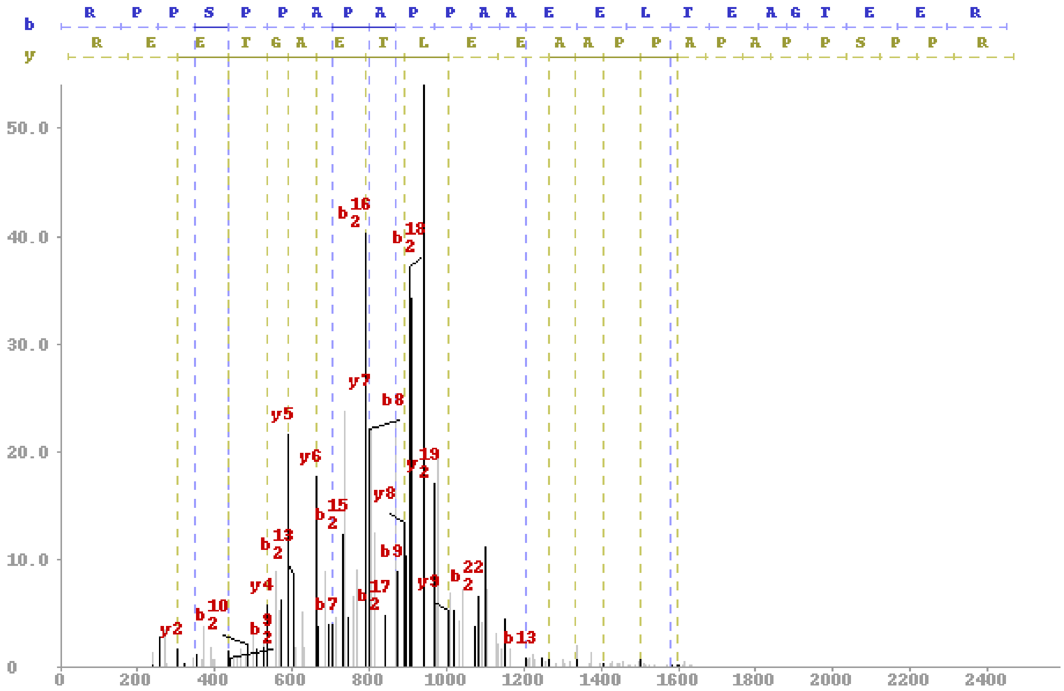


Figure 4.12: The annotated cluster spectrum from maize germ kernel (file 194, scan number 8342). This spectrum is annotated with the peptide in charge state 3.

4.2.4 Identification of Paralagous Locus Groups

Among annotated maize working set proteins, we identify 563 protein groups. Each protein group contains at least 2 proteins each arising from distinct genes. All proteins in a protein group are indistinguishable given the identified peptides. We find 1,899 proteins in the 563 groups defined by 2,992 shared peptides. Though we cannot determine which specific proteins from each protein group are expressed, we are confident that at least one of the proteins in each group exists in the samples.

We expect a similar phenomenon to exist among novel peptides. We create a bipartite graph to identify novel loci that share peptides as described in Methods. In total, our graph contains 3,613 peptides in 3,301 connected components. From the collection of paralogous locus groups, we selected 205 groups that have 2 or more peptides. We identified 55 locus groups containing 822 loci in which all of the loci were at least 3 kilobases from the nearest genes (equivalent to the criteria used to define novel gene events). Not unexpectedly, we find that many of the loci (444) overlap annotated repeats (downloaded from maizesequence.org). Upon analysis using blastx, aligning the nucleotide sequence from the loci to the non-redundant protein database, we found that the loci are likely unannotated retrotransposons.

4.3 Discussion

We presented a probabilistic framework, the *eventProb*, for scoring annotation events derived from mass spectrometry data. Proteogenomic studies, for which broad and deep sampling of the proteome is a key goal, generate enormous data sets. False positive identifications in these experiments arise from the millions of hypotheses that are being tested, one for each peptide-spectrum match. While controlling the false discovery rate at the level of the peptide-spectrum match addresses this problem, errors may be propagated and amplified by protein or event inference. We demonstrated that our eventProb can be used to limit the false discovery rate at the refinement event-level without sacrificing sensitivity.

Zea mays, with a human-scale genome and significant sequence redundancy, presented a challenging target for our proteogenomic methods. The parallization of the database search by InsPecT was crucial to making the analysis tractable. Engineering alone, however, was insufficient to address the problem of repetitive genomic regions. Instead we extended the idea of protein groups from protein inference to enable the identification of paralagous locus groups. While we cannot claim which of the loci are expressed, the identification of these groups improves our understanding of the complement of protein sequences which can be expressed by a genome.

By applying our semi-automated genome annotation method to *Zea mays*, we demonstrated that proteomics provides a much need line of evidence for the identification of protein-coding genes. Over 70% of the novel protein sequences supported by the peptide data were lacking corroborating transcriptomic and homology evidence. However, with current mass spectrometry technology it would be impossible to perform a complete genome annotation. The dynamic range of proteins spans 10 order of magnitude, well beyond the range measurable by modern mass spectrometers. Mass spectrometry is best utilized as a necessary and orthogonal source of genic information, to be coupled with transcriptomics and homology. In future work, we plan to add breadth to our mass spectrometry analysis by sampling additional tissues including leaves, roots, and reproductive organs which has been shown to reveal a distinct cross section of the detectable proteome [48].

4.4 Methods

A total of 109 million tandem mass spectra were acquired as described previously [118]. To reduce the computation resources required to analyze the spectra, we first clustered the raw spectra using MSCluster [77]. Clustering collapses the multiple spectra acquired for a single peptide into a single spectrum with a greater signal to noise ratio. In addition, we only analyzed clustered spectra with quality score larger than 0.2 as determined by PepNovo [126]. All subsequent analyses were performed on the clustered, quality filtered spectra.

First, the spectra were searched against a database containing annotated protein sequences as well as putative protein sequences. Using the putative protein sequences in the database, we will identify novel protein coding regions. The six-frame translation of the genome has previously been used with great success to capture novel coding regions for proteogenomics [48]. To keep the six-frame translation to a reasonable size, we only accepted translated reading frames with length of 40 amino acids greater. In eukaryotic organisms, spliced peptides provide a wealth of information about the structure of a gene. Whereas the six-frame translation does not include peptides which span splice boundaries, an exon splice graph [16] can compactly represent many putative splice junctions.

The exon splice graph takes many putative gene predictions, including alternate transcripts for the same gene that will include significant amounts of sequence redundancy. The graph construction routine then merges transcripts which share sequence, so that every exon prediction appears only once in the graph. To construct the exon splice graph for *Zea mays*, we downloaded approximately 2 million maize cDNA sequences from NCBI RefSeq (2010.10.07) and mapped them to the maize genome release RefGen v2 using BLAT [127]. We used the *ab initio* gene prediction tool, Augustus [69], to interpret the mappings and produce a set of gene predictions. The predictions were produced to include many structures for the same gene. The splice graphs constructed from this input were combined with the six-frame translation to create a database of putative protein sequences.

All tandem mass spectra were searched against the putative protein sequence databases and the annotated protein database (5a.59) using InsPect [7]. The scoring method used by InsPecT is based on a delta-score, the difference is match score between the best peptide and the second best peptide for a spectrum. Our databases contained over 2 billion amino acid, causing the delta-score to fail at distinguishing true and false identifications. Instead, the top 10 candidate peptides ranked by InsPecT were re-scored using MS-GF [60]. The highest scoring peptide was identified for each spectrum, and results were filtered to a 1% false discovery rate (FDR) [58]. A decoy database constructed from reversed sequences was used in the FDR calculation. We then removed peptides with more than one missed trypsin cleavage site. The resulting collection of peptide-spectrum matches had an estimated FDR of 0.4% with a peptide-level FDR of 3%.

Peptides were then labeled as either ‘known’, meaning they can be derived from an annotated protein sequence, or ‘novel’. Novel peptides do not match an annotated protein sequence. In a few cases, a novel peptide can be modified by a single amino acid change to create a known peptide. Several amino acid changes do not alter the mass of the peptide (e.g. Isoleucine and leucine or lysine and glutamine) or can be explained by small chemical modifications (e.g. deamidated asparagine and aspartic acid). If by making a single amino acid change, the novel peptide matches a known protein, it is removed from the set of novel peptides. By our definition, we found 24,782 novel peptides.

4.4.1 Validation of the Known Proteome

Peptides that match to annotated proteins confirm the translation of the protein. We distinguish between three ways in which proteins may be selected. A protein may be selected if it has at least one peptide that matches uniquely to the protein. In the case of protein isoforms from the same gene, there may be no peptide that is unique to each isoform. However, there may be a locus-specific peptide which maps to only proteins produced at that gene. In this case, we say that the locus, or gene, was selected. Finally, we group together proteins which are not produced at the same locus, but are indistinguishable given the identified peptides. Following the parsimony rule, proteins which match a subset of peptides matched to another protein are removed. Shared peptides that match to a previously selected protein are removed before identifying groups.

4.4.2 Defining Novel Events

Novel peptide sequences were then mapped to all locations in our putative protein databases. Both the six-frame translation database and the exon splice graph contain meta-information denoting the location on the genome from which each amino acid was derived. Mapping to the databases, instead of the genome directly, introduces a bias towards regions that have sufficiently large reading frames

Table 4.2: Event types and descriptions

Event Type	**Description**
Antisense Translation	The peptide cluster appears within 3kb of the annotated transcript, but is on the reverse strand.
Frame Change	The peptide cluster overlaps an annotated exon, but in a different frame of translation.
Exon Extension	The peptide cluster overlaps an annotated exon, but extends beyond the exon boundary.
Novel Exon	The peptide cluster falls with the intron region of an annotated transcript.
Novel Splice	The peptide sequences map to exons, but the indicated splice junction is not in the annotated transcript.
Translated UTR	The peptide sequences fall within the untranslated region (UTR) of the transcript.
Gene Extension	The peptide sequences fall within 3kb of the annotated transcript.

or have genomic signals indicating translation. We adopted this strategy to prevent the identification of false positive locations that would be created by allowing all possible splicing patterns. On average, a novel peptide matched to 3.67 locations in the genome. We found that novel peptides often co-located with other novel peptides. For gene annotation, the most important step is interpreting the peptide clusters to identify refinements to the current annotation. We created an automated method for interpreting and scoring the suggested refinements. We call these refinements, ‘novel events’.

Each novel event indicates an update to a specific annotated protein. The exception to this is novel genes, which are refining the annotation as a whole, not a specific gene. In our method, we define seven types of novel events besides novel genes. The event types and a brief description is provided in Table 4.2. The events are ordered by precedence so as to prevent a novel peptide cluster from being interpreted as two distinct types of events while not providing any new information.

Precedence is assigned to remove ambiguity in the identification of events. For example, a peptide may extend an exon and may also be in a discordant frame. In these cases, the ‘Frame Change’ event takes precedence. All events are also defined with reference to a particular transcript. A single gene may have several splice isoforms, and a novel exon event may be defined using the same set of novel peptides for each distinct splice isoform.

4.4.3 Scoring Novel Events

The quality of an event is determined based on the quality of the original peptide identification, the quantity of identifications to each peptide, and the quantity of peptides supporting the event. The score of an event, is computed using a bayesian framework. Consider the event, E, shown in Figure 4.13. We wish to compute the probability that E is correct, $Pr(E)$. Event E has two supporting peptide locations, and we assume that the identification of distinct peptide locations is independent. Therefore,

$$Pr(E) = 1 - \prod_{l \in Loc(E)} (1 - Pr(l))$$

where $Loc(E)$ is the set of peptide locations supporting E, and $Pr(l)$ is the probability of the location l being translated. Intuitively, this equation explains that the probability of an event being correct is the probability that at least one of the locations in the event is expressed. $Pr(l)$ is dependent on the confidence in the peptide sequence identification, and the number of locations the peptide matches in the genome. In 4.13, the blue peptide is uniquely located, only appearing at the location in the event E. However, the green peptide appears in two locations. Since either location is equally likely to be correct, we evenly distribute the confidence in the peptide sequence identification across the two locations. We compute the probability of a location l that matches peptide p being translated as

$$Pr(l) = \frac{1}{m_p} Pr(p)$$

where m_p is the number of genomic locations for peptide p. We chose to consider each location of a peptide equally, but the distribution of probabilities could be extended to take into account neighboring peptides or genomic signals. The probability of a peptide sequence identification being correct, $Pr(P)$, is in-

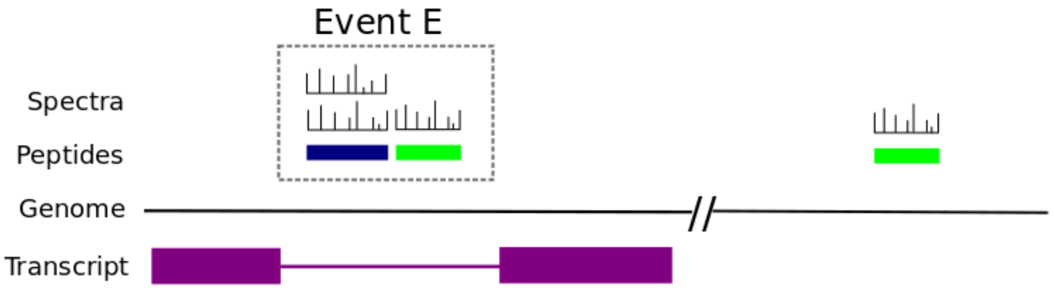


Figure 4.13: An event E is a 'Novel Exon' event and is supported by 2 peptides. One of the peptides also maps to the genome at a location distal to E. The unique peptide was identified by 2 spectra, while the degenerate peptide was identified by 1 spectra.

formed by the quality of the spectra identifying the peptide. Whereas we could treat the identification of distinct peptides as independent events, we cannot make that assumption of multiple spectra identifying the same peptide. Instead, we conservatively evaluate the probability of the peptide sequence being correct as the highest probability of a spectrum matching the peptide.

$$Pr(p) = max_{s \in Spec(p)} Pr(s, p)$$

where $Spec(p)$ are the set of spectra matching p and $Pr(s, p)$ is the probability of the spectrum s being generated from p. We estimate the $Pr(s, p)$ using the local false discovery rate (lFDR) [59] of the match. The lFDR is a measure of the rate of false discoveries among peptide-spectrum matches of similar quality. In our pipeline, we estimated match quality using MS-Generating Function [60], and computed the FDR in fixed-width bins with a minimum of 100 matches per bin. The lFDR of the match of p and s with MS-GeneratingFunction Spectral Probability q is simply the lFDR in the bin containing the score q. We then can compute the probability of a peptide-spectrum match being correct as

$$Pr(s, p) = 1 - lFDR(q)$$

We find that our definition of the probability of an event, or *eventProb*, has several desirable qualities. First, it eliminates the need to choose an arbitrary, and experiment dependent, cut-off for the minimum number of peptides required to trust an event. An event with a single high confidence peptide will receive a high eventProb. This makes the eventProb more robust to varying sizes of experiments. Secondly, it enables us to evaluate events which have no uniquely located peptides.

In order to choose appropriate eventProb cutoffs for the events, we evaluated the both its sensitivity and specificity. To measure sensitivity, we simulated the removal of proteins from the known proteome, and performed event finding anew with the peptides from the removed proteins now marked as novel. We repeated the trial 10 times, each time selecting 1000 proteins to remove. We only removed proteins for which we identified at least one uniquely mapping peptide, the same criteria used for event identification.

We measured specificity by inserting decoy peptide sequences into the set of novel peptides. At the estimated 3% peptide-level FDR, we expect approximately 7,000 false peptides to appear in our dataset. If the incorrect peptide identifications occurred at random among the known and novel identifications, then at the estimated 3% peptide-level FDR, we expect 743 novel peptides to be incorrect. However, since our search database contains more novel sequences than known protein sequences, we expect more incorrect peptides to occur among the set of novel peptides. To be conservative, we assume all of the incorrect peptides occur in the novel peptide set. To estimate our specificity in event identification, we created 7,394 decoy peptides by selecting random genomic locations, choosing a peptide length from the distribution of observed peptides, and translating the genomic location. To make the calculation more tractable, we restricted our decoy peptides to those appearing at no more than 500 locations in the genome. Each decoy peptide was assigned a lFDR sampled from the distribution of expected decoy PSMs at observed lFDRs. Using the combination of novel peptides, and our decoy novel peptides, we performed event identification. We labeled events composed solely of decoy peptides as decoy events, and computed an event-level FDR.

We found that the behavior of the eventProb depends on the type of event, and for our pipeline we chose several eventProb cutoffs which correspond to 90% sensitivity. The eventProb cutoff for Novel Gene events has the highest error rate (eventProb >= 0.95, sensitivity = 0.9, FDR = 8.75%) because the chance of finding co-located decoy peptides in a gene desert is relatively high. In contrast, Gene Extension events, Translated UTR events, and Antisense Translation events require co-located peptides which are proximal to a known gene. This is a more difficult scenario to find by chance, and therefore at the same sensitivity, the FDR is lower (eventProb >= 0.951, sensitivity = 0.9, FDR = 5.93%) . The most relaxed eventProb cutoff (eventProb >= 0.788, sensitivity = 0.9, FDR = 4.88%) is reserved for events with low probability of occurring by chance. These include events that require peptides to appear very close to one another; Novel Exon events, Frame Shift events, Novel Splice events, and Exon Boundary events. We allow a higher FDR among novel gene discoveries in order to achieve good sensitivity. We expect that many of the erroneous discoveries will be removed at the full gene model prediction stage.

4.4.4 Peptide degeneracy

In higher eukaryotes, we observe many peptides that match to multiple genomic locations, termed ‘peptide degeneracy’ in the mass spectrometry community. *Zea mays* has undergone 2 whole genome duplications, and over 80% of the genome is composed of transposable elements [17]. In our proteomic analysis of Maize, we found that 31.14% of novel peptides matched two or more genomic locations. In the identification of events, described above, we found 61,902 clusters of novel peptides containing only degenerate peptides. Since we only accepted events with at least one uniquely located peptide, these clusters were discarded from consideration as novel events.

In the set of proteins identified in the known proteome, we observe small groups of paralagous genes which share peptides. Among novel peptide clusters, we define sets of clusters which share many peptides as paralogous locus groups. To identify the groups, we construct a bipartite graph with one class of nodes representing clusters and the other class of nodes representing peptides. We add an edge between a peptide node and all of the locus nodes in which the peptide appears. To reduce the graph to only shared peptides, we remove all loci which have a unique peptide as well as all shared peptides from the removed loci. What remains is a collection of connected subgraphs, which we label as paralagous locus groups.

4.4.5 Gene Model Prediction

We generate full gene models based on the events that passed our eventProb filters. We used Augustus to produce the gene models due to its ability to accept external evidence. In addition to the peptide evidence, we also provided Augustus with evidence from transcriptomics (2 million maize cDNA sequences from NCBI RefSeq (2010.10.07) and 875 million RNA-Seq reads [128, 129]). We also included evidence from homology by mapping proteins from *Oryza sativa* (genome annotation version 7 containing 66,338 proteins) and *Sorghum bicolor* (genome annotation version 1.4 containing 34,496 proteins). Proteins and ESTs were aligned to the maize genome using BLAT [127] with default parameters. RNA-Seq reads were aligned using Bowtie version 0.12.7 with default parameters and TopHat version 1.4.1. We accepted only reads which mapped uniquely. For gene refinement events, we also include the currently annotated gene model from the 5a.59 protein set as a hint.

4.5 Acknowledgments

Chapter 4, in full, is in preparation for publication as "Novel gene finding in *Zea mays* using an automated proteogenomics pipeline". NE Castellana, Z Shen, Y He, J Walley, LG Smith, SP Briggs, and V Bafna. in preparation The dissertation author is the primary author of this paper. The dissertation author developed the computational methods, ran the computational analysis, and wrote the paper.

Chapter 5

Sequencing whole antibodies

5.1 Introduction

Database-search algorithms, such as Sequest[130], Mascot[5], and InsPecT[7], are the primary workhorses for the identification of tandem mass spectra. However, these methods are limited to the identification of spectra whose peptides are present in the database. It is well recognized that curated protein databases are, at best, an imperfect template for the extant peptides. For example, peptides arising from novel splice-forms or fusion proteins would be difficult to identify using most protein databases.

Recent developments have extended the identifications to peptides that have diverged from the database entry. By allowing divergence, the methods enable the identification of small-scale mutations, and post-translational modifications, albeit with some loss of sensitivity[95, 131, 97, 132]. Among these tools, MS-Blast is able to determine a homologous protein in the related species, but does not report the (diverged) protein in the target organism. The other tools consider variations, including modifications, and mutations, in reconstructing the target sequence. However, these tools will not work if the template (homologous peptide) is missing in the database, or comes from a novel splice-form. Additionally, these tools do not attempt to reconstruct the entire protein target sequence. *De novo* identification of peptide sequences[126, 133] is another possibility, and does not require a protein database. However, these methods are prone to error.

The issue of discovering spliced peptides (more generally, eukaryotic gene structures) has been investigated using a combination of approaches, loosely termed *proteogenomics*. Often, these approaches start by creating specialized databases of splice forms, combining evidence from protein (ex: NCBI nr[134]), and cDNA sequencing[135, 62, 136]. To discover novel splicing events, the tools also search databases derived directly from the genome such as a 6-frame translation or a compact encoding of multiple putative splicing events[47, 48, 16, 66]. For example, Castellana, *et al.* achieved this by constructing a database represented as a graph[16], containing many putative exons and exon splice junctions.

However, this approach also has its shortcomings. The putative gene models are constructed based on prior assumptions about splice junctions, and proximal exons. In addition, recent genomic discoveries point to extensive structural variation in the genome, in the form of large scale deletions, insertions, inversions, and translocations on the genome that might fuse different genic regions, or create non-standard splice-forms[137, 138]. Indeed, many cancers are characterized by such large scale mutations of the genome[19]. A second example of variation that confounds standard database identification techniques are immunoglobulins, or antibodies. Here, recombination events fuse disparate regions of the genome, often inserting non-templated sequence, and creating many novel gene structures in every individual. The common theme in all of the scenarios described is that it is not possible to maintain all possible encodings in a database to allow for a standard proteogenomic search.

In this paper, we ask if the imperfect template provided by the genome can be still used as a basis for peptide (and protein) identification. We are motivated in our approach by the work of Bandeira, *et al*[99], who could sequence monoclonal antibodies *de novo*, not making use of a database at all. In their method, an all-to-all comparison of spectra allows the creation of spectral contigs, similar to sequence contigs in shotgun sequencing projects. The sequences of the spectral contigs are determined *de novo*. Using full antibody sequences as references, they are able to order the contigs and infer the missing sequence. Because the construction and sequencing of the contigs is performed completely *de novo*, Bandeira, *et al* are able to sequence highly divergent proteins or proteins for which there is no database. However, the ordering of the sequenced contigs relies on a database of full antibody sequences for mapping. Sequences that cannot be mapped to an antibody in the database may be discarded. In contrast, the templates used in our method are not full proteins, but substrings of proteins, such as the V,J, and C regions (or exons), that are combinatorially chained together to best explain the spectrometric evidence.

Recently, Liu, *et al*[100] developed a method, *Champs*, for sequencing a divergent protein using a homologous protein database. In their method, a single reference protein is chosen and the *de novo* interpretation of spectra are mapped to the reference. They are able to sequence a protein with high accuracy using a reference protein with only 77% similarity to the target. While *Champs* is able to map peptides that differ from the reference by one or two amino acids, it does not look for large insertions or deletions in the target sequence, as in a novel splice form. In our work, use of the database as an incomplete template lends additional confidence to the target sequencing without substantially limiting the ability to identify diverged sequences.

Here, we describe a novel method for template proteogenomics, implemented in a tool, GenoMS. GenoMS takes as input a collection of spectra (acquired from multiple protease digests), and a collection of imperfect templates and constraints (defined in Experimental Procedures). It returns a target protein sequence. At the heart of the approach is a novel method of extending a target amino acid sequence, by recruiting and aligning spectra that match it partially. By using spectral data-sets with multiple protease digests, we are able to identify many overlapping peptides. We then align the overlapping spectra and produce an extended consensus spectrum. We are able to extend 89% of the target amino acid sequences. Over 40% of these extensions are 3 or more amino acids.

We test the performance of GenoMS in reconstructing monoclonal antibody sequences. Antibodies are a interesting test case owing to their highly variable nature, and no complete antibody database exists.. They are composed of four polypeptide chains; two identical heavy chains and two identical light chains

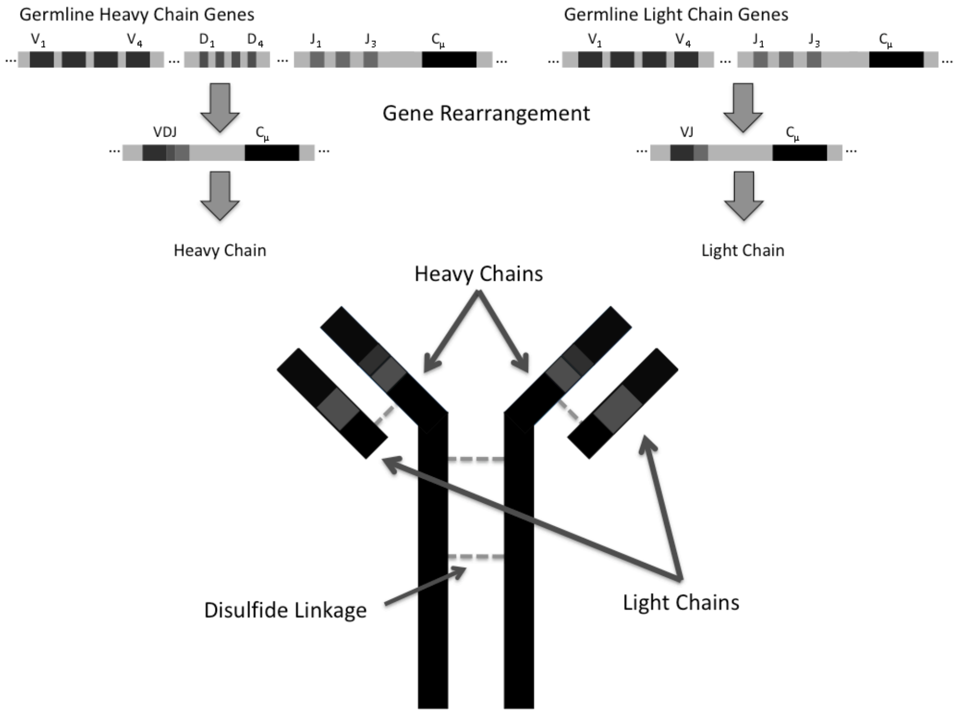


Figure 5.1: Bottom: The mature immunoglobulin protein structure contains two identical light chains and two identical heavy chains. The germline heavy chain and light chain loci (shown top) contain many different gene segments. During heavy chain gene rearrangement, in B cell differentiation, one V, one D, and one J segment gene segment are combined. For light chain gene formation, a V and a J gene are combined. The combined VDJ, or VJ, segment are joined by splice junction to a constant region.

(Figure 5.1). An antibody's preference and efficiency in the detection and removal of encountered antigens is heavily dependent on its amino acid sequence. Consequently, antibodies are extremely diverse. A principal way in which antibody diversity is achieved is through genome rearrangement of the germline locus (Figure 5.1). An antibody's heavy chain is comprised of 4 gene segments; a variable (V) segment, a diversity (D) segment, a joining (J) segment, and a constant (C) segment. Similarly, the light chain is composed of 3 gene segments; a V segment, a D segment, and a C segment. Each segment is chosen from potentially hundreds present in the genome, and many combinations of gene segments may be joined. Imprecise boundaries with the possible insertion of additional nucleotides allows the creation of many sequences from a single germline locus. Somatic hypermutation also plays a role in achieving antibody diversity. While antibody sequence may be determined by sequencing the DNA of the source cell line, few direct protein sequencing options exist when the source is unavailable or for ensuring antibody integrity. The antibody structure provides enough complexity to serve as a test case for template proteogenomics.

Using the technique of extending the peptide sequence without reference to a database, we are able to reconstruct the full protein sequence for the antibody raised against the B- and T-cell lymphocyte attenuator molecule (aBTLA)[99]. We also test our approach by using an available data-set of spectra acquired using multiple protease digests for bovine serum albumin (BSA). The sequence of BSA is determined using the the bovine genome as a template database. Both chains of aBTLA were sequenced using un-rearranged gene segments as templates. An independent reconstruction of the aBTLA heavy chain was performed using the un-rearranged heavy chain genomic locus as a template.

5.2 Methods

Our goal is to reconstruct the *target* amino acid sequence, using a chain of *templates*. Define a template to be an amino acid sequence that may be present in the target protein, though possibly in a mutated or modified form. The target protein might contain multiple templates chained together. We provide additional abstraction to model constraints on the templates. First, the user can specify a partial order $t_1 \rightarrow t_2$ to enforce that template t_1 must precede t_2 in the chain. Second, the user can provide mutual exclusion constraints on a pair of templates, (t_1, t_2), to enforce that only one of the two templates is in the chain. For example, in antibody sequences, all V, D, J, C genes are templates. The constraints help specify the ordering of V, D, J, C genes, and the exclusion of any pair of genes from the same class (Ex: V).

Define an *anchor* to be a substring of a template that is present in the target with no mutations. Each template may contain zero or more anchors. Figure 5.2 describes an overview of our algorithm. GenoMS takes a collection of MS^2 spectra as input, along with a set of templates and their constraints, requiring at least one anchor sequence. It outputs a target protein sequence, using a chain of templates as a guide. There are 3 stages: *template-chain selection*, *anchor extension*, and *sequence construction*, all described below.

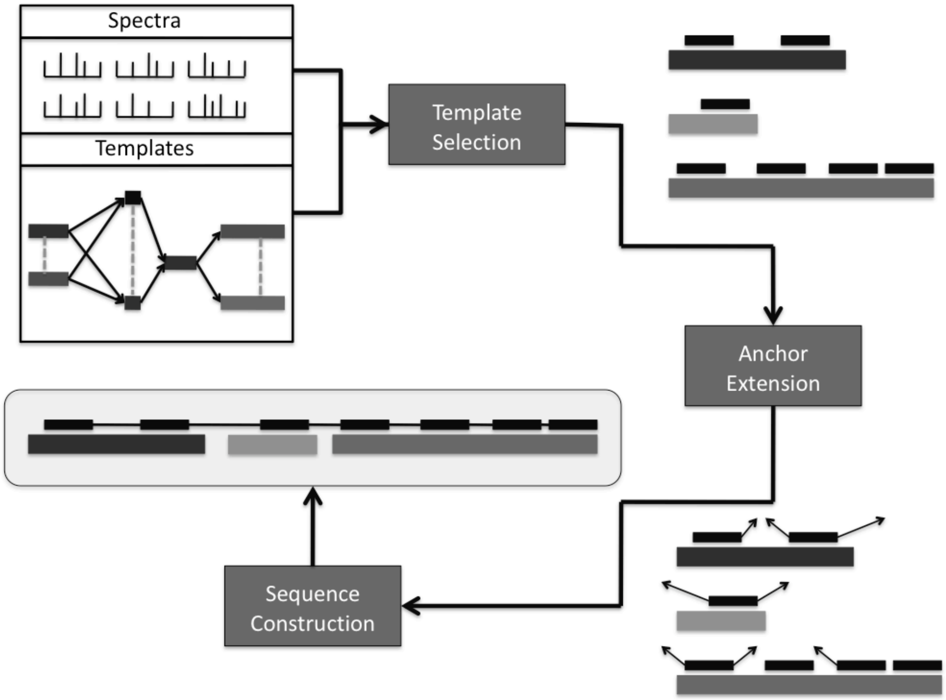


Figure 5.2: The template proteogenomic method reconstructs a target protein sequence using tandem mass spectra and a template database in three steps; Template-Chain Selection, Anchor Extension, and Sequence Construction. The template database specifies ordering and mutual-exclusion constraints between templates. A set of templates are selected that obey these constraints based on peptides identified on them. Anchors are peptides identified by searching spectra against the template database. Anchors are extended by aligning spectra that overlap the anchor. Finally, the sequence is reconstructed by merging the extended anchor sequences.

5.2.1 Template-chain selection

We create a custom database of all template sequences, and use the database search tool, InsPecT to search all spectra against the database[7]. The best templates to use as guide are the ones that show good match to the spectra. Define

Coverage$[t]$ as the number of amino-acids on t that were confirmed by the database search. Peptides that appear in multiple templates count towards the coverage of all of them. This reuse is eliminated in the next step. The goal of the *template-chain selection* phase is to select a chain of templates with maximum coverage, while satisfying all constraints.

To find the chain of templates, we define a graph in which the nodes are templates. There are two sets of edges. Directed edges $t_1 \rightarrow t_2$ model the ordering constraints, while a set of undirected edges, $(t_1, t_2) \in E_f$, model the exclusion. In addition to the constraints specified by the user, we also create forbidden edges between templates that share more than the minimum of 2 peptides or half of the peptides belonging to one of the templates. A chain $T = \{t_1, t_2, \ldots, t_k\}$ is *valid* if $(t_i, t_j) \notin E_f$ for all t_i, t_j in T, and $t_1 \rightarrow t_2 \rightarrow \ldots \rightarrow t_k$. The objective is to compute a valid chain so that $\sum_{i=1}^{k} \text{Coverage}[t_i]$ is maximized.

Solving this problem generally is hard. We use a heuristic method based on dynamic programming to find a valid chain. Let V_j denote the maximum score of a valid chain ending at t_j, and T_j denote the corresponding chain. Then,

$$V_j = \text{Coverage}[t_j] + \max_{i : T_i + \{t_j\} \text{is valid}} V_i$$

and T_j is constructed by chaining t_j to the optimal T_i. The template-chain determined by this heuristic is considered for subsequent stages of GenoMS. For an antibody, the template chain will often link V(D)JC together in that order. However, all templates are not required. Missing templates will be filled in by anchor extension. Second, we are not limited to a single chain. A variant of this heuristic can output multiple chains when needed (Ex: alternative splicing).

5.2.2 Anchor Identification and Extension

Recall that the template chain was created by connecting templates that were well-covered by target peptides. For each selected template in the chain, anchors are created by merging overlapping peptides. Anchors are ordered by their position on the chain. Spectra not annotated using the database search are reconsidered in the subsequent phases of the algorithm.

In the second step, we extend the sequence of each anchor. Prior to extension, all spectra are first clustered to reduce the overall number of spectra and improve spectrum quality[139]. The clustered spectra are converted to Prefix Residue Mass (PRM) spectra[126]. A PRM spectrum is represented by a list of mass values, and a PRM-score function ϕ that computes the likelihood of a mass value being a PRM. To extend the anchors, we perform the following at either end of each anchor:

```
procedure ExtendAnchor
  1. Recruit PRM spectra that overlap the N/C-terminal of the anchor.
  repeat
    1.1. Align recruited spectra.
    1.2. Construct a consensus spectrum from the aligned spectra.
    1.3. Recruit spectra that overlap the N/C-terminal of the consensus spectrum.
  while new spectra can be recruited.
  2. Sequence the consensus spectrum
```

Recruiting PRM Spectra

All spectra that do not contribute to an anchor and have not already been recruited are examined for overlap with each anchor. Any spectra that are recruited in previous rounds to the same terminus of the anchor are eligible for recruitment in subsequent rounds of recruitment for the terminus as well. We determine overlap by using a modified spectral alignment method[140]. When aligning a spectrum to an anchor, we allow the spectrum to only partially overlap the anchor (Figure 5.3). Since the extended target sequence is determined by aligning the recruited spectra, it is critical to reduce false-positive recruitment, while maintaining enough coverage to reliably extend the sequence.

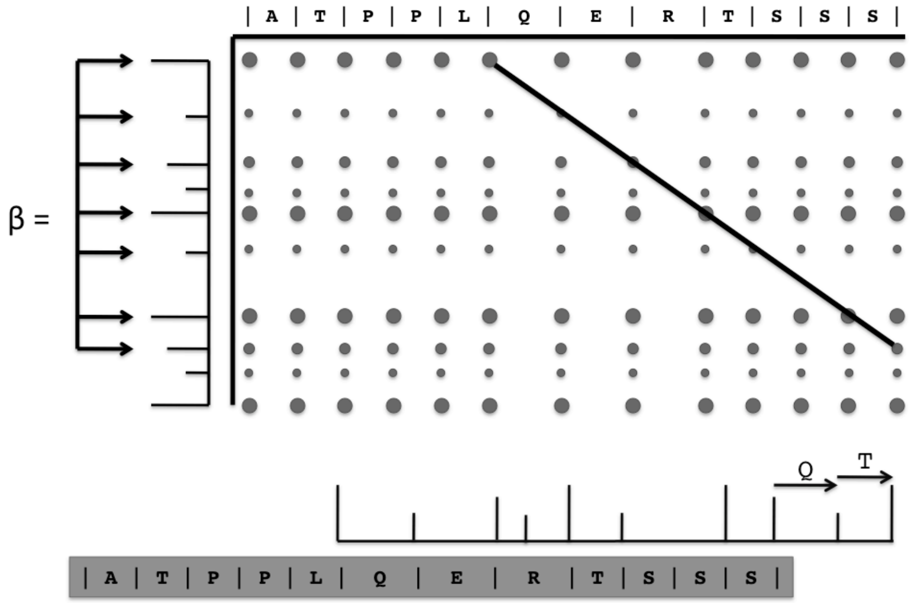

| A | T | P | P | L | Q | E | R | T | S | S | S |
β =
Q
T
| A | T | P | P | L | Q | E | R | T | S | S | S |

Figure 5.3: Top: A PRM spectrum is shown with a partial alignment to the theoretical PRM spectrum of an anchor. The C-terminal of the spectrum is not aligned. Bottom: The overhanging peaks enable the extension of the anchor sequence by two amino acids (QT).

We consider 3 parameters: the minimum additive score of the spectral alignment, Q[140]; the minimum number of overlapping peaks, β, for a spectral alignment to be considered, and the exact number of spectra recruited, N_S. Q could be learned by the algorithm independently for each experiment by looking at the alignment score of spectra identified by InsPecT. We tested for the dependence on β and N_S using a training set of 206 uniformly selected anchor ends from the aBTLA heavy chain sequence. Values $\beta = 4, N_S = 5$ were chosen to balance the accuracy (fraction of recruited spectra that are correct), and sensitivity (fraction of true spectra recruited).

The recruited spectra and the anchor sequence must then be aligned. The sequence helps to anchor the spectral alignment, and the spectral alignment is then used to produce a consensus extension of the sequence. We do this using Hidden Markov Models (HMMs).

Multiple spectrum alignment

Profile hidden markov models (HMMs) are a popular tool for multiple sequence alignment[141]. We alter the scheme slightly in order to perform multiple spectrum alignment. The use of HMMs for scoring peptide-spectrum alignments has previously been proposed[142]. A novel part of our approach is that the HMM is not static, but is updated by model surgery, as we extend the anchor sequence.

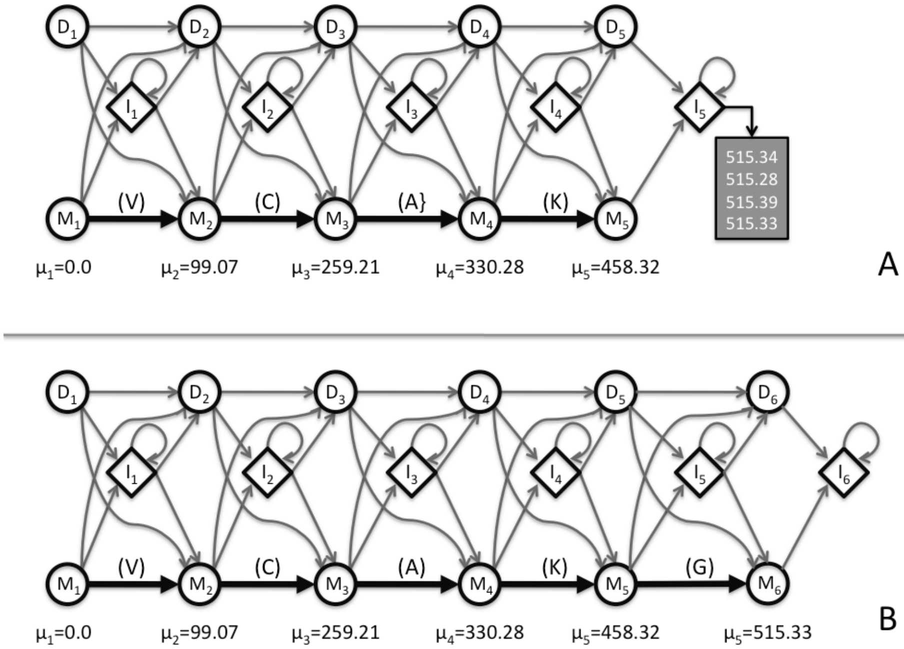


Figure 5.4: A: The spectrum profile HMM derived from the anchor 'VCAK' after aligning 4 spectra, all of which are aligned to the state I_5. The peaks aligned to that state are shown, and suggest a candidate match state at mass 515.33 Da. B: The same HMM, after we perform model surgery to add the new match state.

Recall that the anchor sequence can also be interpreted as a list of PRMs $[m_1, m_2, m_3, \ldots]$. For example, the anchor VCAK corresponds to the PRM list

$$[0, 99.07, 259.21, 330.28, 458.32]$$

Intuitively, the HMM is an automaton that generates these PRMs. See Figure 5.4A. In the absence of noise, we have a set of *Match* states $(M_1, M_2, \ldots)$.

The automaton starts in Match state M_1. In each Match state M_i, the PRM m_i is emitted, followed by a transition to the next Match state. Formally, an HMM is described by a 5-tuple $M = (\Omega, A, B, \pi, \Sigma)$, where Ω is the set of states. The HMM is initially in state $\omega_i \in \Omega$ according to the distribution π. In state ω_i, M emits a symbol $o \in \Sigma$ according to the distribution $B_{i,o}$, and transitions to state ω_j, according to the transition probability $A_{i,j}$. To model measurement errors, the Match state M_j outputs a mass m according to

$$B_{M_j,m} \sim N(m_j, \sigma)$$

where s.d. σ is obtained by empirically measured instrument accuracy. Noise peaks are modeled by *Insert* states in between each adjacent pair of Match states, with the emission probabilities defined by

$$B_{I_j,m} \propto \begin{cases} e^{-\phi(m)} & \text{if } m_j < m < m_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

Missing peaks in spectra are modeled by moving from a Match state to a *Delete* state, where no symbol is emitted. The transition probabilities $A_{i,j}$ are initialized to favor Match transitions, and penalize Delete transitions. All parameters are updated at each iteration using a Bayesian approach described in the next section.

In this generative model, each spectrum is produced by traversing a (hidden) path through the states of the HMM. Reconstructing the most likely path is equivalent to aligning the spectrum to the HMM, and can be determined using the Viterbi algorithm[143]. An Insert states is created after the final Match state for C-terminal extension, or before the initial Match state for N-terminal extension. Model surgery is performed to generate additional Match states from these terminal Insert states, which are used to reconstruct the template extension. The complete procedure is described below.

procedure AlignSpectrum

1. Create an initial HMM using the anchor.

2. For each recruited spectrum, S
   - 2.1. Align S to the model using the Viterbi algorithm.
   - 2.2. Update model parameters.
   - 2.3. Perform model surgery.

Updating model parameters: Transitions $A_{i,j}$ are updated according to

$$\rho_i \leftarrow \frac{1}{\sum_k c_{i,k} + 1} \tag{5.1}$$

$$A_{i,j} \leftarrow \frac{c_{i,j} + \alpha_j + \rho_i A_{i,j}}{\sum_k [c_{i,k} + \alpha_k] + \rho_i} \tag{5.2}$$

where $c_{i,j}$ is the number of aligned spectra with transition from ω_i to ω_j and ρ_i is the 'learning rate'. Low values of ρ favor the observed transitions, while high ρ favor the current transition probability. α_j is the pseudocount for ω_j, described empirically by

$$\alpha_j = \begin{cases} 7 & \text{if } \omega_j \text{ is a Match state} \\ 1 & \text{otherwise} \end{cases}$$

To update $B_{M_j,m}$, the mean is recomputed in each step, using spectral PRMs that were emitted in state M_j. The variance remains unchanged.

Model Surgery: The initial HMM is constructed using the anchor PRMs. The aligned spectra overlap only partially. The PRMs preceding the N-terminal Match state (or succeeding the C-terminal Match state in the case of right extension) are emitted by Insert states. The observed masses emitted by an Insert state cluster around certain PRM values, specifically at the preceding (or succeeding) PRMs of the target sequence. Model surgery is used create a Match state that can emit the cluster of PRMs (See Figure 5.4B). In this way, the HMM is extended to better represent the target sequence.

Let M_I denote the set of mass values emitted by insert state I. Consider a subset $M' \subseteq M_I$. Let $\mu_{M'}, \sigma_{M'}$ denote the mean of the values in M', and the standard deviation, respectively. Define

$$\text{Score}(M') = \sum_{m \in M'} \phi(m)$$

We compute

$$M^* = \arg \max_{\substack{M' \subseteq W_I \\ M\text{'} \geq 2 \\ \sigma_{M'} < 0.25}} \text{Score}(M') \tag{5.3}$$

Note that the computation can be done efficiently by sorting the mass values, and looking at intervals.

If Score(M^*) exceeds the minimum PRM score $\phi(m)$ for any spectrum, we add a new Match state with mean μ_{M^*}, along with the corresponding Delete states and Insert states (Figure 5.4B). All spectra are realigned to the new HMM.

Building a consensus spectrum and extending the anchor sequence

The HMM, once learned from the recruited spectra, is used to produce a consensus spectrum. The consensus PRM spectrum is produced by finding the maximum likelihood path constrained to those paths that begin at the initial Match or Delete state and end at the final Match or Delete state. The peak emissions of this path, omitting noise peaks emitted from Insert states, produce the consensus spectrum. Each peak in the consensus spectrum is associated with a peak score. The PRM score for the mass emitted from state M_i is

$$\Sigma_{(w_i,\phi_i) \in W_{M_i}} \phi_i - \lambda W\_\text{–}D\_i'' \tag{5.4}$$

where W_{M_i} is the set of peaks aligned to state M_i and W_{D_i} is the set of peaks aligned to state D_i. λ is a constant. The peak scores are likelihoods, and λ is the likelihood of a true peak not being observed. We chose λ to be the average score of a PRM in the dataset. The consensus spectrum is then used as the anchor for subsequent rounds of extension. The sequence of the final consensus spectrum, once no more spectra can be recruited, is determined *de novo* by constructing a spectrum graph allowing edges for single and double amino acid masses[144]. The sequence is then recovered from the highest scoring path in the spectrum graph.

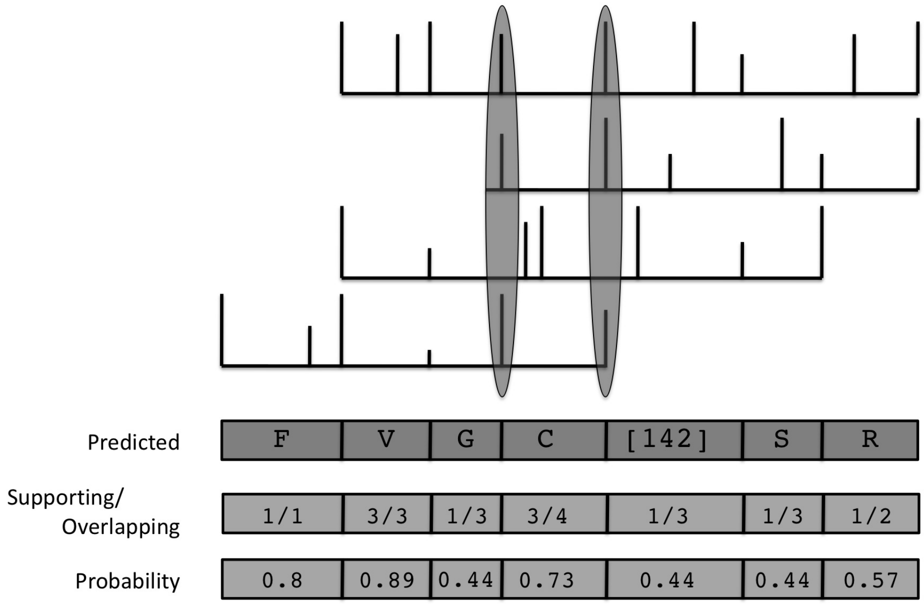
5.2.3 Protein Sequence Reconstruction

Once anchors have been extended until no spectra can be recruited, the extended anchor sequences are merged into a single protein sequence. If confident overlap between extended anchor sequences exists, then the sequences are merged. Each extended anchor sequence is considered for merging with all eligible anchors, that is the C-terminal extension of an anchor can only be merged with the N-terminal of another anchor. Similarly, if template ordering is provided, anchors can only be merged in accordance with the template order constraints.

5.2.4 Confidence estimates

The output of GenoMS is a sequence of mass intervals, some of which are represented by amino-acids in the final sequence while others which can be explained by two AAs are presented as masses. Each interval is inferred from a pair of adjacent PRMs. All spectra that were used to create an anchor or extend an anchor can be mapped to the sequence of intervals. We compute the *confidence* of a mass-interval as the fraction of overlapping spectra with PRMs mapping to the two adjacent PRMs. Figure 5.5 illustrates the computation of the site-wise confidence from the set of mapped spectra.

For spectra which are identified using InsPecT and are used for anchor creation, all mass intervals that are overlapped are also considered supported. In some cases, the confidence estimate is conservative as many spectra may support a larger mass interval that contains the correct one. For example, the first spectrum in Figure 5 supports the large interval ‘SR’, but is not counted towards either mass interval since it is missing the PRM between ‘S’ and ‘R’. We use pseudo-counts 1 and 1.5 for the number of supporting and overlapping spectra, respectively.



Predicted	F	V	G	C	[142]	S	R
Supporting/ Overlapping	1/1	3/3	1/3	3/4	1/3	1/3	1/2
Probability	0.8	0.89	0.44	0.73	0.44	0.44	0.57

Figure 5.5: A set of spectra is shown overlapping a region of the predicted sequence. A spectrum supports a mass interval in the predicted sequence if both adjacent PRMs to the interval are matched in the spectrum. The confidence of each mass interval is the fraction of overlapping spectra that support the interval (with pseudocounts). The PRMs of the overlapping spectra which are necessary to support the mass interval corresponding to 'C' are circled.

5.2.5 Genomic Templates

Immunoglobulins are an excellent candidate for template proteogenomics, with templates selected from translated germline segments. However, for other applications of protein sequencing, such as gene annotation, a protein template database may be difficult to produce. To handle these situations, GenoMS also accepts genomic sequence as input. It automatically generates templates and constraints as follows: the template database is a 6-frame translation of a gene locus, with each open reading frame (ORF) describing a template. Templates that overlap or are on different strands are mutually exclusive. Templates are ordered according to their genomic coordinates. Once the template database and constraints are produced, the same algorithm is used to reconstruct the target. For flexibility, we do not consider splice-junction signals in selecting template boundaries. However, users have the option to input customized template and constraint files.

The output of the GenoMS from genomic templates is the sequence of the target protein, as well as the genomic coordinates of the exons or gene segments selected as templates. In this way, the precise exon and splice boundaries for a gene may be discovered. Because template proteogenomics does not require the template genome to be an exact match for the target protein, it is possible to sequence the target protein of one species using the genomic template database derived from the genome of a related, and more comprehensively studied species.

5.2.6 Constructing a Divergent Sequence Database

In order to assess the ability of GenoMS to sequence more divergent proteins, we alter the template database to contain sequences with less similarity to the target. We construct the divergence sequence database from the known target protein sequence appended with the Mouse IPI database (v3.54) that contains 56,551 proteins once immunoglobulin sequences are removed. We simulated degrees of mutation by replacing regions of the target sequence with nonsense amino acids, "XXXX". The regions were selected at random positions on the heavy chain, with a normal length distribution with mean 7, standard deviation 2, and minimum length of 4 amino acids similar to the hypervariable complementarity defining regions (CDRs)[145].

5.2.7 Mass spectrometry analysis

Spectral datasets derived from aBTLA described in Bandeira, *et al*[99] were used for evaluating anchor extension and for full antibody sequence reconstruction. The dataset consisted of 44,985 MS/MS from the heavy chain and 39,135 MS/MS from the light chain acquired on either an LTQ-Orbitrap or LTQ-FTMS instrument. Heavy chain samples were prepared using four different protease digestions (trypsin, chymotrypsin, pepsin, and AspN) while light chain samples were prepared with three different proteases (trypsin, chymotrypsin, and AspN). 5,154 MS/MS spectra acquired on an Orbitrap instrument from three digestion conditions using GluC, LysC, and trypsin[100] were used to determine the gene structure of BSA.

All spectra were first clustered to reduce the overall number of spectra and improve spectrum quality[139] and converted to Prefix Residue Mass (PRM) spectra[126].

5.3 Results

5.3.1 Anchor Extension

Anchor extension is the mainstay of our algorithm. We measured the accuracy and length of GenoMS's extension of arbitrary anchors. Certain discrepancies between the target sequence and the predicted sequence were not considered errors, such as substitutions of amino acids with similar mass or mass shifts owing to common post-translational modifications.

From the known sequence of the aBTLA heavy chain, we selected every possible 10 amino acid sequence as an anchor. Q was fixed at 70, which is approximately the score cutoff chosen for the heavy chain. We then performed one round of recruitment, alignment, and reconstruction, as described above. Of the 413 anchors, 89% were extended. The average extension length was 2.56 amino acids. 41% of the extendable anchors were extended by 3 or more amino acids, while 12% were extended by 5 or more amino acids. Across all extensions, 95% of the amino acids were correctly predicted. Errors generally occurred in regions with one or more prolines, which hinders peptide fragmentation.

The accuracy depends greatly on the position in the extension. Figure 5.6 shows the accuracy as a function of the position from the *tail* of the extension. To explain, consider an anchor that is extended by two amino acids, g_1g_2, and another that is extended by three amino acids, $h_1h_2h_3$. The accuracy at position -1, the last position of the extension, considers the accuracy (fraction of residues predicted accurately) of g_2 and h_3. The accuracy at position -2 is determined by the accuracy of g_1 and h_2, while the accuracy at position -3 is only determined by the accuracy of h_1. As expected, the number of predictions decreases from

−1 onwards, while the accuracy increases. The length of extension depends upon the availability of overlapping peptide spectra, which in turn depends upon the protease mixtures. Our results indicate that with a large number of overlapping peptides, the extensions are accurate.

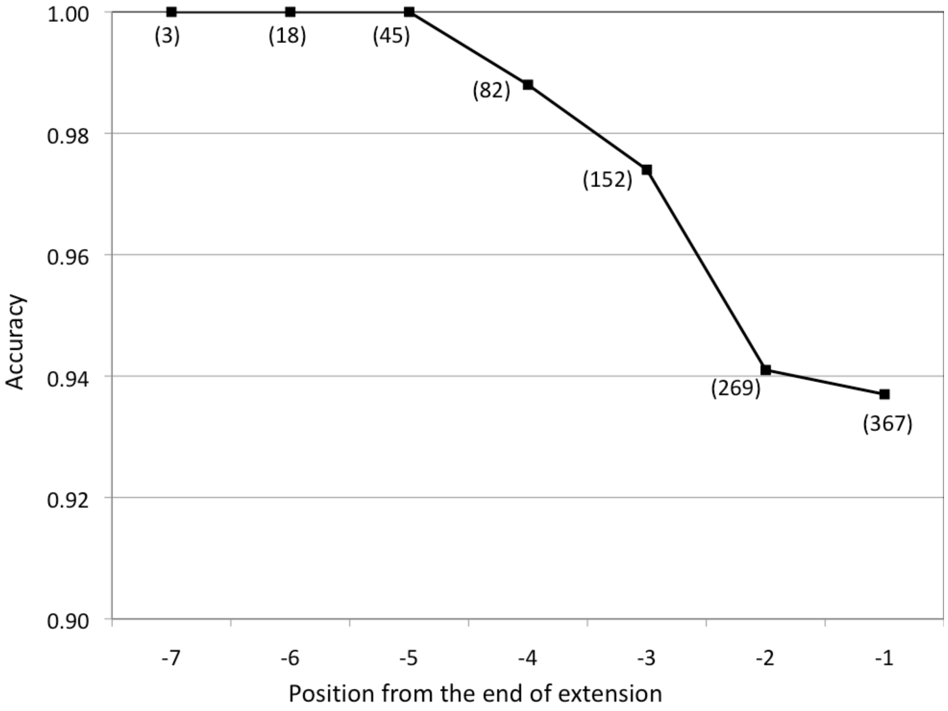


Figure 5.6: The average accuracy of each position in the extension. The accuracy of the extension degrades for positions close to the end of the extension, while the number of predictions increases. Each data point is annotated with the total number of anchors extended to that position or further.

Complete Protein Sequence Reconstruction The International Immunogenetics Information System(IMGT) GENE-DB[146] contains immunoglobulin genes observed in human, mouse, rat, and rabbit. We used the mouse genes in GENE-DB as templates for full protein sequencing. These templates contain sequences that are highly similar, but not identical to the specific aBTLA antibody used to acquire spectra. The heavy chain sequence and light chain sequence of the target, determined previously by Edman sequencing, are 443 and 221 AA in length, respectively. We tested if GenoMS could reconstruct the aBTLA targets using the MS^2 spectra, and the GENE-DB templates.

We constructed a database, IgH-DB, containing all mouse immunoglobulin heavy chain genes in GENE-DB (v.20090331), and a database, IgLK-DB, containing all mouse immunoglobulin light chain genes in GENE-DB (v.20090320). IgLK-DB contained both Lambda and Kappa light chain genes. Each V, D, J, and C segment was a template, and constraints were created according to two rules. Templates of the same type (e.g. V segments) were mutually exclusive, and the templates were ordered so that all V segments preceded D segments, D segments preceded J segments, and J segments preceded C segments. IgH-DB contained 479 templates and IgLK-DB contained 177 templates.

Figure 5.7A contains the results of full protein sequencing for the heavy and the light chains. The grey boxes correspond to anchors, annotated by the GI number and position in the sequence. Arrows extending and linking anchors in Figure 5.7 are annotated with the sequence that was determined by anchor-extension and sequence-reconstruction. A red sequence indicates error in extension. If the arrow is continuous from one anchor to the next, then there was sufficient overlap in the extensions to allow merging of anchor sequences. Mass gaps in the consensus spectrum that could not be resolved to a single amino acid are indicated with brackets ([XX]). If the mass gap correctly identifies a pair of amino acids from the sequence, the mass in brackets is replaced by the amino acids symbols in brackets.

Nearly all of the heavy chain sequence was recovered (99%) with 99% accuracy. One pair of anchors had sufficient overlap to be merged. The C-terminal extension of the third anchor and the N-terminal extension of the fourth anchor also overlap and could be merged by eye but not by our conservative merging criteria. The chain consisted of a template from each of *V* (gene IGHV2-3, GenBank Accession Number[1]), *J* (IGHJ4*01, GenBank Accession Number[2]), and *C* (IGHCG1*02, GenBank Accession Number[3]) segments. No database sequence matched the D gene, but the automated extension could reconstruct much of it. The only error in the D gene sequence, which appears in the C-terminal extension

[1] AC090887

[2] V00770

[3] L35252

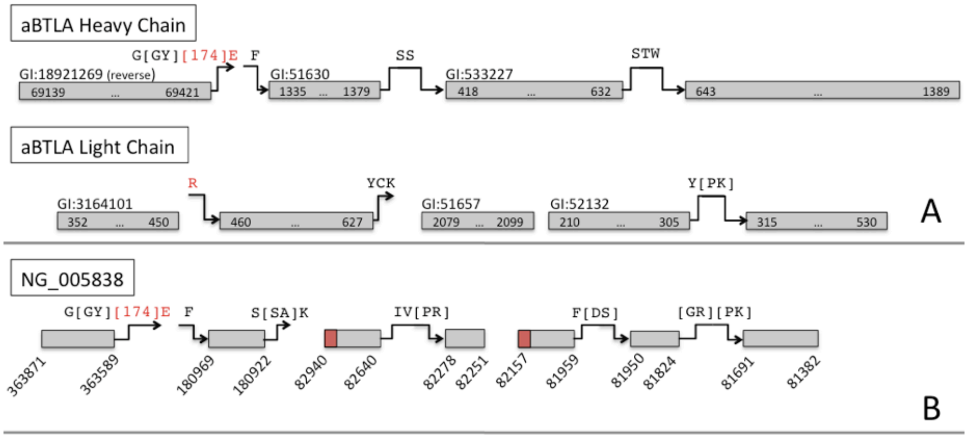


Figure 5.7: A: The aBTLA heavy chain and light chains reconstructed from protein template databases. The grey rectangles are anchors while the arrows, annotated with sequence, are the extended and merged sequence. Text above the anchors indicates the GI number of the template used and coordinates within or below the anchors indicate their position within the template. Red amino acids were incorrectly predicted. B: The aBTLA heavy chain identified using a genomic template database. The anchors were identified using templates from the locus reverse strand. Anchor ordering and genomic position is annotated with reference to the forward strand. The coordinates of each anchor on the chromosome are shown. Red portions of the anchors are incorrectly incorporated anchor sequence. C: The heavy chain sequence produced by using increasingly divergent templates. The reconstructions at 85%, 75%, and 65% similarity to the aBTLA heavy chain sequence are shown.

of the first anchor, has one incorrect PRM. '[174]E' has the same mass as the correct sequence 'RF'. The incorrect intervals receive a lower site-level confidence (0.62 and 0.73, respectively) than the rest of the sequence. Missing sequence occurred at the N-terminus (3 AA), owing to modification of the leading glutamine to glutamic acid.

Light chain templates were chosen in the same manner as for the heavy chain. The V, J, and C templates correspond to genes IgKV8-21*01(GenBank Accession Number[4]), IgKJ4*02(GenBank Accession Number[5]), and IgKC*01(GenBan k Accession Number[6]) (Figure 5.7A). Sequencing construction determined 96% of the sequence with 99% accuracy.

[4] Y15982

[5] V00777

[6] V00807

There was little gap between any of the anchors. A gap of two amino acids between the end of the first anchor and the start of the second anchor could be filled in correctly by inferring the sequence from the template. The N-terminal extension of the first anchor 'NN' has the same mass as the correct sequence 'DI', but the internal PRM is off by one Da. These two incorrect amino acids get very low site-level confidence (0.08 and 0.16, respectively). Five amino acids are missing between the end of the extension of the second anchor and the start of the third anchor. This gap corresponds to the joining boundary of the V and J gene segments. N-terminal extension of the third anchor is prevented by the incorrect incorporation of 'L' during anchor construction as the first AA in the anchor. The third and fourth anchors are directly abutting, with no missed amino acids.

aBTLA Heavy Chain: Genomic templates

We tested if the target could be reconstructed in the absence of protein templates. The mouse heavy chain genomic locus, GenBank Accession Number[7], was used to construct a template database, as described in Experimental Methods (Genomic Templates). The database contained 87,265 templates. Spectrum identifications were filtered to a 1% false discovery rate. Figure 5.7B shows the reconstruction of the heavy chain. We identified seven anchors comprising 95% of the target sequence. Each anchor was identified from a different ORF template on the reverse strand except for the 5th and 6th anchors which were from the same ORF. Each ORF revealed an exon in the final rearranged immunoglobulin heavy chain gene. Gaps between anchors, which are sequenced via anchor-extension, determine the exact splice boundaries between exons.

The predicted target antibody sequence contained 443 AA, with 98% accuracy. Two pairs of anchors were able to be merged. The first anchor of the sequence was identical to the first anchor identified against the IMGT database and consequently the error at the C-terminal extension of the anchor is also the same. The boundaries of two anchors were mis-identified by InsPecT. Four and three amino acids were incorporated into the third and fifth anchors, respectively,

[7] NG005838

and are denoted by the red shaded portions of those anchors. The summed masses of the incorrect amino acids were no more than one Da different from the true sequence, but they prevented the N-terminal extension of both anchors. The second anchor's C-terminal extension overlaps the correct portion of the third anchor, but the incorrect anchor boundary prevents merging.

5.3.2 Protein Sequence Reconstruction with Template Divergence

The template databases used contained sequences that were highly similar to the target sequence, presenting us with an easier test case. We tested accuracy by comparing against a diverged template database. Various levels of divergence (based on similarity to the original template) were created by introducing nonsense mutations in the template database (See Experimental Procedures). At each level of similarity, 20 independent diverged database results were averaged in Table 5.1. Q varied between experiments with mean 68.9.

Table 5.1: 'Sequence Similarity' is the identity of the target sequence to the closest mutated sequence in the database. '#Peptides', and '#Anchors' refers to the number of unique peptides, and anchors, respectively, identified on the mutated templates. 'Target Length' refers to the length of the reconstructed sequence, while 'Anchor Sequence' refers to the fraction recovered from the anchors. 'Target Accuracy' is the percentage of amino acids predicted correctly. While the anchor sequence drops rapidly, a significant fraction of the target is reconstructed accurately.

Sequence Similarity	#Peptides	# Anchors	Anchor Sequence	Target Length	Target Accuracy
95%	527	4	90.8%	429	99.1%
90%	443	6	83.6%	406	98.3%
85%	364	8	77.1%	402	97.6%
80%	286	9	68.2%	375	96.9%
75%	245	10	62.8%	368	95.1%
70%	201	10	56.1%	337	96.1%
65%	181	10	52.2%	324	95.3%

As the similarity of the database to the target decreases, more of the sequence was determined by automated *de novo* extension (Table 5.1). Three reconstructions with 85%, 75%, and 65% database similarity to the target sequence are shown in Figure 5.7C. Table 5.1 demonstrates how the number of peptides decreases with greater target sequence divergence. The disjointness of the peptides is indicated by the increasing number of anchors. Though the accuracy is diminished as the target sequence becomes more divergent, it is never below 95%. As the amount of sequence recovered in anchors decreases ('Anchor Sequence' in Table 5.1), the portion of the sequence recovered by extension increases. In Figure 5.7C, the longest extension in the case of 85% similarity is 14 AA, and most of the anchor extensions could be merged. Once the similarity drops to 65%, the longest extension is 22 AA, fewer extensions could be merged.

5.3.3 Gene Annotation

Recently, Liu, et al [100] published an algorithm for sequencing a diverged protein using a homologous protein. The target sequence used was BSA, and purified spectra were derived from three protease digestions. The complete BSA protein consists of 608 AAs, however the first 25 residues are cleaved as a signal peptide. Therefore, we only consider the 583 AAs following the cleavage site. Liu, et al is able to sequence the target protein with 100% accuracy and over 99% coverage by using a close homolog in sheep (¿90% similarity). In contrast, our method does not require a protein reference sequence, but can build from the genome directly.

In addition to sequencing the target protein, identifying templates from a genomic template database gives the positions of exons in the gene whose product is the target protein. We sequenced BSA using a database constructed from the bovine genome. We created a genomic template database from the 6-frame translation of the BSA locus (GenBank Accession Number[8]) containing 559 templates. From the genomic template database, we identified 91% of the sequence with 98% accuracy. We recover ORFs for 12 exons (Figure 5.8). The lack of overlapping spectra prevented the merging of all but one pair of anchors, however, some

[8] NC_007304.3:91,461,065-91,479,638

splice junctions could still be determined. For example, the N-terminal extensions ‘PCTEDY’ and ‘[LGE]YGFQNA’ span the splice junctions between the tenth and eleventh exons, and the ninth and tenth exons, respectively. This allows us to determine boundaries of the splice junctions and infer the missing exon sequence from the template.

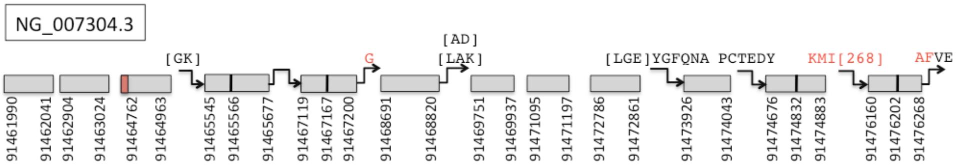


Figure 5.8: The annotation of the BSA gene using a genomic template database. Twelve exons for the gene are shown, with corresponding extensions. Each anchor is annotated with its genomic coordinates.

5.4 Discussion

Since the first sequencing of the human genome in 2001[147], we have witnessed an explosion in the number of species with partial and fully sequenced genomes. Gene and proteome annotation, however, have not been able to keep pace. Mass spectrometry and advancing computational tools, as a complement to cDNA sequencing, have been shown to greatly improve the accuracy and efficiency of the annotation process[47, 87, 48]. At their core, these methods rely on the assumption that the genome is an adequate database for the identification of peptides. It is nearly impossible to create a database that encodes all possible gene splice variants as well as small and large scale genome rearrangements. One alternative is to use *de novo* methods[148, 126] for peptide and protein sequencing. These algorithms make no such assumptions, but are plagued with low accuracy.

We have presented a novel method for protein sequencing that draws from the strengths of both the database- and *de novo* -based approaches. Template proteogenomics improves upon prior proteogenomic efforts by eliminating the need for custom databases that anticipate splice junctions and mutations[16]. Our method makes use of the genome as an imperfect template, while employing *de novo* techniques to sequence the divergent portions of the protein. By utilizing available sequence information, we are able to increase confidence in the final sequence, while not relying on the existence of a complete and accurate database.

Antibodies are highly diverse proteins that have confounded past attempts to construct a complete sequence database. We are able to use known antibody gene segments as templates to sequence proteins with up to 35% sequence divergence from the templates. The utility of the template proteogenomic method for gene annotation has also been demonstrated. From the final protein sequence, we were able to determine many exon boundaries and splice-junctions by constructing a template database from the 6-frame translation of the aBTLA heavy chain locus.

The alignment of overlapping spectra derived from a mixture of proteases lends additional confidence to full protein sequence. However, the errant portions of the alignment provide useful information as well, yet are often ignored. Post-translational modifications may be identified by observing both modified and unmodified spectra aligned. Complex protein mixtures containing both modified and unmodified spectra, or spectra from alternatively spliced peptides may be lost when an alignment is reduced to a consensus spectrum. In these scenarios, the correct output of the template proteogenomic method would be multiple sequences, not a single protein. In future work, template proteogenomics will be extended in order to capture post-translational modifications and sequence higher complexity samples.

5.5 Extensions to GenoMS

The description of GenoMS above requires the construction of a single, directed, acyclic graph consisting of a single component (representing either the heavy chain or light chain of any antibody). In the application of GenoMS to the sequencing of the anti-LT-a antibody [149] required the simultaneous sequencing of the heavy and light chains. We extended the algorithm to act on two connected components, one representing the heavy chain gene segments and the other representing the light chain gene segments. Therefore, the extended GenoMS is able to sequence two antibody chains despite the increased complexity of the mixture of spectra.

A second extension was required to sequence the anti-LT-a antibody. While aBTLA differs from the database in the CDRs, the anti-LT-a antibody had point mutations spread throughout the sequence, preventing the identification of anchors. We modified the template-chain selection method to perform a mutation-tolerant database search. By allowing each identified peptide in an anchor to differ by an amino acid from the peptide in the database, we improved the sequencing coverage of the antibody.

5.6 Acknowledgements

This chapter, in part, was published as “Template Proteogenomics: sequencing whole proteins using an imperfect database”. NE Castellana, V Pham, D Arnott, JR Lill, and V Bafna. *Molecular & Cellular Proteomics*, vol. 9, no. 6, pp. 1260-1270, 2010. The dissertation author was the primary author of this paper. The dissertation author developed the algorithm, implemented the algorithm, ran the experiments, and wrote the paper.

Bibliography

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198–207, Mar 2003.

[2] P. Mallick and B. Kuster, "Proteomics: a pragmatic perspective," *Nat. Biotechnol.*, vol. 28, pp. 695–709, Jul 2010.

[3] X. Liu, Y. Sirotkin, Y. Shen, G. Anderson, Y. S. Tsai, Y. S. Ting, D. R. Goodlett, R. D. Smith, V. Bafna, and P. A. Pevzner, "Protein identification using top-down spectra," *Mol. Cell Proteomics*, Oct 2011.

[4] J. Eng, A. McCormack, and J. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, vol. 5, pp. 976–989, 1994.

[5] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551–3567, Dec 1999.

[6] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, pp. 1466–1467, Jun 2004.

[7] S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna, "InsPecT: identification of posttranslationally modified peptides from tandem mass spectra," *Anal. Chem.*, vol. 77, pp. 4626–4639, Jul 2005.

[8] M. R. Brent, "Steady progress and recent breakthroughs in the accuracy of automated genome annotation," *Nat. Rev. Genet.*, vol. 9, pp. 62–73, Jan 2008.

[9] V. M. Markowitz, I. M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides, "IMG: the Integrated Microbial Genomes database and comparative analysis system," *Nucleic Acids Res.*, vol. 40, pp. D115–122, Jan 2012.

[10] H. W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K. F. Mayer, V. Stumpflen, and A. Antonov, "MIPS: curated databases and comprehensive secondary data resources in 2010," *Nucleic Acids Res.*, vol. 39, pp. D220–224, Jan 2011.

[11] A. A. Salamov and V. V. Solovyev, "Ab initio gene finding in Drosophila genomic DNA," *Genome Res.*, vol. 10, pp. 516–522, Apr 2000.

[12] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "AUGUSTUS: ab initio prediction of alternative transcripts," *Nucleic Acids Res.*, vol. 34, pp. W435–439, Jul 2006.

[13] A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky, "Gene identification in novel eukaryotic genomes by self-training algorithm," *Nucleic Acids Res.*, vol. 33, pp. 6494–6506, 2005.

[14] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, pp. 57–63, Jan 2009.

[15] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, "dbEST–database for "expressed sequence tags"," *Nat. Genet.*, vol. 4, pp. 332–333, Aug 1993.

[16] S. Tanner, Z. Shen, J. Ng, L. Florea, R. Guigo, S. P. Briggs, and V. Bafna, "Improving gene annotation using peptide mass spectrometry," *Genome Res.*, vol. 17, pp. 231–239, Feb 2007.

[17] P. S. Schnable *et al.*, "The B73 maize genome: complexity, diversity, and dynamics," *Science*, vol. 326, pp. 1112–1115, Nov 2009.

[18] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, *et al.*, "What is a gene, post-ENCODE? History and updated definition," *Genome Res.*, vol. 17, pp. 669–681, Jun 2007.

[19] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, *et al.*, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nat. Genet.*, vol. 40, pp. 722–729, Jun 2008.

[20] P. J. Farabaugh, "Programmed translational frameshifting," *Annu. Rev. Genet.*, vol. 30, pp. 507–528, 1996.

[21] V. Curwen, E. Eyras, T. D. Andrews, L. Clarke, E. Mongin, S. M. Searle, and M. Clamp, "The Ensembl automatic gene annotation system," *Genome Res.*, vol. 14, pp. 942–950, May 2004.

[22] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, *et al.*, "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science*, vol. 252, pp. 1651–1656, Jun 1991.

[23] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, *et al.*, “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project,” *Nature*, vol. 447, pp. 799–816, Jun 2007.

[24] E. Huala, A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, *et al.*, “The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant,” *Nucleic Acids Res.*, vol. 29, pp. 102–105, Jan 2001.

[25] E. Pennisi, “Genomics. DNA study forces rethink of what it means to be a gene,” *Science*, vol. 316, pp. 1556–1557, Jun 2007.

[26] N. Gupta, S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, *et al.*, “Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation,” *Genome Res.*, vol. 17, pp. 1362–1377, Sep 2007.

[27] Z. Kan, E. C. Rouchka, W. R. Gish, and D. J. States, “Gene structure prediction and alternative splicing analysis using genomically aligned ESTs,” *Genome Res.*, vol. 11, pp. 889–900, May 2001.

[28] L. Florea, V. Di Francesco, J. Miller, R. Turner, A. Yao, M. Harris, *et al.*, “Gene and alternative splicing annotation with AIR,” *Genome Res.*, vol. 15, pp. 54–66, Jan 2005.

[29] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, *et al.*, “Distinguishing protein-coding and noncoding genes in the human genome,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, pp. 19428–19433, Dec 2007.

[30] C. Mathe, M. F. Sagot, T. Schiex, and P. Rouze, “Current methods of gene prediction, their strengths and weaknesses,” *Nucleic Acids Res.*, vol. 30, pp. 4103–4117, Oct 2002.

[31] G. Oshiro, L. M. Wodicka, M. P. Washburn, J. R. Yates, D. J. Lockhart, and E. A. Winzeler, “Parallel identification of new genes in Saccharomyces cerevisiae,” *Genome Res.*, vol. 12, pp. 1210–1220, Aug 2002.

[32] C. Burge and S. Karlin, “Prediction of complete gene structures in human genomic DNA,” *J. Mol. Biol.*, vol. 268, pp. 78–94, Apr 1997.

[33] J. R. Yates, J. K. Eng, and A. L. McCormack, “Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases,” *Anal. Chem.*, vol. 67, pp. 3202–3210, Sep 1995.

[34] B. P. Lewis, R. E. Green, and S. E. Brenner, "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 189–192, Jan 2003.

[35] R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, *et al.*, "The Universal Protein Resource (UniProt) in 2010," *Nucleic Acids Res.*, vol. 38, pp. D142–148, Jan 2010.

[36] J. R. Yates, C. I. Ruse, and A. Nakorchevsky, "Proteomics by mass spectrometry: approaches, advances, and applications," *Annu Rev Biomed Eng*, vol. 11, pp. 49–79, 2009.

[37] P. Flicek *et al.*, "Ensembl's 10th year," *Nucleic Acids Res.*, vol. 38, pp. D557–562, Jan 2010.

[38] A. Goffeau *et al.*, "Life with 6000 genes," *Science*, vol. 274, pp. 563–567, Oct 1996.

[39] C. elegans Sequencing Consortium, "Genome sequence of the nematode C. elegans: a platform for investigating biology," *Science*, vol. 282, pp. 2012–2018, Dec 1998.

[40] T. A. G. Initiative, "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana," *Nature*, vol. 408, pp. 796–815, Dec 2000.

[41] M. D. Adams *et al.*, "The Genome Sequence of Drosophila melanogaster," *Science*, vol. 287, no. 5461, pp. 2185–2195, 2000.

[42] J. C. Venter *et al.*, "The sequence of the human genome," *Science*, vol. 291, pp. 1304–1351, Feb 2001.

[43] E. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb 2001.

[44] R. A. Holt *et al.*, "The genome sequence of the malaria mosquito Anopheles gambiae," *Science*, vol. 298, pp. 129–149, Oct 2002.

[45] D. E. Kalume, S. Peri, R. Reddy, J. Zhong, M. Okulate, N. Kumar, and A. Pandey, "Genome annotation of Anopheles gambiae using mass spectrometry-derived data," *BMC Genomics*, vol. 6, p. 128, 2005.

[46] E. Brunner *et al.*, "A high-quality catalog of the Drosophila melanogaster proteome," *Nat. Biotechnol.*, vol. 25, pp. 576–583, May 2007.

[47] K. Baerenfaller *et al.*, "Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics," *Science*, vol. 320, pp. 938–941, May 2008.

[48] N. E. Castellana, S. H. Payne, Z. Shen, M. Stanke, V. Bafna, and S. P. Briggs, “Discovery and revision of Arabidopsis genes by proteogenomics,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 21034–21038, Dec 2008.

[49] M. P. Washburn, D. Wolters, and J. R. Yates, “Large-scale analysis of the yeast proteome by multidimensional protein identification technology,” *Nat. Biotechnol.*, vol. 19, pp. 242–247, Mar 2001.

[50] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, vol. 20, pp. 3551–3567, Dec 1999.

[51] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth, and S. P. Gygi, “Intensity-based protein identification by machine learning from a library of tandem mass spectra,” *Nat. Biotechnol.*, vol. 22, pp. 214–219, Feb 2004.

[52] Y. Wan, A. Yang, and T. Chen, “PepHMM: a hidden Markov model based scoring function for mass spectrometry database search,” *Anal. Chem.*, vol. 78, pp. 432–437, Jan 2006.

[53] M. Bern, Y. Cai, and D. Goldberg, “Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry,” *Anal. Chem.*, vol. 79, pp. 1393–1400, Feb 2007.

[54] A. A. Klammer, S. M. Reynolds, J. A. Bilmes, M. J. MacCoss, and W. S. Noble, “Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification,” *Bioinformatics*, vol. 24, pp. i348–356, Jul 2008.

[55] A. M. Frank, “A ranking-based scoring function for peptide-spectrum matches,” *J. Proteome Res.*, vol. 8, pp. 2241–2252, May 2009.

[56] A. M. Frank, “Predicting intensity ranks of peptide fragment ions,” *J. Proteome Res.*, vol. 8, pp. 2226–2240, May 2009.

[57] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, “Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search,” *Anal. Chem.*, vol. 74, pp. 5383–5392, Oct 2002.

[58] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry,” *Nat. Methods*, vol. 4, pp. 207–214, Mar 2007.

[59] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, “Empirical Bayes analysis of a microarray experiment.,” *J. Am. Stat. Assoc.*, vol. 96, pp. 1151–1160, 2001.

[60] S. Kim, N. Gupta, and P. A. Pevzner, "Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases," *J. Proteome Res.*, vol. 7, pp. 3354–3363, Aug 2008.

[61] G. Alves and Y. K. Yu, "Statistical Characterization of a 1D Random Potential Problem - with applications in score statistics of MS-based peptide sequencing," *Physica A*, vol. 387, pp. 6538–6544, Nov 2008.

[62] D. Fermin, B. B. Allen, T. W. Blackwell, R. Menon, M. Adamski, Y. Xu, P. Ulintz, G. S. Omenn, and D. J. States, "Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics," *Genome Biol.*, vol. 7, p. R35, 2006.

[63] M. A. Grobei, E. Qeli, E. Brunner, H. Rehrauer, R. Zhang, B. Roschitzki, *et al.*, "Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function," *Genome Res.*, vol. 19, pp. 1786–1800, Oct 2009.

[64] J. D. Jaffe, H. C. Berg, and G. M. Church, "Proteogenomic mapping as a complementary method to perform genome annotation," *Proteomics*, vol. 4, pp. 59–77, Jan 2004.

[65] F. Desiere, E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, *et al.*, "Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry," *Genome Biol.*, vol. 6, p. R9, 2005.

[66] N. J. Edwards, "Novel peptide identification from tandem mass spectra using ESTs and sequence database compression," *Mol. Syst. Biol.*, vol. 3, p. 102, 2007.

[67] N. de Bruijn, "A combinatorial problem," *Proc. Kon. Ned. Akad. Wetensch.*, vol. 49, pp. 758–764, 1946.

[68] E. Blanco, G. Parra, and R. Guigo, "Using geneid to identify genes," *Curr Protoc Bioinformatics*, vol. Chapter 4, p. Unit 4.3, Jun 2007.

[69] M. Stanke, O. Schoffmann, B. Morgenstern, and S. Waack, "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources," *BMC Bioinformatics*, vol. 7, p. 62, 2006.

[70] B. Kuster, P. Mortensen, J. S. Andersen, and M. Mann, "Mass spectrometry allows direct identification of proteins in large genomes," *Proteomics*, vol. 1, pp. 641–650, May 2001.

[71] "The 1000 genome project."

[72] C. Kumar-Sinha, S. A. Tomlins, and A. M. Chinnaiyan, "Recurrent gene fusions in prostate cancer," *Nat. Rev. Cancer*, vol. 8, pp. 497–511, Jul 2008.

[73] R. S. Mani, S. A. Tomlins, K. Callahan, A. Ghosh, M. K. Nyati, S. Varambally, N. Palanisamy, and A. M. Chinnaiyan, "Induced chromosomal proximity and gene fusions in prostate cancer," *Science*, vol. 326, p. 1230, Nov 2009.

[74] J. S. Choudhary, W. P. Blackstock, D. M. Creasy, and J. S. Cottrell, "Interrogating the human genome using uninterpreted mass spectrometry data," *Proteomics*, vol. 1, pp. 651–667, May 2001.

[75] M. Bern, D. Goldberg, W. H. McDonald, and J. R. Yates, "Automatic quality assessment of peptide tandem mass spectra," *Bioinformatics*, vol. 20 Suppl 1, pp. 49–54, Aug 2004.

[76] A. I. Nesvizhskii, F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, W. Gruissem, *et al.*, "Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides," *Mol. Cell Proteomics*, vol. 5, pp. 652–670, Apr 2006.

[77] A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith, *et al.*, "Clustering millions of tandem mass spectra," *J. Proteome Res.*, vol. 7, pp. 113–122, Jan 2008.

[78] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner, "Identification of post-translational modifications by blind search of mass spectra," *Nat. Biotechnol.*, vol. 23, pp. 1562–1567, Dec 2005.

[79] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Anal. Chem.*, vol. 75, pp. 4646–4658, Sep 2003.

[80] A. Savidor, R. S. Donahoo, O. Hurtado-Gonzales, N. C. Verberkmoes, M. B. Shah, K. H. Lamour, *et al.*, "Expressed peptide tags: an additional layer of data for genome annotation," *J. Proteome Res.*, vol. 5, pp. 3048–3058, Nov 2006.

[81] G. E. Merrihew, C. Davis, B. Ewing, G. Williams, L. Kall, B. E. Frewen, W. S. Noble, P. Green, J. H. Thomas, and M. J. MacCoss, "Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations," *Genome Res.*, vol. 18, pp. 1660–1669, Oct 2008.

[82] M. Baudet, P. Ortet, J. C. Gaillard, B. Fernandez, P. Guerin, C. Enjalbal, *et al.*, "Proteomic-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons," *Mol. Cell Proteomics*, Oct 2009.

[83] X. J. Wang, T. Gaasterland, and N. H. Chua, "Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana," *Genome Biol.*, vol. 6, no. 4, p. R30, 2005.

[84] E. Lasonder, Y. Ishihama, J. S. Andersen, A. M. Vermunt, A. Pain, R. W. Sauerwein, *et al.*, "Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry," *Nature*, vol. 419, pp. 537–542, Oct 2002.

[85] A. de Groot, R. Dulermo, P. Ortet, L. Blanchard, P. Guerin, B. Fernandez, *et al.*, "Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti," *PLoS Genet.*, vol. 5, p. e1000434, Mar 2009.

[86] S. Gallien, E. Perrodou, C. Carapito, C. Deshayes, J. M. Reyrat, A. Van Dorsselaer, *et al.*, "Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol," *Genome Res.*, vol. 19, pp. 128–135, Jan 2009.

[87] N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, *et al.*, "Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes," *Genome Res.*, vol. 18, pp. 1133–1142, Jul 2008.

[88] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0," *J. Mol. Biol.*, vol. 340, pp. 783–795, Jul 2004.

[89] K. Hiller, A. Grote, M. Scheer, R. Munch, and D. Jahn, "PrediSi: prediction of signal peptides and their cleavage positions," *Nucleic Acids Res.*, vol. 32, pp. W375–379, Jul 2004.

[90] S. Bonissone, N. Gupta, M. Romine, and P. Pevzner, "Comparative proteogenomics reveals a possible functional role of N-terminal methionine excision." submitted.

[91] J. M. Asara, M. H. Schweitzer, L. M. Freimark, M. Phillips, and L. C. Cantley, "Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry," *Science*, vol. 316, pp. 280–285, Apr 2007.

[92] J. M. Asara, J. S. Garavelli, D. A. Slatter, M. H. Schweitzer, L. M. Freimark, M. Phillips, and L. C. Cantley, "Interpreting sequences from mastodon and T. rex," *Science*, vol. 317, pp. 1324–1325, Sep 2007.

[93] M. Buckley *et al.*, “Comment on ”Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry”,” *Science*, vol. 319, p. 33; author reply 33, Jan 2008.

[94] P. A. Pevzner, S. Kim, and J. Ng, “Comment on ”Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry”,” *Science*, vol. 321, p. 1040; author reply 1040, Aug 2008.

[95] A. Shevchenko, S. Sunyaev, A. Loboda, A. Shevchenko, P. Bork, W. Ens, *et al.*, “Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching,” *Anal. Chem.*, vol. 73, pp. 1917–1926, May 2001.

[96] S. Na, J. Jeong, H. Park, K. J. Lee, and E. Paek, “Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach,” *Mol. Cell Proteomics*, vol. 7, pp. 2452–2463, Dec 2008.

[97] Y. Han, B. Ma, and K. Zhang, “SPIDER: software for protein identification from sequence tags with de novo sequencing error,” *J Bioinform Comput Biol*, vol. 3, pp. 697–716, Jun 2005.

[98] S. Dasari, M. C. Chambers, R. J. Slebos, L. Zimmerman, A. J. Ham, and D. L. Tabb, “TagRecon: high-throughput mutation identification through sequence tagging,” *J Proteome Res*, Feb 2010.

[99] N. Bandeira, V. Pham, P. Pevzner, D. Arnott, and J. R. Lill, “Automated de novo protein sequencing of monoclonal antibodies,” *Nat. Biotechnol.*, vol. 26, pp. 1336–1338, Dec 2008.

[100] X. Liu, Y. Han, D. Yuen, and B. Ma, “Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy,” *Bioinformatics*, vol. 25, pp. 2174–2180, 2009.

[101] N. E. Castellana, V. Pham, D. Arnott, J. R. Lill, and V. Bafna, “Template proteogenomics: sequencing whole proteins using an imperfect database,” *Mol Cell Proteomics*, Feb 2010.

[102] R. J. Ram, N. C. Verberkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, *et al.*, “Community proteomics of a natural microbial biofilm,” *Science*, vol. 308, pp. 1915–1920, Jun 2005.

[103] E. S. Klaassens, W. M. de Vos, and E. E. Vaughan, “Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract,” *Appl. Environ. Microbiol.*, vol. 73, pp. 1388–1392, Feb 2007.

[104] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, "Sequencing and comparison of yeast species to identify genes and regulatory elements," *Nature*, vol. 423, pp. 241–254, May 2003.

[105] M. F. Lin, J. W. Carlson, M. A. Crosby, B. B. Matthews, C. Yu, S. Park, K. H. Wan, A. J. Schroeder, L. S. Gramates, S. E. St Pierre, M. Roark, K. L. Wiley, R. J. Kulathinal, P. Zhang, K. V. Myrick, J. V. Antone, S. E. Celniker, W. M. Gelbart, and M. Kellis, "Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes," *Genome Res.*, vol. 17, pp. 1823–1836, Dec 2007.

[106] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler, "Using native and syntenically mapped cDNA alignments to improve de novo gene finding," *Bioinformatics*, vol. 24, pp. 637–644, Mar 2008.

[107] S. H. Payne, M. Yau, M. B. Smolka, S. Tanner, H. Zhou, and V. Bafna, "Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis," *J. Proteome Res.*, vol. 7, pp. 3373–3381, Aug 2008.

[108] B. B. Wang and V. Brendel, "Genomewide comparative analysis of alternative splicing in plants," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 7175–7180, May 2006.

[109] N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, and S. R. Wessler, "Pack-MULE transposable elements mediate gene evolution in plants," *Nature*, vol. 431, pp. 569–573, Sep 2004.

[110] K. Hanada, X. Zhang, J. O. Borevitz, W. H. Li, and S. H. Shiu, "A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection," *Genome Res.*, vol. 17, pp. 632–640, May 2007.

[111] S. L. DeBlasio, D. L. Luesse, and R. P. Hangarter, "A plant-specific protein essential for blue-light-induced chloroplast movements," *Plant Physiol.*, vol. 139, pp. 101–114, Sep 2005.

[112] S. B. Cannon, A. Mitra, A. Baumgarten, N. D. Young, and G. May, "The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana," *BMC Plant Biol.*, vol. 4, p. 10, Jun 2004.

[113] S. R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nat. Rev. Genet.*, vol. 2, pp. 919–929, Dec 2001.

[114] N. Castellana and V. Bafna, "Proteogenomics to discover the full coding content of genomes: a computational perspective," *J Proteomics*, vol. 73, pp. 2124–2135, Oct 2010.

[115] G. E. Merrihew, C. Davis, B. Ewing, G. Williams, L. Kall, B. E. Frewen, W. S. Noble, P. Green, J. H. Thomas, and M. J. MacCoss, "Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations," *Genome Res.*, vol. 18, pp. 1660–1669, Oct 2008.

[116] E. Venter, R. D. Smith, and S. H. Payne, "Proteogenomic analysis of bacteria and archaea: a 46 organism case study," *PLoS ONE*, vol. 6, p. e27587, 2011.

[117] D. S. Kelkar, D. Kumar, P. Kumar, L. Balakrishnan, B. Muthusamy, A. K. Yadav, P. Shrivastava, A. Marimuthu, S. Anand, H. Sundaram, R. Kingsbury, H. C. Harsha, B. Nair, T. S. Keshava Prasad, D. S. Chauhan, K. Katoch, V. M. Katoch, P. Kumar, R. Chaerkady, S. Ramachandran, D. Dash, and A. Pandey, "Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry," *Mol Cell Proteomics*, Oct 2011.

[118] J. Walley, Z. Shen, and S. Briggs, "An Atlas of Maize Seed Proteotypes," *in preparation*, 2012.

[119] E. G. Wagner and R. W. Simons, "Antisense RNA control in bacteria, phages, and plasmids," *Annu. Rev. Microbiol.*, vol. 48, pp. 713–742, 1994.

[120] B. A. Williams, C. H. Slamovits, N. J. Patron, N. M. Fast, and P. J. Keeling, "A high frequency of overlapping gene expression in compacted eukaryotic genomes," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 10936–10941, Aug 2005.

[121] R. Yelin, D. Dahary, R. Sorek, E. Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, and G. Rotman, "Widespread occurrence of antisense transcription in the human genome," *Nat. Biotechnol.*, vol. 21, pp. 379–386, Apr 2003.

[122] M. E. Fahey, T. F. Moore, and D. G. Higgins, "Overlapping antisense transcription in the human genome," *Comp. Funct. Genomics*, vol. 3, no. 3, pp. 244–253, 2002.

[123] C. R. Sanna, W. H. Li, and L. Zhang, "Overlapping genes in the human and mouse genomes," *BMC Genomics*, vol. 9, p. 169, 2008.

[124] O. Borsani, J. Zhu, P. E. Verslues, R. Sunkar, and J. K. Zhu, "Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis," *Cell*, vol. 123, pp. 1279–1291, Dec 2005.

[125] N. Osato, Y. Suzuki, K. Ikeo, and T. Gojobori, "Transcriptional interferences in cis natural antisense transcripts of humans and mice," *Genetics*, vol. 176, pp. 1299–1306, Jun 2007.

[126] A. Frank and P. Pevzner, “PepNovo: de novo peptide sequencing via probabilistic network modeling,” *Anal. Chem.*, vol. 77, pp. 964–973, Feb 2005.

[127] W. J. Kent, “BLAT–the BLAST-like alignment tool,” *Genome Res.*, vol. 12, pp. 656–664, Apr 2002.

[128] P. Li, L. Ponnala, N. Gandotra, L. Wang, Y. Si, S. L. Tausta, T. H. Kebrom, N. Provart, R. Patel, C. R. Myers, E. J. Reidel, R. Turgeon, P. Liu, Q. Sun, T. Nelson, and T. P. Brutnell, “The developmental dynamics of the maize leaf transcriptome,” *Nat. Genet.*, vol. 42, pp. 1060–1067, Dec 2010.

[129] R. Davidson, C. Hansey, M. Gowda, K. Childs, H. Lin, B. Vaillancourt, R. Sekhon, N. de Leon, S. Kaeppler, N. Jiang, and C. Buell, “Utility of RNA sequencing for analysis of maize reproductive transcriptomes,” *Plant Gen.*, vol. 4, pp. 191–203, 2011.

[130] J. Eng, A. McCormack, and J. Y. III, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,” *J. Am. Soc. Mass Spectrom.*, vol. 5, pp. 976–989, 1994.

[131] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner, “Identification of post-translational modifications by blind search of mass spectra,” *Nat. Biotechnol.*, vol. 23, pp. 1562–1567, Dec 2005.

[132] B. C. Searle, S. Dasari, P. A. Wilmarth, M. Turner, A. P. Reddy, L. L. David, *et al.*, “Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm,” *J. Proteome Res.*, vol. 4, pp. 546–554, 2005.

[133] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, “PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry,” *Rapid Commun. Mass Spectrom.*, vol. 17, pp. 2337–2342, 2003.

[134] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, “GenBank,” *Nucleic Acids Res.*, vol. 36, pp. 25–30, Jan 2008.

[135] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, “dbEST–database for ”expressed sequence tags”,” *Nat. Genet.*, vol. 4, pp. 332–333, Aug 1993.

[136] R. Menon, Q. Zhang, Y. Zhang, D. Fermin, N. Bardeesy, R. A. DePinho, C. Lu, S. M. Hanash, G. S. Omenn, and D. J. States, “Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer,” *Cancer Res.*, vol. 69, pp. 300–309, Jan 2009.

[137] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nat. Genet.*, vol. 36, pp. 949–951, Sep 2004.

[138] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler, "Large-scale copy number polymorphism in the human genome," *Science*, vol. 305, pp. 525–528, Jul 2004.

[139] A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith, and P. A. Pevzner, "Clustering millions of tandem mass spectra," *J. Proteome Res.*, vol. 7, pp. 113–122, Jan 2008.

[140] P. A. Pevzner, V. Dancik, and C. L. Tang, "Mutation-tolerant protein identification by mass spectrometry," *J. Comput. Biol.*, vol. 7, pp. 777–787, 2000.

[141] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis.* Cambridge University Press, 1998.

[142] Y. Wan, A. Yang, and T. Chen, "PepHMM: a hidden Markov model based scoring function for mass spectrometry database search," *Anal. Chem.*, vol. 78, pp. 432–437, Jan 2006.

[143] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Infor. Theory*, vol. 13, pp. 260–269, April 1967.

[144] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner, "De novo peptide sequencing via tandem mass spectrometry," *J. Comput. Biol.*, vol. 6, pp. 327–342, 1999.

[145] R. M. MacCallum, A. C. Martin, and J. M. Thornton, "Antibody-antigen interactions: contact analysis and binding site topography," *J. Mol. Biol.*, vol. 262, pp. 732–745, Oct 1996.

[146] M. P. Lefranc, V. Giudicelli, C. Ginestoux, J. Bodmer, W. Muller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbie, and D. Chaume, "IMGT, the international ImMunoGeneTics database," *Nucleic Acids Res.*, vol. 27, pp. 209–212, Jan 1999.

[147] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau,

V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The sequence of the human genome," *Science*, vol. 291, pp. 1304–1351, Feb 2001.

[148] N. Bandeira, K. R. Clauser, and P. A. Pevzner, "Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified

proteins," *Mol. Cell Proteomics*, vol. 6, pp. 1123–1134, Jul 2007.

[149] N. E. Castellana, K. McCutcheon, V. C. Pham, K. Harden, A. Nguyen, J. Young, C. Adams, K. Schroeder, D. Arnott, V. Bafna, J. L. Grogan, and J. R. Lill, "Resurrection of a clinical antibody: template proteogenomic de novo proteomic sequencing and reverse engineering of an anti-lymphotoxin-antibody," *Proteomics*, vol. 11, pp. 395–405, Feb 2011.