Developing a bioinformatics framework for proteogenomics

This thesis was submitted in fulfilment of the requirements for the degree of Doctor of Philosophy of Murdoch University

Presented by:

Brett Steven Chapman

BSc. (Hons), GradDip

School of Engineering and Information Technology

Murdoch University

Murdoch, Western Australia

September 2015

DECLARATION

I declare that this thesis is my own account of my research and it contains as its main content work, which has not previously been submitted for a degree at any tertiary education institution.

Brett Steven Chapman

ABSTRACT

In the last 15 years, since the human genome was first sequenced, genome sequencing and annotation have continued to improve. However, genome annotation has not kept up with the accelerating rate of genome sequencing and as a result there is now a large backlog of genomic data waiting to be interpreted both quickly and accurately. Through advances in proteomics a new field has emerged to help improve genome annotation, termed proteogenomics, which uses peptide mass spectrometry data, enabling the discovery of novel protein coding genes, as well as the refinement and validation of known and putative protein-coding genes.

The annotation of genomes relies heavily on *ab initio* gene prediction programs and/or mapping of a range of RNA transcripts. Although this method provides insights into the gene content of genomes it is unable to distinguish protein-coding genes from putative non-coding RNA genes. This problem is further confounded by the fact that only 5% of the public protein sequence repository at UniProt/SwissProt has been curated and derived from actual protein evidence.

This thesis contends that it is critically important to incorporate proteomics data into genome annotation pipelines to provide experimental protein-coding evidence. Although there have been major improvements in proteogenomics over the last decade there are still numerous challenges to overcome. These key challenges include the loss of sensitivity when using inflated search spaces of putative sequences, how best to interpret novel identifications and how best to control for false discoveries.

This thesis addresses the existing gap between the use of genomic and proteomic sources for accurate genome annotation by applying a proteogenomics approach with a customised methodology. This new approach was applied within four case studies: a prokaryote bacterium; a monocotyledonous wheat plant; a dicotyledonous grape plant;

and human. The key contributions of this thesis are: a new methodology for proteogenomics analysis; 145 suggested gene refinements in *Bradyrhizobium diazoefficiens* (nitrogen-fixing bacteria); 55 new gene predictions (57 protein isoforms) in *Vitis vinifera* (grape); 49 new gene predictions (52 protein isoforms) in *Homo sapiens* (human); and 67 new gene predictions (70 protein isoforms) in *Triticum aestivum* (bread wheat). Lastly, a number of possible improvements for the studies conducted in this thesis and proteogenomics as a whole have been identified and discussed.

GENERAL ACKNOWLEDGEMENTS

I would like to thank my supervisor Matthew Bellgard for his guidance over the course of my PhD. I would also like to thank my colleagues Adam Hunter, Paula Moolhuijzen, Michael Black, Roberto Barrero, John McCooke and Kathryn Napier for their advice and assistance throughout my PhD.

I would also like to thank Natalie Castellana, Samuel Payne and Sunghee Woo for the development of their proteogenomics tool and for their patience with questions through correspondence, as well as Steven Van Sluyter, Ryan Ghan, Nicola Vitulo, Morgan Giddings, John Wrobel and Rudi Appels, for providing me access to datasets I wouldn't have otherwise been able to obtain at the time.

I would also like to thank the Centre for Comparative Genomics and the Pawsey Supercomputing Centre for access to their compute resources, without which this thesis would not have been possible.

Lastly, my deepest gratitude goes to my family: my wife Rhonda Chapman, son Ethan Chapman, parents Lee and Judy Chapman, and sister Carla Corbitt, who have all provided me support, listened to my troubles and endured the years of my PhD candidature.

DISCLAIMER

This PhD thesis consists of chapters that have been accepted for publication (Chapter 4) and also in preparation for publication (Chapters 5 to 7), with similar or the same applied methods. To reduce repetition between chapters, the core methodology applied throughout all chapters is compiled in Chapter 3 and as a result, each of the final published manuscripts may differ slightly.

PUBLICATIONS ARISING FROM THIS THESIS

Chapman, B., Castellana, N., Apffel, A., Ghan, R., *et al.*, Plant proteogenomics: from protein extraction to improved gene predictions. *Methods in molecular biology* 2013, *1002*, 267-294.

Chapman, B., Bellgard, M., High-throughput parallel proteogenomics: A bacterial case study. *Proteomics* 2014, *14*, 2780-2789.

Mayer, K., Rogers, J., Doležel, J., Pozniak, C., **Chapman, B.**, Bellgard, M., *et al.*, A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 2014, *345*, 1251788.

OTHER PUBLICATIONS ARISING FROM THE PERIOD OF CANDIDATURE

Bellgard, M., Taplin, R., **Chapman, B.**, Livk, A., *et al.*, Classification of fish samples via an integrated proteomics and bioinformatics approach. *Proteomics* 2013, *13*, 3124-3130.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Annotation event The result of a peptide cluster conflicting with the current known annotation. Examples of annotation events include: exon boundaries, novel genes, gene boundaries, frame-shifts, reverse strands, translated UTRs, and novel splice junctions.

Combined FDR When an FDR cut-off is applied across all PSM identifications.

Event probability The probability that an annotation event is correct, based on the product of all posterior probabilities or local FDRs of PSMs (1 – local FDR) divided by the number of co-locations across the genome.

Two-stage FDR When an FDR cut-off is applied to different sets of PSM identifications, particularly all known PSMs and all novel PSMs.

Two-pass search When an initial database search is conducted to define the search space for a second search, thereby reducing the final search space size and improving the significance and rate of identifications.

Peptide linkage distance Determines how peptides are grouped together to form a peptide cluster (intragenic space) and the distance between peptide clusters (intergenic space). The value impacts the number and type of annotation events identifiable from each peptide cluster.

Protein isoforms A term used to describe all proteins from a single gene, generally at the genetic level due to genetic variants and alternative RNA splicing.

Proteoforms A term used to describe all possible proteins derived from a single gene, including due to genetic variants, alternative RNA splicing, and post-translational modifications.

Proteotypic peptide A peptide that can uniquely identify a single protein.

LIST OF ABBREVIATIONS

aa	Amino acid
Advanced APS	Advanced Average Peptide Score
AIF	All-Ion Fragmentation
Aldente	Advanced Large-scale iDENTification Engine
ANDI-MS	ANalytical Data Interchange Mass Spectrometry
API	Application Programming Interface
AS	Alternative Splicing
ATP	Adenosine Triphosphate
AutoFACT	Automatic Functional Annotation and Classification Tool
BAM	Binary Alignment/Map
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
bp	Base pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
CCD	Charge-Coupled Device
cDNA	Copy DNA
CDS	Coding DNA Sequence
CEA	Clique-enrichment approach
CE	Capillary Electrophoresis

ChiTaRS	Chimeric Transcripts and RNA-Sequencing data
CID	Collision Induced Dissociation
CIEF	Capillary Isoelectric Focusing
COPaKB	Cardiac Organellar Protein Atlas Knowledgebase
CRF	Conditional Random Field
CRIBI	Centro di Ricerce Interdipartimentale per le Biotecnologie Innovative
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSV	Comma Separated Values
CTP	Cytidine triphosphate
Da	Dalton
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxy-ribose nucleic acid
DRMAA	Distributed Resource Management Application API
EBP identifier	Empirical Bayes Protein identifier
EBV	Epstein Barr Virus
ECD	Electron Capture Dissociation
EMBL	European Molecular Biology Laboratory

ENCODE	ENCyclopedia Of DNA Elements
ePST	expressed Peptide Sequence Tag
ESI	Electro-Spray Ionisation
EST	Expressed Sequence Tag
ETD	Electron Transfer Dissociation
FASP	Filter-Aided Sample Preparation
FDR	False Discovery Rate
FL-cDNA	Full length cDNA
FPR	False Positive Rate
FT-ARM	Fourier Transform-All Reaction Monitoring
GA	Genome Analyzer
GAPP	Genome Annotating Proteomic Pipeline
GATK	Genome Analysis Toolkit
GB	Gigabyte
Gbp	Gigabase pair
GEO	Gene Expression Omnibus
GFF	General Feature Format
GFS	Genome-based peptide Fingerprint Scanning
GHMM	Generalised Hidden-Markov Model
GPF	Gas-Phase Fractionation

GPF	Genomic Peptide Finder
GPM	Global Proteome Machine
GridFTP	Grid File Transfer Protocol
GTP	Guanine Triphosphate
HAVANA	Human and Vertebrate Analysis and Annotation
HCD	Higher-energy Collisional Dissociation
$HDMS^E$	Ion mobility mass spectrometry assisted MS^E
HPC	High-performance computing
HPLC	High Performance Liquid Chromatography
HTML	Hyper Text Markup Language
HUPO	Human Proteome Organisation
Hybrid-FT	Hybrid-Fourier Transform
IGV	Integrative Genomics Viewer
IMG	Integrated Microbial Genomes
IMS	Ion Mobility Spectrometry
INDEL	Insertion and deletion
InsPecT	Interpretation of Spectra with Post-Translational modifications
IPI	International Protein Index
ISB-SPC	Institute for Systems Biology - Seattle Proteome Centre
IWGSC	International Wheat Genome Sequencing Consortium

JCAMP-DX	Joint Committee on Atomic and Molecular Physical Data
JSON	JavaScript Object Notation
Kbp	Kilobase pair
kDa	kilodalton
keV	Kilo electron volt
LC	Liquid Chromatography
lFDR	Local FDR
lincRNA	Large intergenic non-coding RNA
lncRNA	Long non-coding RNA
LTQ	Linear Trap Quadrupole
MALDI	Matrix-Assisted Laser Desorption/Ionization
MB	Megabyte
Mbp	Megabase pair
MDIP	Minimum acceptable Detectability for Identified Peptides
MGF	Mascot Generic Format
MIPS	Munich Information Center for Protein Sequences
miRNA	micro RNA
MMD	Maximum Mass Deviation
MOWSE	MOlecular Weight SEarch
MPI	Message Passing Interface

mRNA	Messenger RNA
MScDB	Mass Spectrometry-centric sequence Database
MSE	MS everything
MS-GF	MS-Generating Function
MS	Mass Spectrometry
MS/MS	Tandem mass spectrometry
MudPIT	Multidimensional Protein Identification Technology
MW	Molecular mass
NCBI	National Centre for Biotechnology Information
NetCDF	Network Common Data Form
NGS	Next Generation Sequencing
NIST	National Institute of Standards and Technology
NME	N-terminal Methionine Excision
NPCR	Novel Protein Coding Region
NR	Non-Redundant protein database
NTP	Nucleoside triphosphate
OMIM	Online Mendelian Inheritance in Man
OMSSA	Open Mass Spectrometry Search Algorithm
ORF	Open Reading Frame
PacBio	Pacific Bioscience

PAcIFIC	Precursor Acquisition Independent From Ion Count
PASA	Program to Assemble Spliced Alignments
PBS Pro	Portable Batch System Professional
PCR	Polymerase Chain Reaction
PE	Paired-end
PEPPI	PEPtidomics Protein Isoform
PEP	Posterior Error Probability
PFF	Peptide Fragment Fingerprinting
PGM	Personal Genome Machine
PG Nexus	Proteomic-Genomic Nexus
PGTools	Proteogenomic Tools
pI	Isoelectric point
PIITA	Precursor Ion Independent Top-down Algorithm
PMD	Protein Mutant Database
PMF	Peptide Mass Fingerprinting
PNNL	Pacific Norwest National Laboratories
PPGP	Prokaryote Proteogenomics Pipeline
PPM-Chain	Probability Profile Method – Chain
ppm	Parts per million
pre-mRNA	Precursor messenger RNA

PRIDE	PRoteomics IDEntifications database
PrSM	Protein-Spectrum-Match
PSI	Proteomics Standard Initiative
PSM	Peptide-Spectrum Match
PST	Peptide Sequence Tag
PTM	Post-Translational Modification
QIT	Quadrupole Ion Trap
QqQ	Triple Quadrupole
QTOF	Quadrupole Time-of-Flight
RAId	Robust Accurate Identification
RAST	Rapid Annotation using Subsystem Technology
RefSeq	Reference Sequence protein database
RNA	Ribonucleic acid
RP	Reverse Phase
S3	Amazon Simple Storage Service
SAM	Sequence Alignment/Map
SAP	Single Amino acid Polymorphism
SBS	Sequence By Synthesis
Scufl	Simple Conceptual Unified Language
SCX	Strong Cation Exchange

SEC	Size Exclusion liquid Chromatograph
SE	Single-end
SFTP	SSH File Transfer Protocol
SGE	Sun Grid Engine
SLURM	Simple Linux Utility for Resource Management
SMFS	Single Molecule Florescent Sequencing
SMRT	Single Molecule, Real Time sequencing
SMS	Single Molecule Sequencing
SNP	Single Nucleotide Polymorphism
SOLiD	Sequencing by Oligo Ligation Detection
SpectraST	Spectra Search Tool
SRA	Sequence Read Archive
SRM	Selective Reaction Monitoring
SSH	Secure Shell
SSM	Spectrum-Spectrum Match
STAR	Spliced Transcripts Alignment to a Reference
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVG	Scalable Vector Graphics
SVM	Support Vector Machine
SWATH-MS	Sequential Window Acquisition of all THeoretical Mass Spectra

TDA	Target-Decoy Approach
TIS	Translation Initiation Start
TOF	Time-of-Flight
TPP	Trans Proteomics Pipeline
tRNA	Transfer RNA
TSV	Tab-Separated Value
TTP	Thymidine Triphosphate
2D-PAGE	Two-dimensional Polyacrylamide Gel Electrophoresis
UCSC	University of California Santa Cruz
UniProtKB	UniProt Knowledgebase
UniProt	Universal Protein Resource
URGI	Unité de Recherche Génomique Info
USDA	United States Department of Agriculture
USTag	Unique Sequence Tag
UTR	Un-Translated Region
VEGA	Vertebrate Genome Annotation Database
XML	eXtensible Markup Language
ZMW	Zero-Mode Waveguide

1 INTRODUCTION

Proteogenomics is a multi-omics approach for genome annotation, integrating proteomics and genomics, as well as evidence from transcriptomics. Over the last decade proteogenomics has become a popular field of research, and is becoming a key player in global genome annotation efforts at an accelerating rate. Advances in proteogenomics have been driven by the demand to meet the ever-advancing rate of genome sequencing technologies, as many diverse species from the tree of life are sequenced [1, 2]. Proteogenomics, and other related approaches such as RNA-seq technology, are now increasingly being employed in genome annotation efforts. As a result, the impact of the "non-model" organism will be significantly reduced or non-existent in the not too distant future [3], as every genomics study begins to meet the same level of annotation as that of model organisms.

Genome annotation often uses public curated protein sequence repositories such as UniProt/SwissProt, however only 5% of the entries are derived from actual protein evidence, while the remaining 95% have been inferred from genomics. There are a number of challenges that need to be addressed in order to enable proteogenomics to resolve such an anomaly and to provide genome annotation with direct protein-coding evidence.

This thesis provides an in-depth review of the main challenges in proteogenomics and through the course of this thesis a new methodology evolved to improve the peptide identification rate from MS-experiments on large genome and transcriptome sequence data. The viability of the new methodology was demonstrated through four case studies: *Bradyrhizobium diazoefficiens* (nitrogen-fixing bacteria), *Vitis vinifera* (grape), *Homo sapiens* (human) and *Triticum aestivum* (bread wheat). Lastly, from insights gleaned from this thesis, there were considerations for how proteogenomics could be conducted in the future, such as using technologies to

supplement the database with sequence variants to account for sequence variations between the MS/MS spectra and target genome, considerations with the use of top-down proteomics and multiplexed data-independent acquisition (DIA) to improve depth and breadth of coverage, and spectral archives to improve specificity, sensitivity and enable a comprehensive screening of annotation events.

In Chapter 2, an overview of molecular biology, genomics, MS-based proteomics, and proteogenomics and bioinformatics workflow environments is presented. Within Section 2.1, the basics of molecular biology are covered in detail; Section 2.2 encompasses genomics, including the current methods employed in genome sequencing and genome annotation; Section 2.3 details MS-based proteomics, which includes technologies and tools involved in bottom-up proteomics, top-down proteomics, methods of MS/MS spectral interpretation, MS/MS spectral pre-processing procedures, database searching and statistical analyses involved in MS/MS spectral and peptide identifications, as well as outlining the latest emerging proteomics technologies; Section 2.4 provides an in-depth coverage of proteogenomics, which includes current trends in methodology, statistical analysis, the level of annotation which can be applied in proteogenomics, and the current tools which are available; and lastly in Section 2.5, bioinformatics workflow environments are discussed in terms of how they could impact analysis across a wide variety of -omics platforms, including proteogenomics, and include a range of the current popular workflow environments.

In Chapter 3, the methodologies employed within the thesis are described in detail, which includes pre-processing of datasets, conversion of data formats, RNA-seq alignments, database preparation, database searching, optimisation steps, the proteogenomics workflow and parameters used, different strategies applied in false discovery rate (FDR) filtering, screening of events, and gene prediction.

In Chapter 4, the proteogenomics analysis of *Bradyrhizobium diazoefficiens*, a nitrogen-fixing bacterium, was conducted using the current NCBI genome annotation and MS/MS spectra from the studies conducted by Delmotte et al. [4] and Koch et al. [5]. A total of 259 novel peptides were identified, contributing to 155 novel annotation events consisting of 9 frame shifts, 22 exon boundaries, 19 gene boundaries, 45 reverse strands, and 60 novel genes, annotating a total of 145 genes. A two-pass search approach to improve on sensitivity of the analysis was also investigated and a possible sequencing error was identified. This analysis contributed to a publication in the journal Proteomics [6].

In Chapter 5, the proteogenomics analysis of *Vitis vinifera* (grape) was conducted, improving on the current 12Xv2.1 genome annotation from the study conducted by Vitulo et al. [7] representing the variety Pinot Noir, an extension of the dissertation author's previous proteogenomics study [8]. A number of datasets were utilised, including a large MS/MS spectral dataset from [9], derived from Cabernet Sauvignon shoot tips and berry skins from the study in [8], while RNA-seq data was obtained from a variety of tissues [10], as well as a large, currently unpublished RNA-seq dataset from the berry skins of a range of grape varieties. A total of 133 novel peptides were identified, contributing to 341 novel annotation events, consisting of 5 frame shifts, 37 translated UTRs, 16 exon boundaries, 1 novel splice, 9 novel exons, 160 gene boundaries, 112 reverse strands and 1 novel gene event, annotating a total of 216 genes and 326 protein isoforms. These annotations led to a total of 110 novel peptides which contributed to validate 57 Augustus predicted proteins. In addition, a possible over-assembly of the genome and a reduction in sensitivity of analysis compared to smaller genomes was identified, indicating a need to further refine the method of FDR filtering.

In Chapter 6, the proteogenomics analysis of *Homo sapiens* (Human) was conducted, improving on the GENCODE v19 annotation. A large MS/MS spectral dataset from the ENCODE study in [11], derived from lymphoblastoid cell line GM12878, and RNA-seq data obtained from NCBI GEO accession GSE30567 were utilised. A total of 77 novel peptides were identified, contributing to 617 novel annotation events, consisting of 7 frame shifts, 4 translated UTRs, 27 exon boundaries, 23 novel exons, 289 gene boundaries, 262 reverse strands, and 5 novel gene events, annotating a total of 147 genes and 609 protein isoforms. These annotations led to a total of 66 novel peptides which supported 52 Augustus predicted proteins. In addition, a two-pass search approach with improved two-stage FDR strategy was established, which was able to identify 35 more novel peptides compared to previous established methods, and identified 15,020 more peptides than the previous ENCODE study [11].

In Chapter 7, the proteogenomics analysis of *Triticum aestivum* (bread wheat) was conducted, improving on the Wheat MIPS v2.2 genome annotation and also extending the proteogenomics analysis conducted by the dissertation author within the study by Mayer and colleagues [12]. A range of MS/MS spectra derived from different tissues and protease digests were derived from Wheat flour from cultivar Butte 86 in the study from Dupont and colleagues [13], digested with trypsin, chymotrypsin and thermolysin and meiotically developing anthers from a cross between rye *(Secale cereale* cultivar Petkus) and wheat (*Triticum aestivum* cultivar Chinese Spring), digested with trypsin and AspN [14]. RNA-seq data were also utilised, derived from Chinese Spring from [12, 15], including a range of other cultivars downloaded from the Sequence Read Archive (SRA). A total of 290 novel peptides were identified, contributing to 189 novel annotation events, consisting of 46 frame shifts, 9 translated UTRs, 17 exon boundaries, 39 novel exons, 17 gene boundaries, 24 reverse strands, and 37 novel gene events, annotating a total of 96 genes and 189 protein isoforms. These

annotations identified a total of 180 novel peptides which validated 70 Augustus predicted proteins. In addition, a two-pass search approach with improved two-stage FDR strategy was also applied, reducing the impact of the inflated search space. Based on the methodology applied, and comparing the same dataset, there was an ~8x improvement in sensitivity compared to the study reported in [12]. The benefits of merging proteogenomics results derived from multiple protease digests was also highlighted, as well as the negative impact and resulting difficulties that the highly fragmented wheat genome had on the accuracy of proteogenomics analysis.

In Chapter 8, the general conclusions outline the findings, caveats and conclusions drawn from the case studies, and provide suggestions to revisit some of these case studies in the future with more datasets and new methodologies, as well as provide suggestions for improvements and future directions that proteogenomics will likely take. These new methodologies and suggestions take the form of: 1) improving the accuracy of defining annotation events; 2) further refining the search space; 3) the addition of multiple other types of annotation events; 4) the use of variant information in splice graphs; 5) accounting for missing genomic regions in highly fragmented genomes using homologous sequences; 6) leveraging N-terminomics to help define events and confidently identify translation initiation start (TIS) sites; 7) ways to more accurately discriminate peptides as known and novel; 8) ways to improve the sensitivity of event identification with little or no impact to the false positive rate; 9) ways to improve the functionality of the proteogenomics workflow; 10) implement new technologies such as top-down proteomics and multiplexed data-independent acquisition (DIA) to improve depth and breadth of coverage; and finally 11) the use of highly specific MS/MS search tools such as spectral library search tools or spectral archives to bring proteogenomics into the domain of spectrum-spectrum matching, to vastly improve specificity and discrimination between true and false positives.

2 LITERATURE REVIEW

2.1 MOLECULAR BIOLOGY: A PRIMER

All organisms have a genome, which contains information to construct the building blocks of an organism. The genome itself is made up of deoxy-ribose nucleic acid (DNA) sequences, which is comprised of Adenine (A), Thyamine (T), Cytosine (C), and Guanine (G) nucleotides. Shorter stretches of sequences contained within are called genes and are categorized as protein coding and non-protein-coding, which lead to the construction of proteins and functional Ribonucleic acid (RNA) molecules, respectively.

The general model of protein production is the same for both eukaryotic and prokaryotic organisms, but the structure and organization of genes are quite different. Firstly, the genic region of DNA is transcribed or copied into messenger RNA (mRNA), also called a transcript, as it is ‘transcribed’ from the gene, and then each transcript is ‘translated’ by ribosomes into proteins, the functional building blocks of an organism. Each gene is also in proximity to a regulatory region, termed ‘enhancer’, ‘promoter’, and ‘repressor’ regions, which contain sequences responsible for the regulation of transcription of the gene and by extension regulation of the expression of proteins. This standard model of molecular biology is also referred to as the “central dogma of molecular biology” [16, 17], the fundamentals of which are summarized in Figure 2.1.

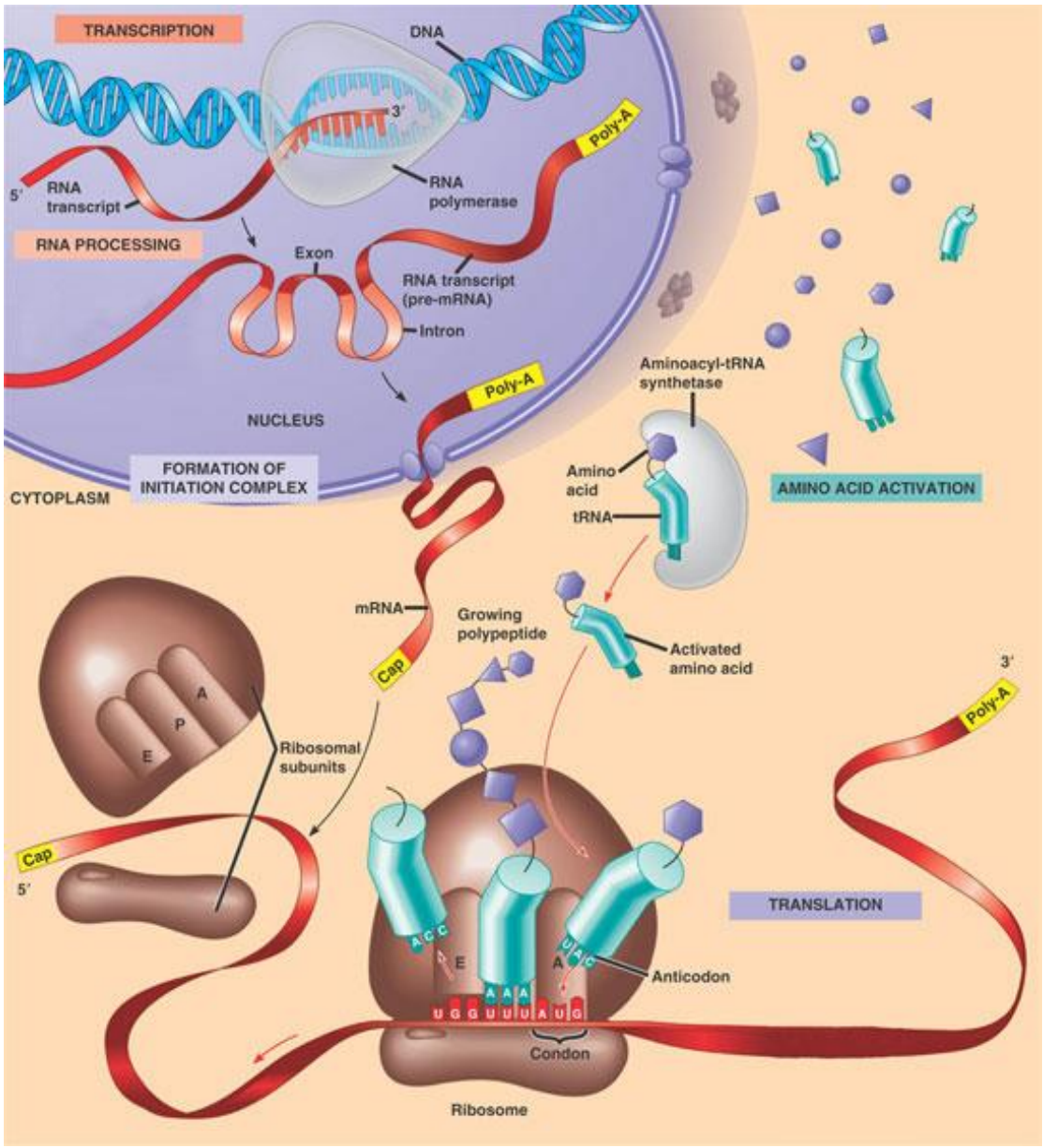


Figure 2.1 Central dogma of molecular biology

Genes are transcribed into pre-mRNA, processed by splicesomes into a variety of mature mRNA, which are then transported to the cytoplasm where the transcripts are translated into polypeptides, later forming into complex protein structures (image courtesy of the Online Computational Biology Textbook at PBWorks (http://compbio.pbworks.com/f/central_dogma.jpg)).

A more detailed examination of the processes highlighted in Figure 2.1, reveals a web of complex interactions involving RNA transcription and protein translation. These processes differ significantly between eukaryotes and prokaryotes. The protein-coding genes of eukaryotes are transcribed to a primary transcript or precursor messenger RNA (pre-mRNA), which takes place in the nucleus of the cell and is produced by an enzyme called RNA polymerase, which synthesizes an RNA strand in a 5' to 3' direction complementary to the DNA strand. The pre-mRNA then has its 5' end

capped with 7-methylguanosine by the enzyme guanylyltransferase which prevents degradation during translation, regulates exportation from the nucleus, promotes translation and 5' proximal intron excision [18, 19]. Multiple adenosine monophosphates are then added to its 3' end (poly(A) tail), through a process called polyadenylation, catalysed by polyadenylate polymerase. The addition of a poly(A) tail to mRNA, occurs in both eukaryotes and prokaryotes, and assists with a number of functions, which include transportation from the nucleus into the cytoplasm [20], translation [21], stability [22, 23], and conversely, its degradation [24].

The resulting polyadenylated RNA strand or premature mRNA consists of long stretches of both exons and introns and in eukaryotes the exons are spliced out by an enzyme called a spliceosome, which cuts out the introns at specific donor sites at the 5' end of the transcript and at the acceptor site of the 3' end of the transcript. Donor and acceptor sites are commonly of the GU-AG (donor-acceptor) variety, although more rarer versions do exist [25].

The resulting exons are then reassembled into a variety of different mature mRNAs containing one or many exons and in some rare cases introns are retained. The final structure of mRNA contains a 5' un-translated region (UTR) upstream and a 3' untranslated region downstream, flanking a region called an open reading frame (ORF). This entire process from precursor mRNA to the wide variety of mature mRNAs, is called alternative splicing (AS) [26, 27]. The mature mRNAs are then transported from the nucleus into the cytoplasm, in which their ORFs are translated into the final protein products. This step is accomplished by the ribosomes, composed of protein and RNA, which reads the ORF, three nucleotides at a time, referred to as a triplet codon, and translates the codon into an amino acid using different transfer RNAs (tRNAs) containing different bound amino acids. The tRNAs bind to three complementary bases or the anticodon as they pass through the ribosome, at which point the amino acids

ligate, forming a growing polypeptide chain which when complete folds into its functional three-dimensional structure and final protein product (Figure 2.1). The resulting functional proteins, which originate from the same gene, but are translated from different spliced mRNAs are known as protein isoforms, which can differ significantly in their sequence structure and biological functions [26]. In the case of non-protein coding RNA genes, the protein translations are skipped, and they are simply transcribed to RNA molecules. These non-protein-coding RNA molecules have been found to outnumber the protein-coding variety by a significant amount in vertebrate genomes [28] and are involved in regulatory processes, such as tRNAs important in the synthesis of proteins mentioned above and in higher eukaryotes micro RNAs (miRNAs), which are involved in the regulation of gene expression through the suppression of mRNA translation [29]. The distinction between protein-coding and non-protein-coding RNA was further explained by the recently completed ENCODE human genome project [30], which found that 80% of the human genome which was previously classified as "junk DNA", actually plays a functional role [31]. However, the exact proportion, which is functional, is highly debatable and would require further investigative analysis.

In complex organisms, such as humans, AS is clearly the source of such complexity and a dominant factor in regulation and function, as well as a source of protein diversity and organism complexity in higher eukaryotes [26, 32, 33]. The role that AS plays in protein diversity and complexity is illustrated when comparing the relatively few genes discovered in the human genome (~23,000 genes), where there are approximately 6 or more protein coding transcripts expressed per gene [34-36], to the similar number of genes in a simpler organism such as *C. elegans* (~20,000 genes), where there are 1 to 2 protein coding transcripts expressed per gene [37].

In the case of prokaryotes (bacteria), transcription and translation is much simpler because AS does not take place, each gene contains only one exon and there is no 5' capping and polyadenylation, with the flow of information from gene, to mRNA, to protein being significantly more direct and linear. The differences between the 'central dogma' of a bacterium and a eukaryote are quite apparent (Figure 2.2).

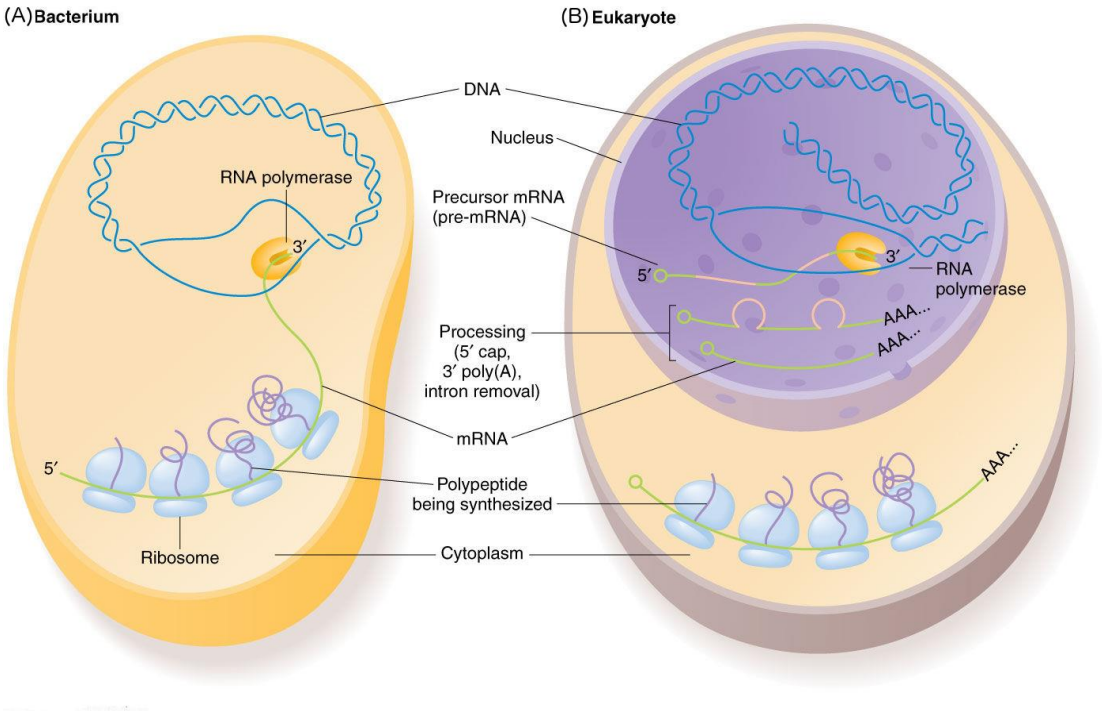


Figure 2.2 Comparative molecular machinery of a eukaryote and prokaryote

The differences between the 'central dogma' of a prokaryote (bacterium) (A) and a eukaryote (B), illustrates the clearer linear relationship between gene and protein product in bacteria, whereas a gene in eukaryotes can represent many multiples of protein products (image courtesy of PJ Russell, iGenetics 3rd edition [38] (http://www.mun.ca/biology/scarr/iGen3_05-09.html)).

The genome is considered a relatively static and unchanging component of an organism (apart from random mutation and somatic recombination of immunoglobulin genes [39]), which comprises all genes (protein and non-protein coding) and regulatory micro RNAs, promoter, suppressor, and enhancer regions. The proteome of an organism on the other hand is very dynamic, as it represents the complement of all protein-coding genes being expressed across different environmental conditions and developmental time-points over the organism's life cycle.

Different fields of research have developed around the understanding of each step of this "central dogma of molecular biology", resulting in the field of genomics, transcriptomics and proteomics, all of which can feed into each other to provide a better understand of the underlying biochemical mechanisms and biological functions of genes. It is the interplay between genomics, transcriptomics and proteomics, and how they can be combined using cutting edge bioinformatics tools to further genomic annotation, which is the focus of this thesis.

2.2 GENOMICS

This section provides an in-depth background on genomics, covering genome sequencing and genome annotation as in the context of proteogenomics.

2.2.1 Genome sequencing

The field of genome sequencing experienced a surge of technological advancement in the late 1990s, with the first genome being sequenced, a bacterium called *Haemophilus influenzae Rd.* [40], which was rapidly succeeded by the sequencing of further bacterial genomes [41, 42], yeast [43], worm [44], fruit fly [45], plant [46], mouse [47], rat [48] and culminated with the sequencing of the complete human genome [35, 49]. Since that time many more organisms have been sequenced.

As of September 2015, 7,256 genome-sequencing projects have been completed and 25,173 have permanent drafts, with 32,633 genomes still pending completion (http://www.genomesonline.org/). These numbers will continue to increase at an exponential rate as more genome sequencing projects begin and as the technologies improve to sequence them.

Sequencing technologies have advanced significantly over the last 38 years (Table 2.1). Frederick Sanger pioneered the first technology in 1977, which used a chain-termination method (Sanger sequencing) and in that same year Walter Gilbert

developed a different technology using chemical modification of DNA sequences and subsequent cleavage at specific bases. Sanger sequencing was ultimately adopted as the universal standard for DNA sequencing due to its ease of use, efficiency, and its capacity to reach sequence lengths of 600 bp with an error rate of only 0.001% and yield of 1.9 - 84 Kbp per run [50]. Many automated Sanger sequencers were introduced over the years to improve the sequencing speed, and ultimately they were used in the sequencing of the human genome [51]. However, it wasn't until 2005 and later that the genome sequencing landscape dramatically changed with the introduction of the 2nd generation of sequencing technology, termed "next generation sequencing" (NGS) technologies. The first of the NGS technologies to arrive was 454 in 2005 (now Roche 454), which was initially capable of reaching read lengths of 100 - 150 bp and then later upwards of 700 bp with an error rate of 0.1%, but with higher error rates observed with polybases of over 6 bp and with a yield of 0.7 Gbp per run [50]. The technology is based on pyrosequencing, which detects the pyrophosphate released (detected as light) during nucleotide incorporation, as opposed to Sanger sequencing which uses dideoxynucleotides to terminate chain amplification. The following year saw the release of the Genome Analyzer (GA) by Solexa (now Illumina GA/HiSeq/MiSeq) with a large improvement in throughput and yield compared to previous technologies, capable of reaching read lengths of 150 bp with <2% error rate and a yield of 30 Gbp per run [50, 52]. The technology uses a technique termed sequence by synthesis (SBS), which uses a library with fixed adapters, denatures them into single strands and attaches them to a flowcell, where bridge amplification occurs to create clusters containing clonal DNA fragments. The library is then spliced into single fragments using a linearization enzyme, dideoxynucleotides (ddATP, ddGTP, ddCTP, ddTTP) containing different cleavable fluorescent dyes and a removable blocking group which anneals to a template one base at a time, detected by a charge-coupled device (CCD). The same year also saw

the development of SOLiD (Sequencing by Oligo Ligation Detection) (now Applied Biosystems (AB) SOLiD), initially capable of reaching read lengths of 35 bp and later upwards of 85 bp with a 0.01% error rate and a yield of 120 Gbp per run. The technology uses a technique termed ligation sequencing, which is a two-base sequencing technology. Once the libraries attached to a flowcell they were then sequenced by an 8 base-probe ligation, incorporating a fluorescent dye, which produces a signal when annealing the last few bases to the template. Collectively these technologies, now termed 1st generation NGS technologies have shown improvements over the years in throughput, accuracy and cost compared to the now dated Sanger technologies [50].

A few years later, in 2010 and 2011, saw the arrival of more advanced genome sequencing technologies, which straddled the divide between 2nd and 3rd generation genome sequencing technologies. The Helicos Genetic Analysis System developed by Helicos BioSciences [53] was capable of reaching read lengths of 15 - 50 bp with a variable error rate depending on read length: an error rate of 0.60% was achieved at a read length of 30 bp and an achieved yield of 35 Gbp per run [54]. The other new genome sequencing technology was the Ion Personal Genome Machine (PGM) developed by Illumina and Ion Torrent [55] which was capable of reaching read lengths of 200 bp, with an error rate of 1.71% [52]. Unfortunately, the Helicos Genetic Analysis System was never commercially realised, as Helicos BioSciences was delisted in April 2010.

The two main differences in these technologies compared with the previous NGS technologies were, in the case of Helicos Genetic Analysis System, the use of a form of Single Molecule Sequencing (SMS) called Single Molecule Florescent Sequencing (SMFS), which obviated the need for DNA amplification through the Polymerase Chain Reaction (PCR) along with its inherent errors and biases', and in the

case of Ion PGM, real time signal detection which provided faster sequencing and significantly lower upfront costs [56, 57]. With Ion PGM, each time a nucleotide was incorporated onto the end of a DNA strand during DNA synthesis by DNA polymerase, a proton was released which altered the pH of the solution, and was subsequently detected by a semiconductor [55]. With the Helicos Genetic Analysis System on the other hand, DNA molecules were hybridized to a primer and each time a Virtual Terminator nucleotide [54] with fluorescent dye was incorporated and the complementary strand extended by DNA polymerase an image was taken. The fluorescent dye was then cleaved off followed by the addition of another Virtual terminator nucleotide with fluorescent dye and DNA polymerase and the process would repeat. The process can also be produced in parallel across many different DNA fragments, significantly improving the throughput and overall yield compared to previous NGS technologies [53].

In recent years a number of newer technologies have emerged, termed 3rd generation technologies, which were a technological leap forward compared to all the previously mentioned NGS technologies. These 3rd generation technologies provide both real time sequencing, measuring the incorporation of each nucleotide along the strand one at a time, and single molecule sequencing where DNA amplification through PCR is not needed. In early 2009 Pacific Bioscience (PacBio) first published their work on a single molecule, real time sequencing (SMRT) technology, which could reach read lengths ≥3,000 bp, however these were plagued with error rates of around 13% and the technology could achieve a yield of 100 Mbp per run [52]. The SMRT technology was capable of measuring the incorporation of each fluorescently labelled deoxyribonucleoside triphosphate (dNTP) along a template bound DNA fragment. A SMRT cell comprises millions of 50 nm wells called zero-mode waveguides (ZMWs), consisting of a set of enzymes, a single DNA fragment and a template bound to the

bottom. The ZMW allowed for precise detection of nucleotide incorporation each time a single fluorophore (fluorescent dye) was excited by a laser by preventing no other sources of light into the well. The fluorescent signal was then detected in real-time by a camera. The millions of ZMWs running in parallel allowed for very high throughputs [50, 58, 59].

Another 3rd generation technology, called Nanopore sequencing has been in development for most of a decade [60-62] and has matured with the release of the first working machine developed by Oxford Nanopore Technologies in 2012 [63], capable of reaching read lengths >5,000 bp with speeds of 1 bp per 10 nanoseconds [62]. The technology relies on passing a DNA sequence through a nanopore, a biopore with a diameter on the nanoscale, in which the sequence is then deduced by measuring changes in the ion current as each nucleotide (A, G, C, T) passes through the narrowest constriction of the pore. The technology negates the previous requirements for DNA amplification by PCR, fluorescent labelling and optical measurements. A number of challenges still need to be overcome with this exciting new technology with further likely improvements in the foreseeable future [50, 64].

Table 2.1 Genome sequencing technologies

Sequencing technology	Sequence length (bp)	Sequencing time	Error rate (%)	Yield per run	Cost per Mbp*	Year of emergence
Sanger	600	20 min - 3 hours	0.001	1.9-84 Kbp	$2400	1977
454	100-700	10 hours	0.1	0.7 Gbp	$13	2005
Illumina Genome Analyzer	150	10 days	<2.0	30 Gbp	$0.02	2006
SOLiD	85	7 days	0.01	120 Gbp	$0.04	2006
Helicos	15-50	5 hours - 1 day	0.60	35 Gbp	NA	2011-2012
Ion Torrent	100	2 hours	1.70	0.02 – 1 Gbp (depends on chip format)	$1	2011-2012
PacBio	≥3,000	2 hours	13.0	0.1 Gbp	$2	2009
Nanopore	≥5,000	1 bp per 10 nanoseconds	NA	NA	NA	2010-2012

* All costs based on sequencer manufacturer and sequencing provider [50, 52].

NA: Not available

In recent years, the throughput of next generation sequencing has surpassed the speed and cost-benefit barrier, with the sequencing of more than 2 human genomes per day, during the 1,000 Genome Project, which sequenced and compared 1,092 human genomes [65].

Together, these technologies highlight the incredible rate at which genome sequencing is progressing. With this progress in mind, the question then needs to be posed as to how to process and understand the level of information produced.

2.2.2 Genome annotation

The speed at which genomes are now being sequenced and their reduced costs poses a problem of how to deal with the large quantities of data being generated. Genomic assembly and annotation significantly lags behind the progress that the field of genome sequencing has obtained. Many genomes over the last decade have been assembled from reads generated using 2nd generation technology, using shorter reads (e.g. Illumina), often leading to highly fragmented assemblies, particularly in genomic regions containing large numbers of repeat elements [66] or high/low GC content due to PCR bias [67, 68]. Depending on the size of the genome, this can result in a highly fragmented genome, and in the case of *de novo* assembly, chimeric scaffolds and contigs can form where the *de novo* algorithm used for the assembly is unable to resolve large repeat regions when the read length is equal to or less than the repeat length [69, 70]. In recent studies it has been shown that using mate-pair information (where a reads location is known in relation to another read), can be used to improve the assemblies and resolve repeat regions (as long as the insert size between reads is longer than the repeat region) [71], and in addition a concept called *path extension* is able to resolve longer repeat regions where the insert sizes are inadequate [72]. Fortunately, smarter ways to use short reads is not the only solution, as many of these problems will disappear in the coming years as the 3rd generation technologies mature, with longer

reads being obtainable (easier to assemble), and with no PCR bias (see previous Section 2.2.1).

As a consequence from the use of shorter sequence reads over the last decade and with many larger genomes being sequenced and assembled, genomic annotation, which is the act of finding the locations of genes and annotating them with structural and biological functional information, has now become more difficult than it has been in the past [73]. This is mainly due to the highly fragmented genomes that are now being assembled, as a result of short reads and large repeat regions in some genomes, such as with the early genome projects, Drosophila melanogaster [74, 75] and the human genome project [35, 76]. Additionally, gene finding and annotation isn't a clear-cut process, and can often lead to false positives, misannotations or completely missed genes [77-84]. In addition, our understanding of what defines a gene is not clear-cut, especially since the conclusion on the ENCODE project [30] where it is now understood that the boundaries of a gene are poorly defined due to overlapping protein products [85].

The statement that a genome is "complete" is often misleading, when there is no certainty that a genome annotation is truly complete in its entirety. This is certainly more true with eukaryotes, due to their complex gene structure as a result of AS. Consequently, misannotations are more often found in bacterial genomes where the error can be more readily detected. Additionally, misannotations can propagate errors in the annotations of other genomes in comparative genomics studies. Using comparative genomics to assign annotations to other genomes should be undertaken with caution, as errors can be introduced from sequencing errors, alignment errors, and changes in nucleotide function and differing biological functions between features such as splice donor sites. Iteratively using more genomes in such as study can lead to more noise from such errors [86], transferring the error to other genome annotations, termed

'transitive disaster'. Therefore, it is best to use at best only a few closely related genomes, or one that is much more highly conserved, with the genome of interest.

To correct all these genomic misannotations and prevent the scenarios as outlined above would be a very large undertaking, incurring high costs and time involved. Better genomic annotation practices and standards would result [87-89], but the end-product could not completely ensure further errors would not occur, or assist in speeding up the annotation process. Therefore, a highly accurate, faster and automated approach is needed.

A number of gene prediction tools have been developed over the years, which apply various approaches, classified as: '*ab initio*', 'combiner', and 'similarity and homology'. *Ab initio* gene prediction tools identify genes by following established rules, such as imposed limits on gene size, GC content, transcription start and stop sites and also apply mathematical models. A number of approaches include the Generalised Hidden-Markov Models (GHMM) [90, 91], machine-learning techniques such as support vector machines (SVMs) [92], which were one of the first types of gene prediction tools made available [93-96] and a more recent mathematical approach called conditional random field (CRF) [97-100].

The *ab initio* tools can provide a fast and easy means to obtain gene predictions and their exon-intron structures without necessarily needing any external evidence, apart from the use of training parameters to improve the prediction for each specific genome, like codon frequency and the distributions of exon-intron lengths, which can often be obtained from closely related genomes if available. *Ab initio* gene predictors however only predict the most likely locations of coding DNA sequences (CDS) and are unable to detect the locations of untranslated regions (UTRs), sites of AS and they have limited accuracy with exon-intron boundaries. A number of *ab initio* gene prediction

tools also include evidence-driven predictions, such as outlined in Table 2.2. By providing weighted (i.e. ranking some evidence higher than others) external sources of evidence to the *ab initio* prediction tools, such as sequence alignments from expressed sequence tags (ESTs), protein alignments, and RNA-seq alignments, the accuracy of the *ab initio* predictions can be improved. Unfortunately, this requires a great deal of work and pre-processing of the data, such as alignment to genomic regions, before being presented to the *ab initio* gene prediction tool. The 'combiner'-based gene prediction tools simply take evidence from multiple different sources, including predictions from other prediction tools, and select a prediction based on a consensus of intron-exon structure. Examples of such tools are outlined in Table 2.2. The 'similarity and homology' based gene prediction tools, as the name suggests use similarity and homology sequence alignments with known gene sequences from other closely related genomic regions, preferentially between conserved genomic regions, potentially identifying genes with identical function with the aim of predicting the sites of genes in the genome of interest. Examples of such tools are outlined in Table 2.2, all of which can also be used to generate evidence for the previously mentioned evidence-driven *ab initio* gene prediction tools.

Table 2.2 Types of gene prediction tools and their employed method of prediction

Gene prediction method	Gene prediction tool	Reference(s)
Ab initio **(evidence-driven)**	Augustus	[101, 102]
	TwinScan	[103]
	FGENESH	[104]
	Gnomon	[105]
	SNAP	[106]
	mGene	[107]
Combiner	JIGSAW, and	[108]
	EVidenceModeler (EVM)	[109]
	GAZE	[110]
	Genomix	[111]
	GLEAN	[112]
	Evigan	[113]
	EGPred	[114]
Similarity and homology	SIM4	[115]
	EST2Genome	[116]
	Procrustes	[117]
	Spidey	[118]
	GeneSeqer	[119]
	AGenDA	[120]
	BLAST	[121]
	BLAT	[122]
	GeneWise	[123, 124]
	Exonerate	[125]
	Bowtie2-TopHat2	[126]
	STAR	[127]
	GIGOgene	[128]
	GMAP	[129]
	GSNAP	[130]

Many of the above mentioned gene prediction tools are often incorporated into genome annotation pipelines, which can be broken down into a number of different types, based on the level of automation: manual, semi-automated, automated and high-throughput automated. The manual approaches tend towards being highly accurate, costly and slow processes, and by contrast, as the annotation pipeline becomes further automated with a magnitude of throughput increase, there is often a trade off in the accuracy of the annotation, depending on how the annotation pipeline is automated and the level of evidence supplied or lack thereof.

Sometimes multiple gene predictions tools can be used together, building on each of their strengths, with the pipeline taking the consensus predictions, which are supported by external evidence from genomic alignments from RNA-seq, ESTs and proteins. Other approaches such as combiners, as mentioned above, incorporate multiple gene predictions.

A number of well-known manual, semi-automated, automated and high-throughput automated genome annotation pipelines, which utilise many of the above gene prediction strategies, are pipelines such as those listed in Table 2.3, some of which target specific classes of genomes, such as bacterial (IMG, Prokka and RAST) or larger complex eukaryotic genomes (NCBI, Ensembl and MIPS), and specifically tailored for niche genomes such as large complex plant genomes (TriAnnot and PlandSEED). Other genome annotation pipelines are manual efforts, like the HAVANA group who curated the human genome and who make their annotations available through the Vertebrate Genome Annotation Database (VEGA) [131].

The outputs from these pipelines can be used to train and improve the accuracy of gene prediction tools, which themselves can be re-used in the pipeline. For example, Maker2 can streamline gene prediction tool training, allowing for easy incorporation into tools such as Augustus or SNAP. A comprehensive overview of genome annotation can be found in [73] and [86].

Table 2.3 Genome annotation pipelines

Level of automation	Annotation pipeline	References
Semi-Automated	MIPS	[132]
Semi-Automated	IMG	[133]
Automated	RAST	[134]
High throughput automated	TriAnnot	[135]
High throughput automated	PlantSEED	[136]
High throughput automated	Maker2	[137]
High throughput automated	Prokka	[138]
Automated	AutoFACT	[139]
Automated	PASA	[140]
Automated	Ensembl	[141]
Automated	NCBI	[142]
Manual	HAVANA	[143]

Because the genome annotation process often uses *ab initio* approaches, based on pre-defined rules, and consequently due to a limited understanding of gene structure, which can vary widely between organisms, such approaches are prone to errors and inconsistencies [144]. Even in cases where the evidence is available it may be incomplete, such as with transcript-based annotation with un-spliced mRNA or nonsense transcripts. The few errors that can be detected are painstakingly corrected through manual annotation. As genomic sequencing costs reduce and the rates at which they are sequenced increases, the use of a number of genome annotation approaches becomes impractical. Therefore, there is now an urgency to produce high throughput automated annotation pipelines of higher quality and throughput than before, accepting multiple lines of evidence before an annotation is considered "complete" and significantly reducing the amount of time and manual annotation required to still achieve a high quality level of curation.

2.3 MASS SPECTROMETRY-BASED PROTEOMICS

This section provides an in-depth background on MS-based proteomics, covering current trends in mass spectrometry technology, proteomics techniques and algorithmic approaches to interpret and statistically validate MS/MS spectra. In relevance to proteogenomics, of particular interest will be given towards the types of mass

spectrometers currently employed in proteomics, as well as the database searching methods used in MS/MS spectral interpretation and statistical validation; as such strategies are employed throughout this thesis.

The field of proteomics is focused primarily on the identification of proteins, the identification of signal pathways, cellular localization, quantification, their interaction in protein-protein networks and complexes and to understand the roles that post-translational modifications (PTMs) play in these complex networks [145].

Mass spectrometry (MS) has been continually improving in throughput, specificity, sensitivity, and dynamic range [146] for well over a century since its inception, and it has been increasingly used for proteomics, to quantify and identify proteins [147-151]. Mass spectrometry has also benefitted greatly from improvements in sample handling (generally a clean, dust-free work environment and good sterile techniques) and separation techniques (forms of Liquid Chromatography (LC)) and the high performance of various MS instruments.

There are a wide variety of mass spectrometers available, ranging in sensitivity and throughput. The most commonly and currently used MS instruments in the field of proteomics are ion trap mass spectrometers [152]. These mass spectrometers have high data acquisition rates and are known to generate enormous amounts of data of low resolution and of low mass accuracy, which affects the level of confidence assigned to peptide sequences, with just 10-15% of peptide assignments being regarded as correct [153, 154]. An example of a mass spectrometer and its general workflow can be seen in Figure 2.3.

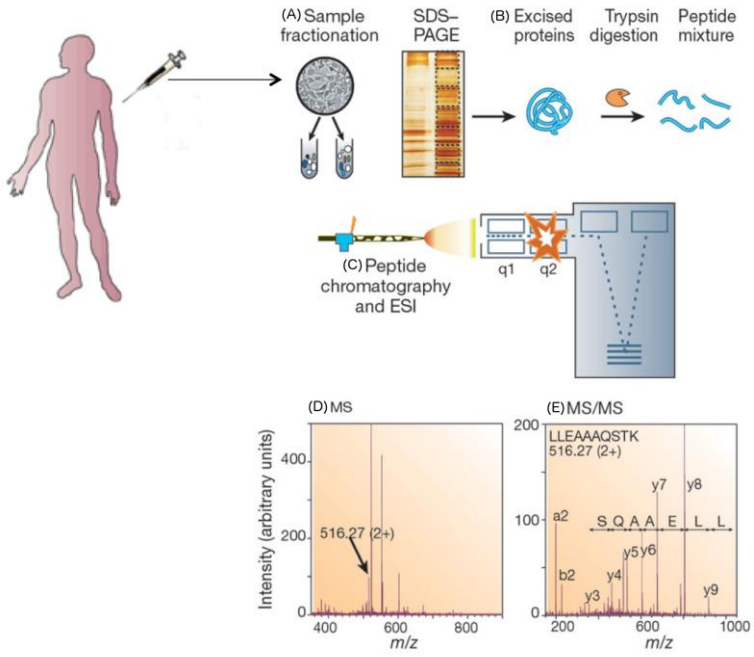


Figure 2.3 General workflow of MS-based proteomics

The workflow applied in MS-based proteomics, in this example, with a Quadrupole Time-of-Flight (QTOF) mass spectrometer, generally takes the form of: (A) preparing the sample and forming technical replicates, running on a SDS-page gel (or taking whole cell lysates); (B) digesting the proteins with a protease such as trypsin; (C) followed by injection of the sample onto an LC column followed with electro-spray ionisation (ESI) (alternatives being MALDI, with ionisation in a matrix compound), with MS1 taking place in the first quadrupole (q1), and MS2 (MS/MS) taking place in the second quadrupole (q2), entering the TOF chamber, followed by; (D) detection of the precursor peptide ion mass and; (E) fragmented ions masses, where *de novo* sequencing can take place (modified from [147]).

A more advanced range of mass spectrometers called hybrid-fourier transform (hybrid-FT) mass spectrometers [155, 156] are capable of high mass resolutions of 30-500 kilodaltons (kDa) and very high mass accuracy within a few parts per million (ppm). The throughput and sensitivity is maximized on these mass spectrometers by collecting MS data at a higher resolution and accuracy, and recording the MS/MS data at a higher speed, but at a lower resolution and accuracy [157]. A higher resolution of MS/MS spectra allows for precursor mass ions to have their charge states determined [158, 159]. This allows for the detection of higher mass ranges at higher charge states and lower cost, with each mass peak having a m/z (mass to charge) ratio, and with a higher mass accuracy, coupled with stricter mass tolerance windows allows search algorithms to narrow down the number of possible peptide candidates, consequently improving the confidence level for peptide matches [160, 161].

2.3.1 Bottom-up proteomics and strategies

In an MS-based proteomics experiment a number of approaches can be applied depending on the desired outcomes. In one approach a protein mixture is first fractionated via two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), which separates the proteins in two dimensions, by their isoelectric point (pI) and molecular mass (MW) [162]. What follows is termed bottom-up proteomics, whereby each spot on the gel is excised and the proteins within are then enzymatically digested by a protease such as trypsin, chymotrypsin, thermolysin or endopeptidase V8 (Glu-C), each of which cleaves proteins at various specific sites. Trypsin is used in most cases, producing peptides with a median length of 15 residues with a basic C-terminal end that can trap a charge [163] and is approximately 1,600 Daltons (Da) in size, which is well within the detectable mass range of most mass spectrometers.

Once the peptides are ready for analysis, there are a number of options from which to choose. Often if all that is desired is protein identification, a technique called peptide mass fingerprinting (PMF) can be applied which looks at intact peptides. The common approach is to use a time-of-flight mass spectrometer (TOF-MS), which can have a High Performance Liquid Chromatography (HPLC) setup with what is termed electrospray ionization (ESI). This approach separates peptides according to their molecular weight (MW) with a C_{18} column before the peptides elute off and are ionized in a fine spray with an applied electric charge, before entering into a gas phase and passing into the mass spectrometer. Alternatively, the TOF-MS can use a Matrix-assisted laser desorption/ionization (MALDI) setup. Following the extraction of digested peptides from the gel pieces, a compound called a matrix is applied to the sample, which assists with ionization. This method utilizes a laser to ionize the peptides, which then enter the mass spectrometer for analysis. Regardless of the ionization method used, the samples now entering the TOF-MS fly to the end of a long cylinder,

assisted by large electric fields within a vacuum. The time taken to reach the detector and the electric field applied is proportional to the mass of each ionized peptide. The larger the electric field, the larger the flight path of each ionized peptide. This essentially increases the resolving power of the machine. The process is analogous to the length of a C_{18} column in HPLC: the longer the column, the greater the resolution of each mass peak. A similar approach is illustrated in Figure 2.3, applying the use of a TOF, however in PMF only the detection of MS1 is carried out with intact peptide ions detected.

The peptide masses obtained from PMF can then be matched to a database with a list of peptide masses from known proteins, using search tools such as MOWSE [164], Mascot [165], MS-Fit [166], PeptIdent [167], ProFound [168] and Aldente [169]. These peptide masses are calculated from protein sequences, digested *in silico* using a specific software program, which cuts the protein sequences at specific cleavage sites based on the known specificity of different proteases such as trypsin. However, the actual sequences of the peptides in the sample themselves remain unknown.

Apart from protein identification by PMF, other applications include the varietal typing of wheat and barley using pattern matching techniques, where the distinct peaks from a PMF between a number of different varieties can readily be identified [170, 171]. Other applications include a similar concept, using peptide mass values from many known samples to train a naïve Bayes classification algorithm, to classify unknown samples [172], an analogous concept to biomarker discovery in medicine [173], among other similar studies using bottom-up proteomics for understanding disease [174-176].

The use of a 2D gel at the initial stages, followed by HPLC or MALDI, is effective in reducing the sample complexity and purifying the proteins. This method can

be relatively low in throughput and the coverage of the resulting proteome also is relatively poor. However, the MADLI-TOF stage itself can be quite sensitive and of high-throughput, but the use of 2D gel and peptide separation significantly reduces the overall proteome coverage and throughput. To overcome this limitation another method is available called Multidimensional Protein Identification Technology (MudPIT) [151], which makes use of multi-dimensional HPLC and replaces the use of a 2D gel. The technique combined with another approach called shotgun proteomics, where whole protein samples are first digested by a protease, can significantly improve proteome coverage and throughput. Shotgun proteomics is not unlike shotgun genome sequencing, in that fragments of the original sequence are created and then reassembled. Following protease digestion of the protein the peptide mixture is then separated by multidimensional HPLC (2D HPLC), where the peptides are separated by their pI and MW using a HPLC column packed with Strong Cation Exchange (SCX) resin and C_{18} Reverse Phase (RP) material [151, 154]. The peptides are then eluted off the column and ionized, usually by ESI, passed through the mass spectrometer and subsequently analysed by the detector.

Following 2D HPLC in a MudPIT setup, the mass spectrometer normally used is a tandem mass spectrometer (MS2 or MS/MS), which can be designed in many formats including, but are not limited to quadrupole TOF (QTOF) (Figure 2.3) and quadrupole ion trap (QIT), as well as variations on these with additional mass analysers above the MS2. For example, TripleTOF or triple quadrupole (QqQ) also referred to as MS3 or MS/MS/MS and provides further separation of peptide fragment ions and higher resolution.

Such tandem mass spectrometers are becoming a common method for identifying peptides and proteins with high sensitivity, specificity and high throughput [177]. Through this method, the precursor ions obtained by the first pass of MS (MS1)

at low energy collision-induced dissociation (CID) are fragmented, most commonly along the peptide bonds during the second pass of MS (MS2, MS/MS). The resulting fragments of a peptide are measured as a mass over charge ratio (m/z), which mirrors the overall structure of the peptide ion [178, 179]. According to Roepstorff's nomenclature [179], within the MS/MS spectra at around 10-50 electron volts (eV) collision energy the peptide ions are denoted as *a*, *b* and *c* when the charge is on the N-terminal side of the fragmented peptide, and *x*, *y*, *z* when the charge is on the C-terminal side. When the collision energies involved in fragmentation are of orders of magnitude greater than around 1 kilo electron volt (keV), the peptide side-chains are broken generating side-chain ions, denoted as *d*, *v* and *w*, which can be formed by the loss of some or all side-chains [180].

The sequence of the fragmented peptides can be calculated through *de novo* sequencing, which uses first principles to determine the amino acid sequence of a protein from the MS/MS spectra by looking at the mass differences between peaks [181, 182]. This was previously conducted manually using a technique called Edman degradation [183], requiring chemical labelling of amino acids and the cleavage of each amino acid in succession to form amino acid derivatives, followed by electrophoresis to identify the amino acids. The procedure was expensive and slow by today's standards using mass spectrometry-based technology which can sequence many hundreds of thousands of peptide sequences simultaneously, mapping out the entire proteome within a fraction of the time. In addition, the MS/MS spectra can also be matched directly to known sequences through the use of MS/MS database search algorithms, discussed in detail later on.

A number of alternatives to the above mentioned MudPIT setup using 2D HPLC with SCX and RP, can include other alternative fractionation methods, such as size exclusion liquid chromatograph (SEC) [184, 185], capillary electrophoresis (CE) [186,

187], capillary isoelectric focusing (CIEF) [188], and gel-based isoelectric focusing [189] in the 1st dimension, followed by RP chromatography in the 2nd dimension. Another approach called gas-phase fractionation (GPF) [190, 191] can also be applied which uses the mass spectrometer to resolve and separate out the different precursor ions over a set number of *m/z* ranges before fragmentation by mass [192, 193] and ion intensity (a measure of the number of detected fragments) [194, 195], using iteration from different fractions of the same sample by RP chromatography. The proteome coverage obtainable with this approach is limited to the number of fractions (*m/z* ranges) applied and the amount of sample available.

A cost-effective approach to improving proteome coverage is to use a range of proteases to digest the sample. In tandem mass spectrometry, peptides of around 7-35 amino acids (aa), protonated, low charge state (z) and high mass-to-charge ratio (m/z) are ideal. Trypsin meets many of these requirements and so has been successful as a protease, becoming the common choice. Trypsin cleaves specifically after arginine (R) and lysine (K), but not before proline [196] although an absence of proline has recently been identified [197]. However, trypsin does have some disadvantages, such as autolysis in alkaline pH, requiring that lysine and arginine to be evenly distributed across a proteome, which is often not the case, and its thermostability is poor. The use of other proteases in conjunction with trypsin would ensure much better coverage of the proteome [198].

Another approach to improve proteome coverage can be found in the way in which mass spectrometry works at the level of precursor ion fragmentation. Historically, the majority of MS-based proteomics work has used one fragmentation mode, namely CID to derive the MS/MS spectra due to the limitations of the technology at the time, and subsequently many MS/MS search tools were optimized solely for CID MS/MS spectra. A number of alternative fragmentation methods are now becoming

utilized by newer mass spectrometers, such as electron transfer dissociation (ETD), higher-energy collisional dissociation (HCD), CID/ETD or HCD/ETD paired MS/MS spectra and the less frequently used electron capture dissociation (ECD), used only by a few mass spectrometers. The quality and usability of MS/MS spectra depends on the fragmentation method used for each precursor ion, which is highly dependent on various properties; for example, ETD is more suitable for precursor charge states of >2 [199-202], acidity due to PTMs such as phosphorylation [203-205], and is more suitable for *de novo* sequencing due to its high levels of fragmentation [206, 207]. Since the introduction of hybrid MS/MS machines (QTOF, QIT, TripleTOF etc), which can implement a number of fragmentation methods in parallel, an opportunity has arisen to use these different methods in a complementary way. The advantages and disadvantages of using various fragmentation methods on different precursor ions have been thoroughly investigated by Frese et al. 2011 [208]. The combination of these various methods has been shown particularly with CID and ETD, which could be used to significantly improve proteome sequence coverage [201-203, 205, 206, 209-211]. Another approach which leverages mixed CID/ETD MS/MS spectra uses data-dependent decision tree logic during MS analysis and before MS/MS, while the mass spectrometer is running, to determine which fragmentation method is the most appropriate to use based on the precursor ion charge state (z) and m/z value. This method significantly improves the number of peptide identifications significantly over the use of single fragmentation methods alone [206].

Shotgun proteomics, which employs a bottom-up proteomics approach, has often used a data-dependent acquisition (DDA) method for the selection of precursor and fragment ions for analysis. Briefly, the DDA approach scans for precursor ions above certain thresholds for intensity, charge state etc at MS1. The selected precursor ions are then sequenced by product ion fragmentation at MS2. Through shotgun

proteomics, enough of the sample can be supplied for analysis to improve the number of precursor ions passing the selected thresholds, thus improving the overall proteome coverage. However this is not ideal and the approach has a number of drawbacks, including slow speed [212], random selection of precursor ions, and poor reproducibility [213], narrow dynamic range [214], problems with mixture or multiplexed MS/MS spectra (co-fragmentation of ≥2 precursor ions within the same MS/MS spectra) [215, 216], low reproducibility and others.

Mass spectrometry has allowed improvements in throughput, the identification of proteins, and reliability of inferring protein expression levels, however achieving proteome-wide coverage has remained a problem [146, 217]. This is due to different protein extraction techniques and the varying conditions under which proteins are expressed, as well as the difficulty of detecting lowly expressed proteins in a sea of highly abundant proteins.

One of the more pressing limitations of bottom-up proteomics in terms of protein discovery and identification is the ambiguity present with the assignment of peptides to proteins, because many proteins share peptide sequences, known as the protein inference problem [218-220]. Due to the digestion and fragmentation of intact proteins, reassignment back to the proteins becomes problematic if the coverage for a particular protein is limited, especially when databases contain multiple isoforms of the same protein, or particular peptide sequences are common within protein families. This often requires identification of unique peptides or proteotypic peptides, which appear in no other isoform to identify the correct protein. However, as is often the case, not every protein isoform has an identified unique or proteotypic peptide, limiting its application across the whole proteome.

2.3.2 Top-down proteomics: A complement to bottom-up proteomics

A significant development in the field of proteomics, which circumvents the inherent problems in bottom-up proteomics, such as the protein inference problem mentioned above, is top-down MS proteomics. Top-down proteomics is the MS analysis of intact proteins and large peptides [221-232], which is appropriate for the localization of multiple PTMs across the entire length of a protein, and the identification of multiple proteoforms [233], and is highly suitable for proteotyping of diseases in medicine. However, due to the very nature of some proteins that are insoluble, complete analysis of all proteins across the whole proteome is not possible. By contrast, traditional bottom-up approaches can only provide a limited "fragmented" view across the whole proteome, on average, 25% of the full length protein [146], indicating that in a bottom-up experiment many PTMs are not detected but many more peptides overall can be identified due to the higher solubility of smaller peptides and their propensity to ionise. The differences between bottom-up and top-down proteomics is illustrated in Figure 2.4 (modified from [234]).

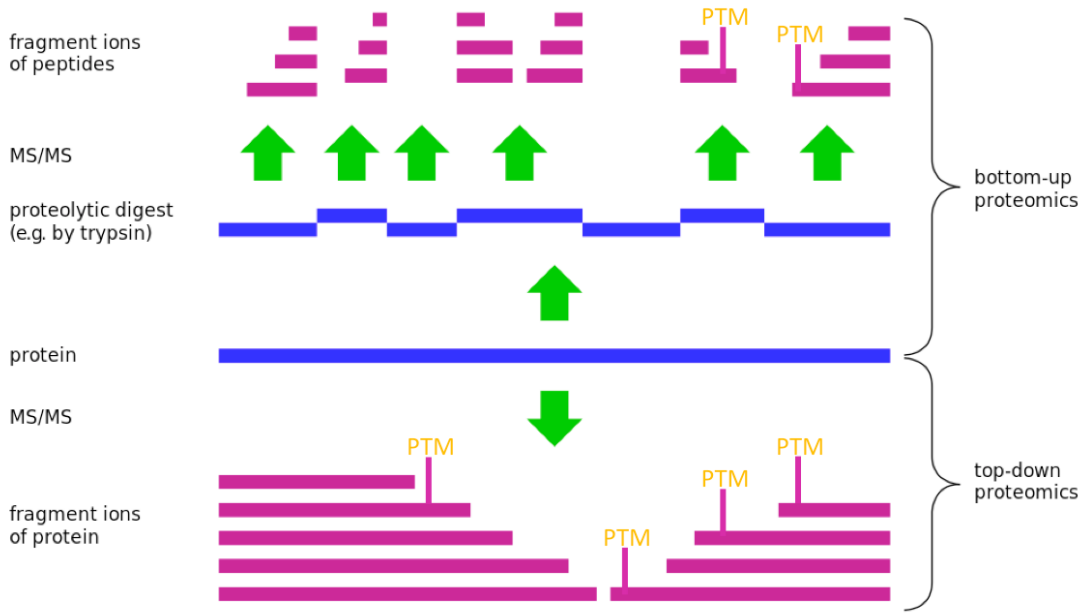


Figure 2.4 Top-down versus Bottom-up proteomics

Fragmentation of proteins in top-down proteomics shows more complete coverage of a single protein, and unambiguous identification, including more identified PTMs, but lacks coverage across the whole proteome. By comparison, bottom-up proteomics shows limited coverage and ambiguous identification of a single protein, identifying only some PTMs, and has much wider coverage of the whole proteome (modified from [234]).

Although top-down proteomics is powerful, it has been less commonly used. Top-down mass spectrometry has generally been limited with separation strategies for whole intact proteins (restricting its capability to single purified proteins), fragmentation, computational tools and with upper limits on the detectable protein mass with significant impacts on throughput [235-239]. However, top-down proteomics can now allow for the analysis of complex mixtures of proteins consisting of many hundreds to thousands of proteins, utilizing recent advances in separation techniques, fragmentation and improved detection limits (<60 kDa) of the MS instrumentation [232, 240-245]. To overcome the restrictions current mass spectrometry technologies have imposed on the size limits of top-down proteomics, a technique called middle-down proteomics was applied, with large peptides generated (<6 kDa) from target specific proteases, followed by analysis with the normal top-down approach. This has the advantage of identifying proteoforms larger than the top-down limit of ~60 kDa, while gaining the sensitivity of bottom-up proteomics [246]. These rapidly changing trends in

top-down proteomics have been outlined in a comprehensive review [233] and in a clinical setting [247].

Top-down proteomics is less sensitive than bottom-up proteomics, due to its inherent insolubility and ionization problems. As both approaches further improve in terms of throughput, they will ultimately be used as standard in a synergistic manner. The use of both approaches can also allow the testing of hypotheses such as the location of the N-terminal end of proteins, the presence of N-terminal methionine excision (NME) [248] and signal peptide cleavage. In addition, in the future there will likely be a merger of multiple approaches, from peptides, whole intact-proteins and whole protein complexes [249].

2.3.3 Next generation bottom-up proteomics

Over the last decade there have been substantial improvements in mass spectrometry. A number of recently developed techniques have overcome the drawbacks of the traditional DDA approaches used in bottom-up proteomics by taking a broader sweep of all peptide ions in a sample, thus providing much greater proteome coverage, by selecting all precursor masses at MS1, and by extension all possible fragment ions at MS2. This approach is known as data-independent acquisition (DIA), originally developed in [250], and subsequently improved [251, 252]. A study demonstrating the advantages of DIA over DDA was undertaken [253]. The DIA approach allows for MS analysis of all precursor ions (no selection process based on intensity etc) in a complex mixture of peptides with no loss in detection, resulting in much higher sequence coverage, with fewer samples being required and with improved reproducibility compared to DDA.

A number of different approaches based on DIA have emerged, which include MS^E [254-258], all-ion fragmentation (AIF) [259], Fourier transform-all reaction

monitoring (FT-ARM) [260], Sequential Window Acquisition of all THeoretical Mass Spectra (SWATH-MS) [261-264], and improvements to this method utilizing multiplexed MS/MS [265], to resolve MS/MS spectra containing multiple precursor ions and spectra. All of these methods have their individual advantages and disadvantages, and by comparison with DDA. In addition, Precursor Acquisition Independent From Ion Count (PAcIFIC) [266, 267] is a DIA method that applies small precursor windows across a small mass range requiring multiple injections of sample, unlike the other DIA methods mentioned. The main advantage over DDA is improved proteome coverage and efficiency of sampling the proteome, with the ability to examine less abundant proteins. This advantage is, however, also a disadvantage, in that the MS/MS spectra are highly complex and noisy, requiring significant pre-processing, and since the number of MS/MS spectra is so large, analysing the data becomes a computational challenge. The large amount of data also brings up the issue of storage, although an advantage of this issue is that it allows further interpretation at a later date without having to re-sample and begin a new bottom-up experiment. Comprehensive reviews of the different DIA approaches to handle multiplexed and non-multiplexed MS/MS spectra can be seen in [268, 269]. A similar problem to multiplexed MS/MS spectra is mixture MS/MS spectra. It is often assumed that a single spectrum contains a single peptide, but in reality there can be multiple peptides residing in a single spectrum, which have the same precursor ion. A number of approaches have been developed over the years to identify multiple peptides from a single spectrum, including ProbIDTree [270], MSPLIT [271], MixDB [272] and more recently MixGF [273].

Approaches to addressing the problem of mixture and multiplexed MS/MS spectra will gradually become more mainstream as the tools improve, with an eventual integration of DIA with both top-down and bottom-up proteomics, thus opening up MS-based proteomics to a whole new level and effectively putting it on the same playing

field as the latest next generation genome sequencing technologies in terms of throughput and coverage.

2.3.4 MS/MS spectral identification strategies

Over the last 20 years MS/MS search tools have evolved along with the capabilities of the mass spectrometers at the time. The vast majority of MS/MS search approaches have used CID MS/MS spectra, with data obtained from DDA mass spectrometry approaches. Search strategies have since been extended to include other fragmentation types, such as ETD and HCD. Search tools have also been developed for MS/MS spectra acquired from DIA and also top-down proteomics.

There are currently a large number of MS/MS identification tools available. These tools can be based either on: 1) identification and scoring of peptide-spectrum matches (PSMs) through basic database searches; 2) directly identifying sequences by *de novo* sequencing of the MS/MS spectra; 3) a hybrid *de novo* sequence/database search approach; or 4) spectral library searching which identifies MS/MS spectra from matches to previously identified MS/MS spectra.

Database searching and scoring of PSMs has been the most widely used and applied method to date for identification of known and novel MS/MS spectra, and therefore greater emphasis will be placed on this technique for the remainder of the chapter. The methods employed to assign significance or confidence to any matches is, however, ubiquitous across many methods.

2.3.5 MS/MS spectral data formats

MS/MS spectra in the majority of cases are generated by mass spectrometers in a number of different proprietary formats depending on the manufacturer and their proprietary software package. To be able to access and mine the MS/MS spectra further on a variety of different pre-processing tools and search algorithms, conversion of the

proprietary MS/MS spectra to an open format is required. The first two open formats to be used in MS-based proteomics were: 1) JCAMP-DX [274], an old format used with early small-scale MS/MS spectra and so was unsuitable for today's large MS/MS spectral datasets; and 2) the ANalytical Data Interchange Mass Spectrometry (ANDI-MS) [275] format which was based on the Network Common Data Form (NetCDF) [276] library used for reading and writing common array-oriented scientific data for sharing. These older formats were used in the early days of proteomics until the arrival of the xml formats. The first xml format to be adopted was mzData [277, 278], developed by the Proteomics Standard Initiative (PSI) from the Human Proteome Organisation (HUPO), which was tasked with creating a standard format for the proteomics community and later was superseded by the mzML and mzXML formats. The mzXML format was developed by the Institute for Systems Biology - Seattle Proteome Centre (ISB-SPC) contemporary with the development of mzXML by HUPO-PSI. The mzXML format was based on the eXtensible Markup Language (XML) [279, 280], and it was later developed as a joint project between the ISB-SPC and HUPO-PSI by combining the concepts of mzData and mzXML [281, 282]. Other popular data formats include Mascot Generic Format (MGF), DTA (Sequest associated format), PKL (Micromass) and MS2 [283].

There are a number of tools capable of converting the proprietary formats to open formats to allow further analysis, such as ReAdW [284], mzWiff [285], and most recently the widely used tool msconvert [286], a part of the Proteowizard tool-kit [287] which is now standard with many proteomics analysis pipelines, such as the Trans-Proteomics Pipeline (TPP) [288]. There have also been a number of new application programming interfaces (APIs) to allow the proteomics community to use global standard approaches towards the reading and writing of these data formats, to allow compatibility of input and output data across different groups, and to allow ease of

sharing data and comparing research results. An example of such an API is the jmzReader Java API to allow parsing of numerous mass spectrometry data formats [289], the jmzML Java API for parsing mzML formats [290], and in addition a new format from HUPO-PSI called mzIdentML which is written and read using jmzIdentML Java API [291]. MS-GF+ [292] is an example of a search tool already implementing the jmzIdentML Java API [292].

2.3.6 MS/MS spectra pre-processing

To improve the confidence of MS/MS spectral identification, it is recommended that the MS/MS spectra go through pre-processing steps to improve their overall quality for database searching. This is due to a variety of reasons including MS/MS spectral noise, which constitutes three types: random noise, chemical noise and non-protein contaminant noise. MS/MS spectra can also be of poor quality due to incomplete fragmentation patterns and low number and/or intensity of peaks, large redundancies within the MS/MS spectra where there are multiple MS/MS spectra for one peptide, the presence of multiple mass peak isotopes (isotopomer envelopes) and multiple different charge state variations amongst the peptide ions in the analysed sample. There are a number of different approaches to address these issues, such as deisotoping, which selects the most representative peak from different isotopes and charge state deconvolution, which is the determination of the mass at a neutral charge state. Additionally, the quality of the MS/MS spectra can be improved by removing noisy MS/MS spectra, improving the signal to noise ratio, and also clustering the MS/MS spectra. Clustering MS/MS spectra merges similar fragmented MS/MS spectra and has the effect of improving the signal to noise ratio, reducing the MS/MS spectral dataset size, improving database search speed and reducing the occurrences of false positives. Overall, depending on the MS/MS spectral dataset and the type of mass spectrometer the MS/MS spectra were generated on, all or a number of these approaches can be used

to reduce the complexity of the MS/MS spectra, allowing for easier interpretation, either for *de novo* sequencing or through database searching. Many of the above mentioned pre-processing approaches are employed within a wide variety of different tools and methods (Table 2.4).

Table 2.4 MS/MS spectral pre-processing tools and methods

Type of MS/MS pre-processing	MS/MS pre-processing tool / method	Reference(s)
Charge state deconvolution	Zscore	[293]
Isotope deconvolution	LASSO	[294]
	MS-Deconv (exclusive to top-down proteomics)	[295]
	Non-negative least squares/non-negative least absolute deviation regression	[296]
Noise reduction	Wavelet transformation	[297-299]
	PeakSelect	[300]
Quality filtering	Binary classification and statistical regression	[301]
	PepNovo	[302]
	Spectrum quality classifier	[303]
	Fisher linear discriminant analysis (FLDA)	[304]
	EagleEye	[305]
	ScanRanker	[306]
Spectral clustering	MS2Grouper	[307, 308]
	Pep-Miner	[309]
	Metric space embedding	[310, 311]
	MS-Cluster	[312]

2.3.7 MS/MS database searching

Once MS/MS spectra have been generated and optionally pre-processed, such as by clustering and quality filtering, a key step is the ability to interpret the data by searching it against a database of known and/or putative proteins. The ability to interpret MS/MS spectra has improved over the last 20 years, with the development of various different MS/MS search algorithms, employing different searching and PSM scoring strategies to assign a best match between a spectrum and peptide sequence.

One of the first search strategies developed was based on a sequence tag approach, which used *de novo* sequencing of the MS/MS spectra, limited to generating short sequence tags of 3-4 residues, called peptide sequence tags (PSTs), flanked by the masses of the N- and C- terminals. By combining this information with fragment ion

masses, the precursor peptide mass and its enzyme specificity, unambiguous matches could be found [313]. Such an approach has seen great improvements over the years in terms of the *de novo* sequencing of the MS/MS spectra [314-322] and its hybrid approach with peptide fragment matching [323-332].

An advantage of using a sequence tag approach is its ability to filter and reduce the size of the database, and by extension reduce the false-positive rate. This also improves the ability to detect PTMs, and to conduct mutation-tolerant searches, improve search speed, improve the accuracy compared to other popular non-tag based search tools [326, 327, 329, 330], as well as to effectively utilise poor quality MS/MS spectra with usually short peptide sequences [333]. At the same time the disadvantages of this approach include the difficulty of *de novo* interpretation of the MS/MS spectra. While there have been great improvements over the years with the accuracy of PST generation, there is still room for MS/MS spectral interpretation to improve. For example, with the sequence tag based tool InsPecT [326] a set number of PSTs are generated for any given spectrum, even if there are likely many more correct PSTs which could be identified. In addition, the ability to generate PSTs is reduced with increasing peptide length (>15aa) and poor quality MS/MS spectra, the efficiency of filtering large databases with short 3-4 residue PSTs is very limited due to their low specificity, particularly when only relatively few PSTs are generated. A similar method to the sequence tag approach was that employed by Look-up peaks, which uses both *de novo* sequencing and database searching [323].

The concept of the sequence tag approach has since evolved. Where the sequence tag approach was limited to a set number of MS/MS spectral interpretations, a new approach called spectral dictionaries reconstructs all possible full-length peptide sequences from a spectrum to ensure at least one of the peptides is correct. The approach sees improvements with efficiency and specificity over the short sequence tag

based approach for filtering large databases, due to the longer sequence tags and accuracy of the peptide reconstruction.

The concept of the spectral dictionary was first introduced [182] and later implemented in a software tool called Robust Accurate Identification (RAId) [334] that performed slow heuristic searches, which reduced its usefulness for larger datasets. Recent advances in this area have seen many orders of magnitude reduction in search time using the spectral dictionary approach, with the tools MS-Dictionary [324] and its successor MS-GappedDictionary [335] (Table 2.5). MS-GappedDictionary is an extension to the spectral dictionary approach using gapped peptide sequences (substrings substituted with mass values) to resolve poor ambiguous MS/MS spectra, different amino acid combinations resulting in the same mass shift, and adaptively generating peptide reconstructions for long peptides (>15aa), using *forward-backward* dynamic programming. Applying this approach permits many more peptide reconstructions for any given MS/MS spectral dataset and it is amenable to mutation-tolerant searches and searches of large databases, such as the six-frame translations of the human genome in relatively short time frames, compared to other sequence tag based and non-sequence tag based search tools [325, 335].

The gapped peptide approach has since been extended to interpret other MS/MS spectral types (CID, ETD, HCD and MS/MS spectral pairs CID/ETD) in the *de novo* sequencing tool UniNovo [336]. Recent developments have pushed the length of *de novo* peptide reconstructions further to sequence grade levels, using multiple MS/MS spectral types (CID, ETD and HCD) in combination with multiple protease digests to achieve many more covering peptides, achieving much longer consensus lengths between 60 aa and 200 aa of length with 99% sequence accuracy [337, 338]. Another development which interprets top-down and bottom-up MS/MS spectra, reconstructs the sequences, and uses the top-down longer peptide sequences as scaffolds with the short

peptide sequences from bottom-up to improve the coverage across the scaffolds, similar to the concept used in genome assembly [339].

During the same time that *de novo* sequencing, the sequence tag approach and PMF were being developed, a different database search approach termed Peptide Fragment Fingerprinting (PFF) [340] was rapidly evolving in parallel. PFF is an analogous concept to PMF, but instead it looks at matching many fragments from a parent peptide. This approach did not require interpretation of the MS/MS spectra as with *de novo* sequencing, but instead relied solely on matching the mass values from all the MS/MS spectra against all calculated mass values interpreted from the database. The basic principle behind the approach is that it mimics the bottom-up experiment by digesting the protein sequences in the database *in silico* with the same protease used in the experiment. All the theoretical peptide sequences that match the experimental peptide mass within a chosen maximum mass deviation (MMD) are then chosen as candidates. Each of these peptide candidates are then investigated further at the MS/MS level by comparing the experimental and theoretical peptide fragmentation patterns and then ranking/scoring the peptide and peptide fragment matches by the level of correlation between the patterns, which often differ between the MS/MS search algorithms employed. To account for PTMs, the theoretical MS/MS spectra derived from *in silico* digestion of the sequence include an additional mass shift for each peptide fragment corresponding to the mass of all (an exhaustively long 'blind' search) or the masses of a chosen list of modification(s).

Yates and Eng first demonstrated the PFF approach for database searching in 1994, employed in the MS/MS database search tool, Sequest [341, 342]. Since that time many similar approaches have been devised, as well as approaches which have incorporated sequence tags, spectral dictionaries, gapped peptides, and/or spectral

probability (Table 2.5). There have been many other reviews in the literature covering other search tools and strategies [343-346].

Table 2.5 Types of MS/MS database search tools and their employed methods

Method	MS/MS database search tool	Reference(s)
PFF	Sequest	[341, 342]
	Mascot	[165]
	Tandem	[347]
	OMSSA	[348]
PFF with sequence tags	InsPecT	[326]
	GutenTag	[327]
	OpenSea	[328]
	MultiTag	[329]
	TagRecon	[332]
	PPM-Chain	[330]
PFF with spectral dictionary	RAId	[334]
PFF with spectral dictionary and spectral probability	MS-Dictionary	[324]
PFF with gapped peptides, spectral dictionary and spectral probability	MS-GappedDictionary	[335]
PFF with spectral probability and scoring for multiple types of spectra	MS-GF+	[292]

For a long time the proteomics community has been split on how best to interpret MS/MS spectra through database searching and the control of false positives. This eventually culminated in a study in 2004, which found many reported results in the proteomics community had a very high false discovery rate (FDR). As a result, stricter guidelines became required for publication [349], and two years later a study conducted by HUPO across a number of laboratories [350, 351] also found widely inconsistent results [352], with less than 50% of results in agreement. This illustrates just how difficult MS/MS spectral identification is and it highlights a number of challenges which still need to be met, including the protein inference problem [220], and the inherently poor reproducibility of bottom-up, DDA-based proteomics. As a result of this study, HUPO-PSI was formed [353], providing community guidance on the standardization and validation of proteomics results. HUPO-PSI further enforced a requirement to share all the raw data associated with any publications through public repositories such as PeptideAtlas [354], PRoteomics IDEntifications (PRIDE) database

[355] and ProteomeXchange [356], and in addition developed standard proteomics data formats such as mzXML, mzML and mzIdentML. HUPO-PSI also concluded that there was a pressing need for standardization across different studies; concluding that any comprehensive proteomics study should endeavour to improve these standards, and apply strategies which can better discriminate between true and false positives. Such approaches could be a means to combine results from multiple search tools, to provide more robust and well-rounded peptide matching strategies, improve search speed and improve the number and confidence of matches found.

To address the needs outlined by HUPO, a number of advances have occurred in the development of database search algorithms. Approaches which improve the sensitivity, speed and number of peptide identifications, include tools such as MS-GappedDictionary, outlined previously, addressing the issue of search speed and improving the number of identifications compared to other non-gapped peptide approaches such as OMSSA, InsPecT and MS-Dictionary (Table 2.5). MS-GappedDictionary achieved this by generating a spectral dictionary and rigorously determining the spectral probabilities (p-values) of each PSM using the generating function approach [357]. However, this approach, and other similar full length peptide sequence and tag based approaches similar to it, were limited when it came to highly charged MS/MS spectra [325]. Approaches that were not limited by MS/MS spectral charge included non-tag-based methods such as MS-GF+ (Table 2.5). The MS-GF+ search tool uses the MS-GF scoring model first implemented in MS-Dictionary, and is able to automatically determine the scoring parameters for a variety of different MS/MS spectral types (CID, ETD, and CID/ETD pairs) and proteases (trypsin, LysN etc.).

Following peptide identification, it is necessary to identify the proteins from which the peptides were derived from. However, confident protein identification can be hampered by the protein inference problem first mentioned in Section 2.3.1. Although

top-down or middle-down proteomics directly addresses the protein inference problem, it is not as widely used and does not provide good overall coverage of the entire proteome compared to bottom-up mass spectrometry. With the protein inference problem still an issue for bottom-up mass spectrometry, algorithmic and database construction approaches have arisen [220], such as resorting to infer only a group of potential proteins, or infer likely candidates using approaches such as those outlined in Table 2.6.

Table 2.6 Algorithmic and database approaches for the protein inference problem

Method	Protein inference tool/approach	Reference(s)
Expectation-maximization	ProteinProphet	[219]
Bayesian	Empirical Bayes Protein (EBP) identifier	[358]
Parsimony	IDPicker 2.0	[359, 360]
Deterministic	PeptideClassifier	[361, 362]
Graph theory	Clique-enrichment approach (CEA)	[363]
Heuristic	Prediction of proteotypic peptides for protein identification	[364]
	Minimum acceptable detectability for identified peptides (MDIP)	[365]
	Minimum protein set with incorporated peptide detectability	[366]
Peptide-centric	IsoformResolver	[367]
	PEPtidomics Protein Isoform (PEPPI) database	[368]
	Mass spectrometry-centric sequence database (MScDB)	[369]

The presence of contaminant MS/MS spectra, derived from non-target proteins is a problem in proteomics, and can often lead to misidentifications. To account for potential contaminants entering a sample it is standard practice to append contaminant sequences to the database before searching. Often this consists of human keratin and trypsin and other enzymes used in the analysis, as well as any contaminants deemed likely to appear due to sampling and handling. MS/MS spectra exclusively matching the contaminants can then be removed during the analysis.

2.3.8 Next generation MS/MS database search technology

As outlined previously, new approaches for MS/MS database searching for bottom-up proteomics are breaking new ground, such as the methods employed by MS-GF+ and MS-GappedDictionary. However, other MS/MS database search technologies are

beginning to look beyond traditional bottom-up proteomics, to how best to manage and interpret the MS/MS spectra from top-down and DIA mass spectrometry. Both methods, which have been around for over a decade, are now becoming viable for more complex samples and throughputs, opening them up to truly mine the proteome. Unfortunately, there has been a lag in the development of MS/MS database search tools tailored towards such datasets.

Top-down MS/MS spectra are highly complex, and to be used in an MS/MS database search tool first requires pre-processing to reduce complexity. The MS/MS spectra are usually first deconvoluted to their monoisotopic masses (determining the mass and charge of the fragment ions from a group of isotope peaks called a isotopomer envelope). This can be done via a variety of tools such as Thrash [370], Xtract [371] and the more recently MS-Deconv [295]. The spectrum now containing only monoisotopic masses is then scored against proteins in a database using a database search tool to generate a Protein-Spectrum-Match (PrSM). A number of top-down search tools have been developed over the last decade, which include ProSightPC [372, 373], PIITA [243] and USTag [241], MS-TopDown [374], top-down versions of Mascot [375], Sequest [240], and OMSSA [348] search tools and more recently the search tools MS-Align+ [376] and its improved version MS-Align-E [377] to identify multiple proteoforms from highly modified proteins. Further developments are needed in this area to more confidently identify all the various proteoforms in complex samples, particularly for large highly modified proteins. In addition there is currently a high-throughput top-down search strategy in development at Pacific Norwest National Laboratories (PNNL) called IQ Top-down, using information from MS1 spectra to validate proteoforms [378].

The range of MS/MS database search tools to directly interpret DIA MS/MS spectra is limited in comparison to those available for DDA. This is because DIA

MS/MS spectra requires significant pre-processing (charge/isotopic deconvolution), and require conversion into DDA-like MS/MS spectra before interpretation [268].

A number of DIA-compatible tools have been developed to meet the need for DIA MS/MS spectral interpretation. For example, the MS^E approach uses the $IDENTITY^E$ database search tool [379] but lacks any estimation on FDR, requiring validation from conventional DDA approaches [380]. Alternatively, a modified approach to MS^E, called Ion mobility spectrometry (IMS) assisted MS^E ($HDMS^E$) [381], applies an additional separation in the gas phase improving proteome coverage by up to 60% and then uses a database search tool called Synapter which, unlike $IDENTITY^E$, allows control of the FDR [382]. For AIF MS/MS spectra there is an MS/MS data processing package called MaxQuant, which in combination with the DDA based Andromeda MS/MS database search tool is able to process and interpret DIA MS/MS spectra from AIF, and produces pseudo MS/MS spectra before interpretation by Andromeda against a database with FDR control [259, 383]. For FT-ARM MS/MS spectra, there is the FT-ARM analysis tool, which creates hypothetical MS/MS spectra from *in silico* enzymatic digests from a database and scores each hypothetical MS/MS spectra against all the acquired multiplexed fragmentation MS/MS spectra, and also applies a control on FDR [260]. There are also a number of tools still under development, for example, SWATH spectral analysis, which does not currently have any established database search tools as it is incompatible with database search approaches [268]. However, recently an open-source tool called SWATH-Umpire, claimed to extract SWATH signal features from MS1 and MS2 and assemble them into pseudo MS/MS spectra compatible with DDA based search tools [384]. Lastly, there are other tools under development from Pacific Northwest National Laboratories (PNNL), such as IC bottom-up, which is a universal tool for both DIA and DDA, and demonstrates improvements over advanced tools such as MS-GF+ [378]. Additionally,

approaches such as PAcIFIC, generate MS/MS spectra compatible with conventional DDA-based search tools, and recent developments have demonstrated how the approach can be applied to top-down mass spectrometry by combining the PAcIFIC approach with the top-down search tool, PIITA [385].

Instead of identifying MS/MS spectra against a database of putative and/or known protein sequences resulting in a PSM, an alternative form of database searching is to match the MS/MS spectra against a spectral library (database) of other curated spectra to identify a spectrum-spectrum match (SSM). The technique is called spectral library searching and was first pioneered by Domokos [386], and later adapted for peptide mass spectrometry [387]. The concept relies on the previous identification of MS/MS spectra to populate the spectral library with identified MS/MS spectra. The searching of spectral libraries is fast, precise, has an improved PTM identification rate and has fewer false positives, rivalling the majority of conventional database search tools and strategies. A number of tools that have been developed to search and identify MS/MS spectra in this way are outlined in Table 2.7. Of particular note is the Tremolo spectral library search tool, which is a recent advancement, utilizing a spectral library generating function approach to identify SSMs. The concept was first demonstrated in MS-GF+, a conventional database search tool, which when compared to spectral library tools such as SpectraST was comparable in sensitivity indicating a likely route for improvement of spectral library searching [292]. A number of spectral libraries have been established over the years, which include PeptideAtlas [354], Cardiac Organellar Protein Atlas Knowledgebase (COPaKB) [388], NIST Libraries of Peptide Tandem Mass Spectra [389] and BiblioSpec [390].

Table 2.7 Spectral library search tools

Search tool	Reference(s)
SpectraST	[391]
NIST Peptide Library Search Engine	A modified tool from [392]
X!Hunter	[393]
BlibSearch	[390]
Pepitome	[394]
Bonanza	[395]
MSplit (multiplexed spectral library search)	[271]
MSplit-DIA (multiplexed spectral library search)	Currently under review [396]
Tremolo	[397]

Recently, an approach called Spectral archives [398, 399] was developed, expanding on MS-Cluster to generate an archive of many large spectral datasets. With this approach the identification of MS/MS spectra becomes rapid, achieves high specificity, and is amenable to the identification of novel MS/MS spectra such as biomarkers and unknown proteins/genes, which would then contribute to a proteomics community consensus of other unidentified MS/MS spectra in the cluster, improving its overall representation and confidence of a real identification. Previously, such an approach only searched MS/MS spectra against a locally known set of curated MS/MS spectra or in the case of a conventional database approach, a protein sequence database. The concept of spectral archives was later extended to molecular networks, where any class of molecule can be rapidly and confidently identified within both the proteomics and metabolomics domains [400].

The identification of MS/MS spectra by searching a conventional sequence database has a number of caveats when compared to spectral library/archive searching: 1) the database is often only limited to 'known' protein sequences, while adding more putative proteins will improve identifications, it will also hamper identifications as the database size increases, reducing sensitivity; 2) the database contains many stretches of homologous sequences (e.g. larger numbers of isoforms) making identifications, particularly of short sequences, difficult; 3) incomplete proteolytic cleavage or fragmentation of MS/MS spectra can lead to misidentifications; 4) the presence of

unaccounted for PTMs and sequence polymorphisms can lead to misidentifications; 5) the potential for erroneous sequences in the database is high, often derived from genomic sequences with potential sequencing errors; and 6) usually the search is limited to peptides which have been proteolytically cleaved (e.g. tryptic peptides) to reduce the occurrence of spurious PSMs and limit the search to only tryptic sequences, improving the search speed and sensitivity.

The sensitivity of a MS/MS spectral identification is inversely proportional to the size of the search space. The search space of *de novo* sequencing is much larger than conventional databases, and more so when compared to spectral libraries. This is because the number of MS/MS spectral interpretations is far greater when interpreting MS/MS spectra into a large number of possible peptides, without being limited to matching to a chosen number of sequences (PSMs) or a select number of curated MS/MS spectra (SSMs). This indicates that of all the methods of MS/MS spectral identification, the spectral library approach is more sensitive, but until the method matures and spectral archives are more widely supported by the proteomics community for the identification of known and unknown MS/MS spectra, the conventional database approach will remain more widely used.

2.3.9 Statistical approaches for peptide and protein identification

The search algorithms that MS/MS database search tools employ, use PSM scores to determine the confidence of a match between the experimental MS/MS spectra and a number of matches to other theoretical MS/MS spectra derived from the protein database. Many different scoring systems have been implemented in search tools outlined in Table 2.8.

Table 2.8 PSM scoring methods in some common MS/MS database search tools

Method	MS/MS database search tool	Reference(s)
Cross correlation between theoretical and observed MS/MS spectra	Sequest	[341, 342]
Bayesian probability based on the number of ions matching a peptide sequence in the protein database	Mascot	[165]
Dynamic programming using k-similarity statistics	MS-Alignment	[401, 402]
Calculated probability factors	PeptideSearch	[313]
Spectral energy (Delta energy between the best *de novo* and database spectral interpretation)	MS-GF+	[292]

Further scoring of PSMs during post-processing is possible, using a probabilistic score (e.g. peptide probability), determined through expectation-maximisation in PeptideProphet [403] and iProphet [404], a support vector machine in Percolator [405], probabilistic network in PepNovo [302] and generating function in MS-GF [357]. The post-processing tool, iProphet, was also able to combine multiple search results, applying a protein or peptide probability as a score derived from across all results [406]. Other re-scoring tools such as Percolator and MS-GF are also able to achieve this, to normalise and combine results.

Of particular note from Table 2.8, is the MS/MS database search tool MS-GF+ (an extension of MS-GF), which when also applying the spectral probability to all identified PSMs as a score for FDR filtering, was shown to identify more peptides than other well established search tools [292, 407], such as Mascot [165], Sequest [341, 342], OMSSA [348] or a combination of these, rescored with PeptideProphet [403], iProphet [404] or Percolator [405].

The scores obtained from various MS/MS database search tools are ranked and in most cases the top matching MS/MS spectra are chosen, which are then further analysed at the MS/MS level by comparing the theoretical and experimental peptide fragmentation patterns. Each match can then also be further ranked by significance estimates via the false positive rate (FPR) or p-value and E-value. The p-value is

essentially the chance that any individual PSM is incorrect, determined from the fraction of all incorrect PSMs above a certain score threshold over all the incorrect PSMs, and can be extended to the E-value in multiple hypothesis testing which is the product of all p-values and the number of tests, or more simply put, the expected number of times that a PSM is observed with a particular score, by chance alone. The p-value threshold, usually either 5% or 1%, can then be applied, delegating matches as significant and rejecting the null hypothesis (all matches known to be incorrect), or in agreement with it, delegating the match as incorrect. A number of methods have been devised to estimate the p-value over the years, mostly through empirical methods, such as those outlined in Table 2.9.

Table 2.9 Examples of methods employed to estimate the FPR

Method	Description	Reference(s)
Score distribution	Models the distribution of all scores to determine the significance at the tale of the distribution.	[408, 409]
Poisson distribution	Models the distribution of false-positive matches given some prior criteria (e.g. peptide length, protein database size etc).	[351]
Bayesian probability model	Probability of the match being correct given certain criteria such as fragmentation ion types, prior MS/MS spectra and peptide knowledge across an experiment.	[219, 403, 410]
Decoy database	A null hypothesis is generated. This is done by reversing, shuffling or randomizing the target sequences.	[411-415]

The decoy database mentioned in Table 2.9, is a more common approach due to its simplicity, and empirically represents the null hypothesis. Using the decoy database approach as an example, the estimated FPR can be calculated by determining the number of matches to the decoy database over the number of total decoys. However, since the number of matches to the decoy is usually very small or zero, to calculate accurate FPRs the decoy database would need to be extremely large, which is impractical (and is usually the same size as the target database), so matches need to be grouped by the same score to estimate the FPR which tends to be inaccurate [357]. In addition, PSM scores from different MS/MS search tools use heuristic methods, which do not correlate well with their FPRs, since many search algorithms usually assign similar scores to the top-most ranking PSMs. As a consequence, the estimation of their

FPRs are more often inaccurate, resulting in an overlap between all correct and incorrect identifications above a certain score threshold. To help reduce the number of false positives, a variety of different search algorithms employ heuristic measures to their scoring, such as taking into consideration the difference between the score of the best match and the second best match, the peptide charge state, peak intensity, the fraction of b ions, and many other properties.

As datasets became larger and multiple hypothesis testing increased it soon became apparent that control of the number of false positives across all tests was needed, which resulted in the now popular approach, called the false discovery rate (FDR), first pioneered by Benjamini and Hochberg [416]. A decoy database is a common method to determine the FDR, which can be either concatenated with the target database and searched or treated as separate databases. Depending on the chosen method, the FDR can be calculated, either by counting the number of decoy matches above a particular score over all matches to the target above the score or by multiplying 2 by the number of decoy matches above a particular score over the combined number of decoy and target matches above the score [411, 412, 417]. The FDR can then be used to remove groups of matches delegated as false, which is usually applied at the PSM-level, or to be more conservative and reduce false positives the FDR is applied at the peptide-level, a common and recommended strategy in proteomics studies. This is due to single peptides often being derived from multiple MS/MS spectra, with some potentially being spurious, resulting in very different numbers of identified PSMs between PSM-level and peptide-level FDRs.

Strictly applying filters using the p-value, E-value or FDR, runs the risk of either being too strict or too lenient, filtering out many true positives or including too many false positives, respectively. An approach found to alleviate this problem was adopted from genome-wide linkage analysis studies. The use of smaller datasets during the early

genomics era, required only p-value cut-offs and they were often made strict to avoid any false positives, a valid approach when there are relatively few hypotheses being tested in a single small study. In recent times, with larger datasets being generated through high-throughput methods, there are now often thousands to millions of multiple hypothesis tests being conducted in a study, with many more genomic features now being considered significant. As pointed out earlier, a high level of statistical significance does not equate to a true positive, and so caution is needed when considering every significant feature. It was found that simply applying strict p-value cut-offs across all hypothesis tests runs the risk of removing true positives, and therefore further approaches need to be applied to retain as many true positives as possible, while minimizing the false positives [418].

Although the FDR measures the significance of a group of PSMs, it does not provide a level of significance for each PSM, and hence a different approach is necessary to improve sensitivity. This can be achieved using the Posterior Error Probabilities (PEP) and q-value. The PEP can be considered as a local FDR (lFDR), and was coined as such by Efron et al. [419]. When using very large datasets the p-value can be very small by chance alone, limiting its use, and hence a different approach is required in the form of q-values. The q-value is a p-value analog based on the FDR, while the p-value is based on the FPR. The q-value, unlike the p-value, includes multiple testing corrections, by determining the minimum FDR of a significant PSM score. In other words, the q-value provides the proportion of incorrect identifications amongst all those considered significant [418]. The PEP or lFDR is the probability that an individual observation (e.g. a single PSM) with a particular score, is found from within the null distribution (e.g. a false peptide within a decoy database) [420, 421], and by summing all the PEPs/lFDRs of significance and dividing by the total number of PEPs/lFDRs, the FDR for that group of PSMs can be determined [422]. The q-value and

PEP/lFDR are usually used in a two-tier approach towards PSM validation. The q-value estimates the rate of incorrect PSMs from a group, while the PEP/lFDR applies significance to whether a particular peptide or protein is present or not. Hence, the first pass of assessing PSM significance should be with a q-value threshold to filter out likely incorrect PSMs, after which a second pass with the PEP/lFDR is used to determine the likelihood of the remaining PSMs being truly present [421].

The FPR is a measure of the quality of a single PSM and is the best approach towards discriminating true and false positives. Recent approaches to derive the FPR or p-value using theoretical means have been developed [357, 423], but they rely on the assumption that each peptide derived from the spectrum is equally likely. One of these approaches uses the generating function, a commonly used combinatorics approach implemented in MS-GF+ [292]. The method determines the spectral probability (p-value) determined from all theoretically possible peptide reconstructions, and spectral energy (score) for each PSM determined from the difference in score between the best peptide reconstruction and the best database peptide. The approach theoretically determines the FPR or p-value, as opposed to empirically using Bayesian algorithms or the use of decoy databases, mentioned previously. The E-value and FDR and consequently the q-value and PEP/lFDR can then be determined from the theoretically derived FPR, thus avoiding the need for a target-decoy approach (TDA) [357] due to its many short comings [424]. The ability to calculate theoretical FPRs, and by association theoretical FDRs, is crucial for studies requiring a reliable distinction between true and false PSMs, such as in metaproteomics, proteogenomics or with the identification of rare PTMs [424]. However, the empirical approach using TDA still remains the standard approach undertaken by many studies, with no study conducted thus far using the FPR alone. This is a direct result of the limitation of the current available MS/MS

database search tools, which still have to rely on empirical approaches to determine an approximate FPR for proteomics and proteogenomics studies.

Recent developments to alleviate this problem eventuated in the study which developed MS-GF+ [292], which found that the spectral probability of a spectrum, independent of the database, was able to closely predict the FDR with low-low precision MS/MS data obtained from LTQ experiments, but produced inaccurate predictions with the more commonly used high-low and high-high MS/MS data obtained from machines such as LTQ Orbitrap and QTOF respectively. Such discrepancies are seen when using high precision MS data with tighter precursor mass tolerances (e.g. parts per million (ppm)) compared to wider fixed value precursor mass tolerances (e.g. Daltons (Da)) for low-low precision MS data. The tighter tolerances reduce the search space and inflates the expected FDR and E-values determined from the spectral probability, while loose tolerances with low precision MS/MS data increase the search space, reducing the factual FDR and E-values determined from the TDA [292].

High precision MS/MS data is important to resolve and identify many multiply charged peptide ions as mentioned previously, and so even though more peptides can be identified with higher precision, at the same time their expected FDRs cannot be determined with the same level of precision using the spectral probability or FPR alone. Consequently, the TDA is still an essential tool in proteomics, until a more robust strategy can be developed to determine the FDR accurately, independent of the database.

Continuing on from the theme of resolving the protein inference problem, highlighted in Section 2.3.7. Apart from the algorithmic methods listed in Table 2.6, there is a general rule in proteomics, to identify proteins considered statistically valid,

by arbitrarily filtering all proteins, which contain at least two identified peptides. This is called the “two-peptide rule”, which essentially filters out all single-hit protein matches, often referred to as “one-hit-wonders”, and which are considered as potential false positives. Another approach that can be combined with the two-peptide rule to further improve confidence of protein identification is percent protein coverage. The larger the proportion of protein sequence which is covered by identified peptides, the more confident that the protein has been identified, and that the peptides have not simply been derived from other similar proteins, protein isoforms or paralogs. In cases where proteins are small, the percent protein coverage may likely be higher, and the level of protein coverage obtainable would also be dependent on the type and number of proteases used.

Other more complex approaches include ProteinProphet [219] mentioned in Table 2.6, which uses an expectation-maximization algorithm to derive a minimum protein list and assigns a protein probability to each identified protein.

Although the two-peptide rule and percent protein coverage are capable of confidently identifying proteins, these approaches also result in losses due to many filtered out one-hit-wonders being real matches. In addition, relying simply on maximizing peptide matches does not equate to protein identifications, as many peptides can be shared across different protein isoforms, which would result in biases skewing the calculated protein FDR. One approach, superior to the two-peptide rule and the ProteinProphet approach is the single-peptide rule in combination with a rigorous FPR calculation, using the generating function approach, and extending it to protein identification to determine the protein FPR. Using this approach Gupta et al. [425] found significant improvements compared to ProteinProphet and the two-peptide rule, confirming that single-hit peptides in orthologs of other species accounted for an additional ~25% of proteins missed by the two-peptide rule. However, to account for

protein isoforms, applying a unique peptide rule on the final aggregate of proteins identified would remove any potentially incorrectly included protein isoforms. A comprehensive review published in 2010, outlines many of the above mentioned statistical methods [426].

As mentioned earlier in Section 2.3.8, the precursor mass tolerance and the size of the database have an effect on the sensitivity of a database search. Adding further PTMs to a search also has the same effect [357, 402, 427-429]. The most widely applied method to improve on the search space is to reduce the database size where possible and to only search for commonly used PTMs such as carbamidomethylation of cysteine, oxidation of methionine and protein N-terminal acetylation, and to limit the number of modifications per peptide to at least two. Further reduction of the database size is possible by applying a two-pass search approach, which was first pioneered by Craig and Beavis [430], and can be applied for particular cases where the database size is overly large. This can be done by first performing a search with no FDR filtering and no decoy database, thus identifying matching protein sequences, and then performing the search a second time using the identified sequences as the target database and their decoys, followed by FDR filtering.

All these approaches would maximize the number of matches obtainable, compared to the inclusion of further PTMs or additional sequences. The study from [427] assessed many of these search strategies and others, and found that for MS-based proteomics there were twelve optimal methods to use (Table 2.10).

Table 2.10 Optimal methods for proteomics analysis

Method number	Method
1	Use a reversed decoy database.
2	Concatenate target and decoy databases.
3	Calculate FDR by counting the number of decoy matches over all matches to the target, above a score.
4	Use the smallest possible database.
5	Filter out of all unidentifiable MS/MS spectra.
6	Apply tighter parent mass tolerances where possible during the database search.
7	Normalise PSM scores when possible. For example use q-values, p-values, local FDRs, peptide probability/posterior probability and spectral probability.
8	Apply peptide-level FDR filtering for protein identification or accurate PSM-level FDR filtering, when warranted, such as with metaproteomics or proteogenomics.
9	Apply a two-pass search approach, although the gains could possibly be reduced for higher complexity samples.
10	Carefully select appropriate PTMs.
11	Choose an appropriate MS/MS acquisition mode, such as MS/MS and MS/MS/MS for higher mass accuracy.
12	Use spectral library searching when possible.

All twelve of the above mentioned methods could have a significant impact on sensitivity of PSM identification during a MS/MS database search. Other approaches to improve upon the sensitivity of a database search include limiting the peptide fractions to a particular range of isoelectric points (pIs), predicting the pI of sequences within the target database and limiting the search space to the same pI range when performing the MS/MS database search [429]. More complicated and extensive proteomics procedures can be used to improve the sensitivity and gains from mining the proteome, which include multi-stage peptide identification. This involves an exhaustive iterative process of performing a general search, further processing of unassigned high quality MS/MS spectra, “blind” PTM searches [402, 431], followed by a more focused search for highly frequent PTMs, spectral library searching, and then any remaining unassigned MS/MS spectra searched against large translated genomic databases [426].

It is the searching of large translated genomic databases, and how to manage such a search, which is the primary focus of this thesis. The following section outlines the history of the approach and the methods employed.

2.4 PROTEOGENOMICS

This section provides an in-depth background on proteogenomics, the focal point of this thesis. It provides an overview of the challenges, important dataset considerations, methods of statistical analysis, how best to interpret the results, and provides a review of a number of proteogenomics tools.

Proteogenomics is a new genome annotation approach that has emerged over the last decade, which merges peptide mass spectrometry with genome and transcriptome data. The traditional approach towards genome annotation particularly for the protein-coding portion of a genome has often been limited to direct protein-level evidence or putative translations from gene predictions. A number of annotation strategies, such as the automated Ensembl Analysis Pipeline [432], HAVANA analysis pipeline [143] and MAKER2 [137], previously mentioned in Section 2.2.2 and Table 2.3, all rely on UniProtKB/SwissProt [433, 434] protein sequences, however only 5% of the sequences are derived directly from proteins (http://www.uniprot.org/faq/37). These strategies are determined from sequencing efforts using Edman degradation or *de novo* sequencing through mass spectrometry, while the majority of remaining protein sequences are putative, derived from translations of cDNA and gene predictions which may be erroneous. This traditional approach to proteomics follows the assumption that all the proteins which make up the protein-coding space are known, and that each protein is accurately defined, and are all mirrored in protein sequence databases, such as UniProtKB, Ensembl and RefSeq [435]. Any protein identifications and quantifications carried out are based on these assumptions. As such, the past few decades of proteomics have been skewed to this view and so they ignore the possibility of a hidden proteome. In reality, many peptides derived from proteomics experiments are not present in the reference database or possibly no known reference database. This is in part due to the fact that many MS-derived peptides contain mutations or are derived from novel protein

isoforms or proteoforms, as well as many translated gene predictions or cDNA sequences. Additionally, many protein sequences populating the databases may be incomplete due to gene prediction biases or missing information at the transcriptional level.

Historically, genomics and proteomics have been conducted separately, with the proteomics being carried out after genome sequencing, assembly and annotation. In recent years a new field has emerged which uses mass spectrometry-based proteomic data to annotate a genome with direct peptide evidence, thus complementing current annotations by unambiguously determining the reading frame, translation start and stop sites, splice boundaries, validation of short ORFs, and the identification of novel genes [436-439]. This new genome annotation strategy was first investigated by Jaffe et al. [440], who coined it proteogenomics, and it has since been applied to many other organisms, such as *Drosophila melanogaster* [441, 442], *Arabidopsis thaliana* [81], *Yersinia pestis* [443], *Pristionchus pacificus* [84], *Mus musculus* [444], *Ruegeria pomeroyi* [445, 446], *Shewanella oneidensis* [447, 448], *Vitis vinifera* [8], *Zea mays* [82], and *Homo sapiens* [11, 449-451].

Proteogenomics is an important new field, because it assists in the improvement of gene predictions as well as the direct validation of known genes as protein-coding. By approaching genome annotation top-down (genomics) and bottom-up (proteomics) a more complete and thorough analysis of gene models can be achieved. Besides the key benefits to refining gene models, other benefits include: 1) spectral counting to infer expression levels; 2) the discovery of alternative translation initiation start (TIS) sites through identification of N-terminal peptides, either through standard identification of N-terminal acetylated peptides by MS/MS database search or with a more rigorous approach employing the enrichment of N-terminal peptides through N-terminomics [452]; and 3) the determination of post-translational/processing modification sites, e.g.

to identify signal peptide cleavage sites. Figure 2.5 (modified from [148]), outlines the key differences between genomics and proteogenomics approaches.

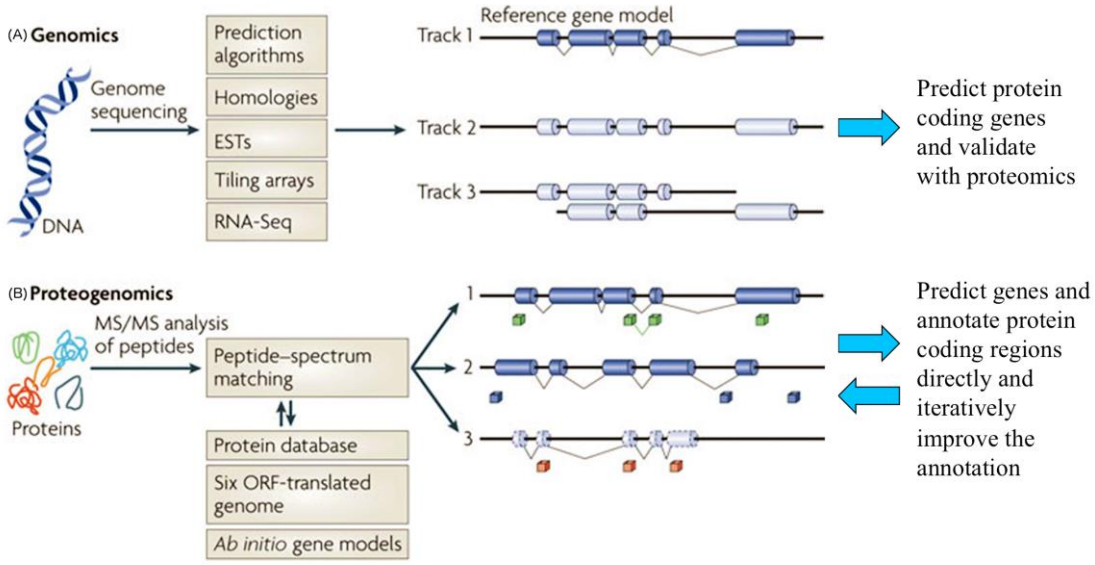


Figure 2.5 Genomics versus Proteogenomics

The (A) genomics approach is linear from gene prediction at the genomics level, with protein coding validation of the predictions at the proteomics level. However, the (B) proteogenomics approach utilises MS-based proteomics to infer the gene predictions and annotations directly in the genome in a complementary way with the genomics approach, iteratively improving the annotations with further peptide mass spectrometry evidence (image modified from [148]).

2.4.1 Defining a proteogenomics search

The aim of proteogenomics is to identify unknown protein-coding regions, and to correct and validate the known gene models. This entails performing database searches of MS/MS spectra against the known proteome and a putative search space to identify peptides and map them to these regions. These unidentified protein-coding regions could identify potential novel protein-coding loci (novel genes), expressed pseudogenes/lncRNA, and intergenic regions, such as between genes. Or by extending the known gene boundaries within a predefined linkage distance [82], within untranslated regions (UTRs), between exons of a gene (i.e. within introns), on different frames as those suggested by the known gene, and across splice junctions (Figure 2.6). In addition known genes can be further refined through the identification of alternative TIS sites inferred from N-terminal peptides to infer over-predicted genes.

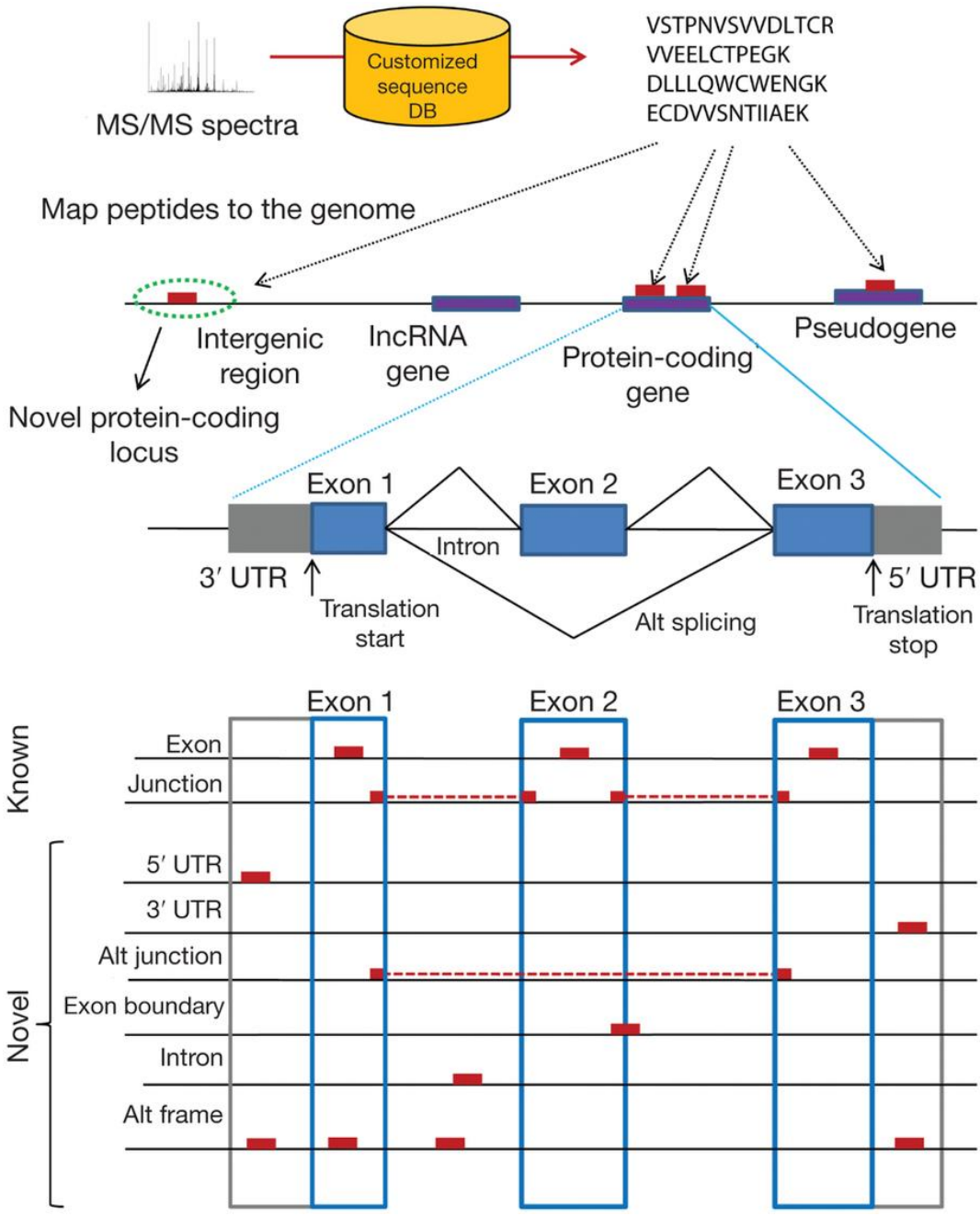


Figure 2.6 Proteogenomics annotation events

The range of proteogenomics annotations, which fall outside of the known annotations, can be classified as a novel protein-coding locus (novel gene), expressed pseudogene/long non-coding RNA (lncRNA), untranslated UTR regions, alternative splice junctions, exon boundaries, novel exons (novel peptides mapped within an intron) and alternative frames. In addition, not illustrated in this figure are gene boundaries (within predefined boundaries closely flanking the gene region) (image courtesy of [439]).

Generating databases to reflect all these possible protein-coding regions is not trivial. They can be defined by 1): a six-frame translation of the genome, which was first pioneered by Yates et al [453] and is usually limited in size to open reading frames (ORFs) of around 30-40 amino acids, to reflect the most likely minimum peptide length and protein coding portion of the genome and to restrict the extent of the inflated

database size. This type of approach is commonly applied in various studies across a range of organisms [81, 82, 443, 453-456]. But this does not cater for peptides spanning exon-exon junctions. In such situations *de novo* sequencing of the un-assigned MS/MS spectra could account for peptides spanning these exon-exon junctions. However, as previously mentioned, *de novo* sequencing has a larger search space and is highly error-prone. Other methods are therefore needed to define the exon-exon junction search space, such as with 2): *ab initio* gene predictions, which are able to predict the gene and the internal exon-exon junctions using numerous tools (as discussed in Section 2.2.2). This limits the size to the most likely coding regions, ORFs and exon-exon junctions given some prior evidence, as demonstrated in a previous study [444]. Although this limits the search space and includes exon-exon junctions, it biases the search to whatever the *ab initio* method considers being a real gene feature and limits the scope for the discovery of novel annotations. Additionally, many incorrect and missing gene annotations are a direct result of the limitations of *ab initio* gene prediction.

Previous studies have applied this approach in conjunction with a six-frame translation to define a separate exon-exon junction search space, derived from cDNA evidence, with each exon represented only once in a compact exon splice graph [81, 82, 457]. An alternative method is to use 3): a six-frame translation of ESTs, which would provide experimental evidence of expressed sequences and limited splicing information (due to sequencing bias at the 3' end), which would include novel splice sites and single amino acid polymorphisms (SAPs). Although this has the same limitations as a pure six-frame translation of ORFs with its large size, EST datasets can be reduced in size by applying a series of stringent criteria to best represent the most likely coding regions and representing the sequences in a de Bruijn graph [458], as well as the option to align multiple ESTs to genomic regions, and determine exon-intron boundaries which can then be represented in a compact exon splice graph [457].

In another alternative method 4): a three-frame translation of curated RNA transcripts derived from databases such as Ensembl or RefSeq would allow the identification of alternative TIS sites and frame-shifts and since the strand orientation is known, it avoids further inflating the search space by not including frames on the opposite strand. Sequences of pseudogenes and lncRNAs could also be included to re-classify such transcripts as coding [459]. A different method is to 5): generate a database of splice junctions, derived from RNA-seq reads aligned to a genome [460, 461] using splice alignment tools such as Bowtie2-TopHat2 [126] and STAR [127], as well as a full transcript assembly through *de novo* methods, or preferably genome-guided methods using tools such as CuffLinks [462], where high abundant transcripts are kept (based on read counts), and a three-frame translation is performed [463]. Additionally, a sequencing strategy, called Ribo-seq or Ribosome profiling [464, 465], can be used with this method, which is the sequencing of mRNA bound to ribosomes, enriched for transcripts destined for translation. This would limit the search space to only transcripts coding for protein, allow for easier identification of reading frame, frame-shift changes, identification of which transcript isoforms are being expressed, and in combination with N-terminomics [452], allow for the identification of alternative initiation codons.

Various other spliced peptide sequence databases can be generated, using RNA-seq data with automated approaches from systems such as Galaxy-P, generating novel SAPs, splice junctions and sequences from highly expressed transcripts [466]. Other tools capable of generating databases from RNA-seq data are customProDB [467], which can generate splice variant sequences from public repositories such NCBI dbSNP [468], the Online Mendelian Inheritance in Man (OMIM) database [469] and the Protein Mutant Database (PMD) [470]. Databases can be generated containing sequences from sites of RNA-editing, identifiable from RNA/DNA comparisons with tools such as

REDItools [471]. Splice sequences can also be extracted from large RNA-seq datasets aligned to the genome and incorporated into a splice graph consisting only of splice regions distinct from sequences found in the six-frame translated genome [472]. This has been demonstrated to incorporate peptide variants, derived from variant calling tools such as the Genome Analysis Toolkit (GATK) [473], in the detection of mutant peptides in cancer [474]. This could also be used to generate sequences containing sites of RNA-editing.

Depending on the study or target organism, other approaches could include the use of 6): curated databases such as the ECgene database with its large numbers of AS sequences [475], the Pseudogene.org database [476], the non-coding RNA sequence database NONCODE [477], a database of large intergenic non-coding RNAs (lincRNAs) located at Harvard [478] and the ChiTaRS database consisting of chimeric RNA transcripts [479].

Many of the above approaches assume a completed genome to work from when defining the search space, either via six-frame translation, mapped ESTs or RNA-seq. However, a complete genome is often not available in the majority of cases, and is only available as draft versions, which can be highly fragmented in early draft versions. This becomes a problem when assigning MS/MS spectra to a genomic sequence, as the MS/MS spectra can potentially be misidentified with a PTM in a different genomic location to that originally derived. A potential way to address this problem, would be to interpret any unassigned MS/MS spectra of sufficient quality by *de novo* sequencing using tools such as PepNovo [314] and UniNovo [336]. The sequences could then be used to search a closely homologous genome sequence or other available sequences, using tools such as MS-BLAST [480], which can also interpret raw MS/MS spectra for searching. In addition, tools such as InsPecT [326] allow for mutation-tolerant searches to find matches within homologous genomes, or to use approaches such as template

proteogenomics, with tools such as GenoMS [481] modified for large scale genome analysis, to find matches to homologous sequences. Strategies of this type allow the assignment of unassigned MS/MS spectra and could also provide a means to assist genome assembly by constructing complete proteins independent of the complete genome, which could then possibly assist with scaffolding where there is minimal read depth or coverage between scaffolds.

2.4.2 Statistical analysis in proteogenomics

Despite the differences between proteomics and proteogenomics, there is often common ground with the application of statistics, taking what is learnt from proteomics, such as FDRs, assignment of FPRs or p-values, and the definition of a null hypothesis through decoy databases [411]. In recent years, the repertoire of tools has increased to deal with the larger search spaces presented by searching the six-frame translation of a genome and addressing the inaccuracies with assigning empirically derived scores and FPRs to PSMs.

In proteogenomics, a PSM level FDR is often chosen as the first step to remove spurious identifications. As pointed out in Section 2.3.9, in proteomics a peptide-level FDR is applied, as the aim is to identify peptides assigned to proteins. By comparison, in proteogenomics, the aim is to identify numerous PSMs across the genome to identify new coding regions, which is likely to contain much more spurious than real identifications. Applying a peptide-level FDR in a proteogenomics study would be overly conservative, and therefore a PSM-level FDR followed with further filtering using probabilistic scores, such as the peptide probability or posterior probability [404] and spectral probability [357], would be a more appropriate strategy.

Besides the previously mentioned FPR (p-value) and FDR, derived from the empirical approach using the TDA, other statistical methods from proteomics can also

be implemented in proteogenomics. Examples of approaches include the q-value [418], PEP/local FDR [421], and theoretical as opposed to empirical approaches to calculate the score and FPR of each PSM using combinatorics [292], which allows for greater discrimination between true and false positives.

A new statistical technique to contribute to this group of tools, specifically developed for proteogenomics, has emerged from concepts in genome linkage analysis studies [418]. Unlike other approaches, which simply apply stringent p-value or score thresholds on peptide identifications and may greatly improve specificity but with a trade off on sensitivity, this approach uses a more holistic approach. Since the aim of proteogenomics is genome annotation by identify protein-coding regions, and not specifically to identify individual peptides or proteins, this new approach uses the annotation *event probability* (i.e. the probability of being correct), which is based on the product of all posterior probabilities or local FDRs of PSMs (1 – local FDR) divided by the number of co-locations across the genome. The *event probability* is assigned to a cluster of peptides identified as an annotation event in the context of the known annotation. The *event probability* is employed in the proteogenomics tool, Enosi [81, 82, 482], which classifies annotation events, such as exon boundaries, novel genes, gene boundaries, frame-shifts, reverse strands, translated UTRs, and novel splice junctions.

The loss in sensitivity with increasing database size has been a challenge in proteomics [357, 427], and even more so in proteogenomics, as large numbers of putative sequences are added, and with an included decoy database the size of an already large database is essentially doubled. Therefore, any increase in database size when identifying novel proteins should be undertaken conservatively, by avoiding the addition of any unnecessary sequences.

The size of the database can significantly affect the scoring schemes used to identify PSMs. A number of PSM scoring methods compare the best scoring match and the second best, as previously highlighted in Section 2.3.9. As many more sequences are added the difference between the best score and second best score reduces and the likelihood of the best scoring match being incorrect increases. In addition, the significance of any matches which are rare diminish with increased database size, making it more difficult to separate out true and false identifications and resulting in many true identifications being removed when FDR filtering is applied [426]. While the number of identifications to the known proteome is relatively large by comparison to the novel identifications from the proteogenomics search, the overall number of identifications across both search spaces decreases [483, 484] and has been shown in some cases to be between 30% (see Section 4.3.13) and 52% (see Section 5.3.14) with a six-frame translation [483].

In addition to a reduction in sensitivity the search time is substantially increased. This can be addressed by splitting the database into smaller separate databases, and searching each one separately, which can become problematic depending on the search tool. When results are merged the score of the best matching PSM, second best matching PSM, and their difference (delta score), need to be adjusted to reflect searching the entire database. This is a necessity for tools such as Sequest [341, 342] and InsPecT [326]. Adjusting the scores can be avoided when the search tool uses a probabilistic scoring method independent of the database, as with MS-GF+ [292]. Other approaches to improve search time include the sequence tag and spectral dictionary approaches, mentioned previously in Section 2.3.7. Examples of tools using the sequence tag approach include InsPecT [326] and GutenTag [327], with InsPecT being demonstrated in proteogenomics studies in *Arabidopsis thaliana* [81] and *Zea mays* [82], while the spectral dictionary approach implemented in MS-Dictionary (later

extended to MS-GappedDictionary [335]) has been demonstrated to quickly search the six-frame translated genome of *Shewanella oneidensis* MR-1 and *Homo sapiens* and was around 40x faster than InsPecT [324]. Other approaches to improve the database search time, include database indexing [485-487], algorithmic improvements to utilise multi-core and multi-threading of high-performance computers [488], proteomics workflows utilising the grid [489], cloud computing [490], as well as other algorithmic approaches [310, 311, 491].

There have been numerous attempts to improve sensitivity when searching large databases used in proteogenomics studies, many of which have previously been mentioned for proteomics studies in Section 2.3.9. These strategies include MS-GF+, with its rigorous approach to define the score and significance of PSMs to better differentiate true and false positives [292], combining database search results from multiple different search tools [406], re-scoring database search results using multiple sources of information from the PSMs and search parameters, using machine-learning approaches such as PeptideProphet [403], iProphet [404] and Percolator [405]. Other methods to improve sensitivity in proteogenomics searches, mentioned previously with proteomics, are to address the problem at the sampling stage by limiting the peptide fractions to a particular range of pIs, or predict the pI of sequences in the database and then search only the appropriate fraction of the database with the same pI [429] as these peptides will likely be more common. The same principle is applied by searching for only tryptic peptides when performing a MS/MS database search using the protease trypsin. An additional method, which can be applied at the sampling stage, is Ribo-seq, mentioned previously, which can significantly reduce the search space to only transcripts destined for translation [464, 465], allowing for more statistically confident identifications.

In addition, the manner in which proteogenomics search results are filtered and processed can also improve the sensitivity. Often in the past a single FDR cut-off has been applied across all known and novel results [81, 82, 443, 450, 451, 482, 492]. A more prudent approach would be to apply FDR cut-offs to the different types of search spaces, by first searching the known database and followed by FDR filtering, with any remaining unassigned MS/MS spectra used for searching the proteogenomics search space for a unified set of novel PSMs [474, 493-495]. This could then have an FDR cut-off applied across all novel PSMs, or across different classes of novel PSMs [439] (e.g. frame-shifts versus novel genes). Control of the FDR across novel PSMs has previously been described, after an initial broad 1% PSM level FDR filtering. However, real identifications may be missed even before annotation event filtering due to prior inaccurate PSM FDR filtering. Such approaches should be applied because different genomic regions have their own FPRs, just as different classes of peptides, such as tryptic, pI ranged peptides etc have their own likelihood of appearing in a sample, as mentioned previously. Compared to a broad FDR cut-off across all known and novel PSMs, applying a FDR cut-off separately on the identified known and a unified set of novel PSMs, such as those demonstrated in [474], would have the benefit of applying a more appropriate threshold for both known and novel PSMs. This would improve FDR accuracy, and by applying a more conservative approach by removing MS/MS spectra already identified as known, would avoid any possible MS/MS spectral misinterpretations among the novel peptides.

This approach to FDR filtering on known and novel datasets separately, by first removing MS/MS spectra assigned to the known proteome before searching the novel dataset, is highly presumptuous that the MS/MS spectra was correctly identified in the known proteome, which could have also been correctly identified to a novel sequence, leading to highly conservative results. However, the approach does have merit when the

dataset contains numerous variant peptides, as was applicable in a study on cancer variant peptides [474], as there would likely be higher numbers of incorrect MS/MS spectral interpretations, due to mass shifts as a result of numerous SAPS mistaken as PTMs. A similar method is applied in [496], where the sequences from the proteogenomics search database are first *in silico* digested, with any peptides found matching known protein databases such as Ensembl, NCBI and UniProt by sub-string matching removed from the proteogenomics database leaving only likely novel sequences. However, this approach neglects filtering out of spurious MS/MS spectra, and assumes that all MS/MS spectra used in a search have no missed cleavages. It also limits the analysis to only *in vitro* protease cleaved peptides, not considering possibly interesting *in vivo* proteolytic peptides which could be included in the analysis.

Another method, which can be employed to improve the sensitivity of a proteogenomics search, is the two-pass search approach, previously mentioned in Section 2.3.9. The method can be applied in proteogenomics to improve sensitivity by performing an initial search of the six-frame translation or other proteogenomics search space, as well as contaminant sequences, to identify putative sequences with high scoring PSMs with no FDR filtering and the removal of any MS/MS spectra not matching the target database or only matching contaminant sequences. A subsequent search is then performed only with sequences identified from the previous search, with the target sequences used in the TDA, followed with FDR filtering [428, 483]. This approach is able to filter out spurious MS/MS spectra and identify more PSMs by reducing the search space, and thereby increasing the significance of rare PSMs.

Other methods can also be employed to reduce the database size by identifying the most optimal sequences to search, e.g. among all frames of a six-frame translation. These include ESTScan2 [497], EORF [498], applying homology searches, predicting coding potential, and *ab initio* predictions [483]. But such approaches, particularly those

applying predictions, may inadvertently bias the search space, by applying predefined assumptions to the data.

Additional sources of false positives to those from an inflated search space, can be derived from random high scoring MS/MS spectra and matches to peptide sequences homologous to the real sequences, which may underestimate the FDR [426]. False positives may also be derived from chemically modified peptides from the known database; with a mass shift matching the unmodified novel peptides [493, 499]. To avoid the inclusion of such peptides the identified peptides from the MS/MS database search should then be searched, either by a simple sub-string search or using the Basic Local Alignment Search Tool (BLAST) against the six-frame and known database, which would identify any erroneous sequences. If, however identification of SAPs is also an aim of the analysis such sites should be marked and also screened for mass shifts explained by selected PTMs during the MS/MS database search, such as oxidation of methionine, deamidation, carbamylation, acetylation etc [82, 429, 493, 494, 499, 500]. If the specific PTMs in a sample are largely unknown, a “blind” search can be run to identify the most common PTMs [402]. But the inclusion of multiple PTMs within a proteogenomics search can negatively impact the sensitivity by further inflating the search space unnecessarily [82, 427], and so it should be done within a proteomics context after the proteome has been improved and defined. As pointed out by Tsur et al [402], for every protein there is likely at least 1 unmodified peptide, and so adding modifications to a proteogenomics search is likely to hamper attempts to discover new proteins and novel annotations.

Once a proteogenomics search is complete with PSMs identified and FDR filtering applied, there are other considerations for filtering before final analysis. Due to an upper limit on the number of identifiable PSMs imposed in a MS/MS database search, identified peptides from a PSM may not be representative across the whole

genome. Therefore, any identified peptides require mapping across the whole genome to ascertain which of the peptides are shared (co-located) and which are unique, along with their respective coordinates. In addition, another filtering consideration is any peptide which contains leucine/isoleucine, as mass spectrometers cannot distinguish between these amino acids since they are isomers with the same atomic mass, and so requires careful interpretation of the fragmentation pattern via multiple MS/MS experiments to distinguish one from the other [501]. Thus any sequences identified as leucine/isoleucine variants should be removed from the analysis.

2.4.3 Defining the level of proteogenomics annotation

In proteomics the level of annotation is at its basic level, with identified proteins and the locations of peptides within those identified proteins. In proteogenomics a broader annotation is required to characterise locations across the genome, flanking gene regions and within genes, such as novel genes, exon boundaries, TIS sites and frame-shifts, along with their peptide evidence, which should be flagged as unique or shared to indicate unambiguously that a particular genomic locus is being expressed [220]. The proteogenomics search space has a much higher false positive rate compared to proteomics search spaces, and so to identify any annotation events with confidence and to unambiguously identify the locus as being expressed, much like in proteomics with the two-peptide rule, the parsimonious presence of two or more unique peptides across the region should be a requirement, particularly for regions with a higher false positive rate, such as intergenic regions – novel genes, translated UTRs and gene boundaries. In addition, there can be multiple simultaneous annotation events inferred for the same locus, depending on the surrounding known reference annotations. For example a novel peptide inferring a frame-shift can also be identified as a novel exon or translated UTR. To account for this overlap and simplify proteogenomics annotation, an order of precedence as outlined in a previous study [82], is required.

To define a peptide as novel relies on the underlying level of completeness of a genome annotation, its quality and the underling proteome database. In some cases the annotation is lacking and therefore needs to be revisited, requiring *ab initio* and evidence-based predictions, which themselves may be lacking in completeness and quality thus introducing errors into the annotation. For example, the protein predictions may be incomplete, containing ambiguous residues, gene models may be incomplete, such as truncated exons not divisible by 3, the CDS phase information may be incorrect, outlining a discontinuity between predicted sequence and entries in the General Feature Format (GFF) files etc (for examples see Chapters 4 and 5). For these reasons a highly curated annotation, containing only high confidence annotations, and which has potentially gone through some manual curation beforehand, is most beneficial to avoid any problems further in analysis. After the initial analysis, any novel annotation events identified should be screened following consideration of their posterior probabilities [404], event probabilities [81, 82, 482], parsimony, and spectral counts [502]. In addition, in support of these annotation events or predictions inferred from the annotation events, screening can be performed using BLASTP against public repositories containing known or predicted sequences from orthologous genomes, to confirm the findings, and by identifying orthologous regions or genes in closely related genomes, annotation can also be performed on these orthologous genomes, termed ortho-proteogenomics [79, 445]. Public repositories may also contain protein sequences derived from the target genome, but not currently incorporated into the current annotation from the study due to different annotation versions or other not well-known parallel genome projects and annotation efforts.

An approach termed comparative proteogenomics [446, 448] can be used to assign confidence to novel annotations, where multiple parallel proteogenomics processes using different MS/MS spectral datasets are carried out on closely related

species, with cross-validation of any annotations applied. This is useful in the case of single peptide hits or “one-hit-wonders”, where the possibility of false positives is more likely. However, there is no currently available scoring scheme for comparative proteogenomics, and the approach has so far only been limited to bacterial studies, with manual inspection and sequence homology searches still necessary to apply confidence to proteogenomics annotations across the comparative studies.

Another approach for conducting proteogenomics is with the annotation of metagenomes using metaproteomics, termed metaproteogenomics [503]. The field of metagenomics is tasked with the sequencing, assembly and annotation of communities of bacterial species, generally from soil samples, which cannot be individually cultured, sequenced and studied on a per species basis [504]. Genome annotation in metagenomics was performed by comparison with already existing genome annotations from individual species [505], however this can heavily bias the annotations and potentially introduce errors, as was outlined in Section 2.2.2. Metaproteogenomics emerged from the introduction of metaproteomics to provide more direct evidence for the expression of proteins across a bacterial community and to an extent indicate the functionality of genes and proteins from each species within that community.

2.4.4 Proteogenomics tools

The main goal of proteogenomics is to conduct genome annotation. Over the last decade the gold standard in genomic annotation involved a concerted effort in manual annotation, often referred to as a ‘genome jamboree’, which cost valuable time and money. A high-throughput proteogenomics approach, with high specificity and sensitivity would alleviate much of the time and financial requirements involved in genome annotation efforts.

A number of proteogenomics strategies have been implemented into tools over the years, with many more tools becoming available during the last few years, offering increased complexity and sophistication. A number of these tools, with their relative advantages and disadvantages are outlined in Table 2.11.

Table 2.11 Comparison between different proteogenomics tools

A brief history of proteogenomics is encapsulated in the following twelve tools, developed over the last decade.

Tool	Advantages	Disadvantages
PeptideAtlas (2004) [354, 506]	• Uses PeptideProphet. • Better specificity due to searching curated protein databases.	• No way to identify unique peptides. • No control on FDR. • Web-based only lacking flexibility. • No six-frame translation search (limited novel discoveries). • Limited to searching protein databases; International Protein Index (IPI) [507] (now discontinued), Ensembl, Vega, RefSeq and UniProtKB. • Inferred annotations limited to peptide coordinates from known proteins.
GAPP (2006) [508]	• Downloadable by request, making it more flexible than PeptideAtlas. • Better specificity due to searching curated protein databases.	• Uses Advanced Average Peptide Score (Advanced APS) [509] to score PSMs when they share a protein match that was obtained from within the same experiment [510]. Potential to identify incorrect proteins and skew the FDR when applying the TDA. • Limited control on FDR. • No measure of significance for each PSM. • No six-frame translation search (limited novel discoveries).
Genomic Peptide Finder (GPF) (2004) [511]	• *De novo* peptide sequencing. • Maps peptides to six-frame translated genome, improving the rate of novel identifications. • Improved throughput by automation, integrating the gene prediction tool Augustus. • Identification of spliced peptides through use of ESTs.	• *De novo* peptide sequencing of MS/MS spectra has a much larger search space than other methods of MS/MS spectral interpretation, impacting sensitivity. • No rigorous control on FDR. • Limited identification of splice junctions using ESTs, as they are limited in coverage and are error prone.
PepLine (2008) [512]	• Maps peptides to six-frame translated genome, improving the rate of novel identifications. • Fast searches of six-frame translated genome by mapping peptide sequence tags (PSTs) onto genome. • PSTs are clustered to identify gene and exon-intron boundaries.	• Sensitivity limited due to peptide sequence tag approach. • Lacks rigorous control on FDR.

Tool	Advantages	Disadvantages
Genome-based peptide fingerprint scanning (GFS) (2008) [513]	• Accepts either MS1 or MS2 spectra to identify ORFs from a six-frame translation. • Flexibility with MS approach: Using MS1, the ORFs are *in silico* digested and MS/MS spectra matched to *in silico* mass peaks. Using MS2 the ORFs are matched within a precursor mass tolerance, using a sequence tag approach or a slower, but more accurate GFS HMM_Score algorithm. • The PSMs are clustered to identify loci. • Can be optionally run on a local machine or cluster.	• Not optimised for larger genomes. • Matches only filtered based on E-values, with no rigorous control using FDR. • No discrimination between unique and shared peptides. • Sensitivity limited due to peptide sequence tag approach. • No identifications of peptides spanning exon-exon junctions. • Does not use known protein sequence and annotation to infer novel and known peptides.
Proteogenomic Mapping Tool (2011) [514]	• Flexible and friendly optional graphical user interface (GUI). • Extends PSMs in a 5' and 3' direction from start to stop or splice junction to generate expressed peptide sequence tags (ePSTs), first coined by McCarthy et al [515]. • Incorporates splicing either using canonical acceptor and donor sites or using imported splice junctions from GeneSplicer [516]. • The identified ePSTs can be used to confirm transcriptional data and/or used to search for homologues in similarly related species.	• Not a fully integrated proteogenomics tool and no integrated MS/MS database search tool. • Lacks specificity, sensitivity and limited in its application. • Does a complete six-frame translation of genome. Not segmented into ORFs of a limited size, resulting in a far larger search space. • Can only map previously identified peptides given in a FASTA file. • No control on FDR or other filtering methods.
Peppy (2013) [11, 449]	• Employed as a proteogenomics tool for ENCODE. • Uses a newly developed PSM scoring approach. • Can run on a desktop for relatively smaller genomes when compute resources are limited.	• Lacks rigorous control on FDR (only 1% PSM FDR applied). • Performs a separate proteogenomics search and proteomics search, requiring manual filtering and identification of any novel PSMs. • All peptides identified. Studies indicated that only unique peptides were focussed on to identify loci. • No indication that peptide clustering was performed.

Tool	Advantages	Disadvantages
Proteomic-Genomic Nexus (PG Nexus) (2013) [517]	• Employs MS/MS database search against six-frame translation, Ensembl and RefSeq public repositories. • Converts identified PSMs into genomic coordinates within a SAM and BED file, which can then be viewed in IGV alongside RNA-seq read alignment results. • Integrated into a workflow environment (Galaxy) for wider accessibility. • Bacterial genomes are sliced into segments of a chosen size (e.g. 900 bp) and six-frame translation performed. ORFs inferred by stitching together sliced segments with peptide matches. An alternative means to cluster peptides on overly long non-genic ORFs due to high GC content bacterial genomes.	• Only limited to the MS/MS database search tool, Mascot [165]. • No rigorous control on FDR, besides that employed by Mascot, which is known to rapidly lose sensitivity and FDR accuracy with larger database searches due to its scoring function [426]. • For eukaryotic genomes, peptides are identified against Ensembl and RefSeq protein sequences, and then mapped onto the genome for annotation. • MS/MS spectra are not directly searched against the RNA-seq data. Only co-visualised with mapped peptides on the genome.
Genosuite (2013) [502]	• Combines multiple MS/MS database search tools and merges results for improved sensitivity, using Combined FDRScore [518]. • Visualisation of novel peptides mapping to genome in HTML file. • Visualisation of novel PSMs for manual validation in HTML file. • High specificity by enforcing ≥2 unique peptides, with single peptides ≥5 significant PSMs.	• Limited to only bacterial genome annotation. • Only limited to MassWiz [519], OMSSA [348], Tandem [347] and InsPecT [326] MS/MS database search tools. • No rigorous control on FDR, besides that employed by the MS/MS database search tools, and the use of a Combined FDRScore [518] when merging results. • ORF size from six-frame translation fixed at ≥50 aa. • Any identified leucine and isoleucine are considered the same, and mapped to the genome/proteome. Could introduce ambiguity to identifications. • Peptides, which are co-located in the genome, are not considered for further analysis. Only ≥2 unique peptides, with single peptides ≥5 significant PSMs, above the desired FDR threshold are considered – this improves specificity, but reduces sensitivity significantly. • Limited categorisation of novel peptides into Novel Protein Coding Region (NPCR) and changes in annotated gene models. Manual interpretation is then needed via comparisons to the NPCRs, ORF and other gene predictions. • Visual inspection in relation to the reference gene model and through sequence homology negatively impacts throughput.

Tool	Advantages	Disadvantages
PGTools (2015) [496]	• Employs MS/MS database search tools Tandem [347], Comet [520] and MS-GF+ [292]. The results of each can be combined using Percolator [405]. • Runs on multiple processors and on a cluster to enable high-throughput. • Modular approach, with some modules running as independent tools, allowing for flexibility with workflow design. • Incorporates OpenMS [521] and ProteoWizard [287] msconvert for file conversion. • Databases accessible from FASTA or SQLite formats. • Source code is freely available as open-source allowing for customisation for non-commercial use. • Uses JSON data format for reading and writing configuration files. Data outputs are in CSV, BED, SVG and HTML format. This allows for compatibility with other tools. • Usable as an automated pipeline or through a simple customisable graphical user interface (GUI). • Translation possible from 1 to 6 frames, not just all 6, for transcripts, ESTs and genome. • At multiple stages of analysis results are reported and data visualised as Venn diagrams of unique and shared peptides between search results, a zoomable treemap showing protein groups, chromosome distribution and Circos plots of identified mapped novel peptides [522], and through tools such as UCSC genome browser and the Integrative Genomics Viewer (IGV) [523]. • Separate modules for a proteomics only search with aggregation into protein groups based on parsimony and unique peptides and a genomics only search for proteogenomics annotation. • Segregates the novel search space prior to searching, by *in silico* digesting sequences from proteogenomics databases, filtering peptide lengths 7-36 aa and removing peptides identified from known proteins in Ensembl, NCBI and UniProt. • Multiple types of annotation events detectable; pseudogene, translated non-coding RNA, novel spliced peptides, translated UTR, novel gene, novel exon, frame-shift, gene extension, exon boundary, mutant peptides (indels) and gene fusions.	• Currently only human focused and requires a database for different annotation event types, such as none coding, pseudogene, UTR, mutation and fusion databases, which are all pre-made and require downloading (although scripts for database generation are available from the authors on request), making it difficult to extend to other organisms. Algorithmic interpretation of the GFF file for these regions from the six-frame translation, RNA-seq and variant calling tools would be more efficient and extendable as a general approach for proteogenomics. • The method employed to split databases to known and novel before MS/MS database searching, neglects filtering out spurious MS/MS spectra, assumes no missed cleavages, and is limited to *in vitro* protease cleaved peptides, as opposed to the inclusion of interesting *in vivo* proteolytic peptides, which could be included in the analysis. • A splice junction database was constructed from exon information extracted from a GFF file and donor and acceptor sites, joining exons together in different arrangements. Although this is able to identify previously unknown isoforms, it also has the potential to inflate the false-positive rate. A similar approach was used in an exon graph from [81], but was abandoned in favour of a FASTA splice graph database derived from RNA-seq read alignments, thus reducing the false-positive rate, by inclusion of experimental evidence, generated with user controlled stringencies. • Only considers unique peptides mapped to genomic coordinates, to improve specificity, but significantly reduces sensitivity.

Tool	Advantages	Disadvantages
Prokaryotic proteogenomics pipeline (PPGP) (2010, 2011, 2014) [443, 492, 524]	• Fast search times with sequence tag approach employed by InsPecT. • Configured to run MS/MS database search on a variety of HPC environments for parallel running on a cluster. • Uses MS-GF to re-score PSMs with MS/MS spectral probability. • Clusters PSMs based on a chosen interpeptide distance to limit ORF size for high GC genomes, which often have long –non-genic ORFs. • Allows identification of signal peptides. • Applies ORF filtering to remove low complexity peptides, non-tryptic peptides, non-unique peptides, and ORFs containing only 1 peptide (one-hit-wonders). • Reports different degrees of conflicting overlapping annotations, due to commonly overlapping genes in bacteria.	• Limited to only bacterial genome annotation. • Only limited to the MS/MS database search tool, InsPecT [326]. • Sensitivity limited due to sequence tag approach. • Filtering of identified PSMs used high stringency of spectral E-values. Improved specificity, but reduced sensitivity. • New annotations only limited to 'novel proteins', 'new starts' and expressed 'pseudogenes' among the known genes.

Tool	Advantages	Disadvantages
Enosi (2008, 2012, 2014) [81, 82, 472, 482]	• Integrates different MS/MS database search tools, initially with InsPecT [81], later used InsPecT with MS-GF to re-score PSMs [82], and has most recently been updated to MS-GF+ [6, 472]. • Two levels of FDR control: 1% PSM level FDR and event level with the event probability. • The use of an event probability is able to save single peptide identifications, which have often been delegated as "one-hit-wonders". • The number of unique peptides and the size of the peptide linkage distance for peptide clusters can be defined by the user, providing control on analysis stringencies. • A minimum stop-to-stop ORF size during six-frame translation can be defined, giving control on identification of short ORFs and database size. • Accounts for peptides spanning splice junctions using an exon graph [8, 81, 82] derived from all previously identified exons, later abandoned in favour of a FASTA splice graph used in [472, 474], derived from large RNA-seq experiments. • A larger number of annotation events are identifiable; novel genes and distal events: reverse strands, translated UTRs and gene boundaries and proximal events: frame-shifts, exon boundaries, novel exons and novel splicing. In the latest release through the use of variant peptides in a splice graph [474], indels and substitutions can be identified. • Configurable to run the MS/MS database search on a SGE cluster. • Reports annotation event types, genomic location, event probabilities, and the presence of the annotation event directly overlapping the gene of interest. All novel and known peptide identifications are provided in a GFF file, suitable for visualisation and as hints for gene prediction. • Novel peptides are verified as truly novel, by searching for those which are homeometric, i.e. those indistinguishable by mass spectrometry and which may be contained in the known proteome, and identified as novel my accident. These are then removed from the analysis after FDR filtering. Such peptides include those containing Isoleucine/Leucine substitutions and Lysine/Pyroglutamate substitutions.	• Many peptide clusters are miscategorised as gene boundary and reverse strand annotation events as a result of a fixed peptide linkage distance too large for some genomic regions, which would suit a smaller linkage distance. This also affects the number of novel genes, which can be categorised. • The latest version 1.0 only runs the MS/MS database search on an SGE cluster, splitting the database up to run on each node, and is not configured for other schedulers such as PBS Pro and SLURM. Running on these would require running the search outside Enosi within an in-house script. • Pre- and post-processing tasks are not multi-threading and do not run on a cluster, such as splice graph generation and FDR filtering. This requires some innovative approaches outside of the Enosi tool to process particularly large datasets from splice graphs and the combined raw results from a six-frame translation search. • In Enosi version 1.0 when using the built-in combined FDR strategy, filtering is applied to each result file based on the number of split databases searched and then merges them post-filtering. This introduces FDR inaccuracies, but can be avoided to a degree by manually combining results prior to FDR filtering.

Many of the tools mentioned above in Table 2.11 do not scale well for large projects, do not easily allow the capture of metadata, offer flexibility to integrate a seamless workflow, or the sharing of these workflows with collaborators, and not all of them leverage HPC clusters. However, there are a few known exceptions: 1) Enosi uses a text file containing parameters which can be shared and reused, is configurable to run on an SGE cluster, and is also accessible from ProteoSAFe [525], a web-based proteomics analysis suite; 2) the prokaryotic proteogenomics pipeline (PPGP) [524] allows the use of high performance computing clusters using InsPecT for MS/MS database searching, and using a text file for parameters, similar to Enosi; 3) PG Nexus has been integrated into Galaxy, a web-based workflow environment; and 4) the recently released PGTools, can run on a HPC cluster, and captures results at each stage of analysis in HTML reports and visual aids.

There is a need for such analysis approaches in proteogenomics, where the capture of meta-data is essential and major collaborations in genome annotation will benefit from the sharing and re-use of workflows to enable large scale proteogenomics efforts.

2.5 BIOINFORMATICS WORKFLOW ENVIRONMENTS

This section provides a background on bioinformatics workflow environments and outlines the means of enabling the wider uptake of proteogenomics in large international and national genomics projects. A number of workflow environments are discussed in detail.

The domains of genomics, transcriptomics, proteomics and proteogenomics all mentioned up to this point share a common theme, that of informatics, or more precisely, bioinformatics, which details how the processing of biological data and their interpretation are conducted. Bioinformatics itself has often been isolated to scientists

with information technology (IT) skills, who also frequently have knowledge in a wide variety of disciplines, e.g. statistical analysis, proteomics, genomics, transcriptomics and metabolomics. The bioinformatics workflow environment is a new tool that simplifies the requirements for bioinformatics analysis, allowing anyone access to tools with much of their complexity abstracted away, the only requirement being the choice of parameters, input and form of workflows. However, many of these prerequisites can become standard with particular workflows, which could be set up once by a specialist, and then reused with minimal tweaking of parameters by the end user. This paradigm is powerful, allowing much broader use of tools with minimal downtime for training.

For proteogenomics to become more accessible to the wider scientific community, anyone who can generate peptide mass spectrometry data, or who has recently sequenced a new genome, either themselves or through external sources, should be able to bring these datasets together and perform analyses without needing to employ the skills of further specialists in bioinformatics.

A number of bioinformatics workflow environments have emerged over the years [526-534], some of which are limited in functionality, flexibility, and user-friendliness and others breaking those trends and setting new benchmarks for a better and useful workflow environment. Three examples of bioinformatics workflow environments are detailed below.

Taverna [529, 530] is a desktop-based workflow environment that is employed in a highly configurable graph of interconnected processes. The execution and management of each process is written in eXtensible Markup Language (XML) called Simple conceptual unified language (Scufl). Each process must be configured in the Scufl language with error handling managed by the user, with the various processes then chained together into complex workflows with multiple branching or iterative

processes running in parallel. Each process can be run on a local machine or submitted to a web service for execution (e.g. NCBI and EMBL web services such as BLAST and ClustalW), and all actions are saved in a log file. In addition, Taverna has integrated workflow saving and offers sharing through myExperiment [535], a web site which shares workflow objects between scientists. Taverna is focused towards scientists who understand programming and probably are familiar with a command line environment, and who essentially become the system administrators of their own workflows. Consequently, many scientists not familiar with these necessary skill-sets may shy away from using such a workflow environment.

Galaxy [532] is an internet-based workflow environment, designed with genomics [531], and more recently proteomics [466] and proteogenomics [536] workflows in mind. Compared to Taverna, the skill requirements are lower and the interface is simplified, with no user requirements to have programming or system administrative skills, and with the maintenance of tools the responsibility of a separate systems administrator. However, adding new tools requires a software developer to integrate the tool using JavaScript Object Notation (JSON), which becomes problematic when many different tools are needed for various different groups. In addition, as with Taverna, all work history is recorded allowing easy access to previous run workflows and their use as templates for similar workflows, which can then be processed independently or processed together to find their union, intersect, subtraction etc. Also similarly to Taverna, Galaxy allows the sharing of workflows through myExperiment and Galaxy Published Workflows and, just as Taverna runs on a local computer, Galaxy runs on a single server on the Internet with a single location for the storage of data input and output of results. The workflow design scheme can be complex, with multiple branching and parallel processes, but it is limited by comparison with Taverna, since the

user has less control on how to configure the processes, leaving that responsibility to a programmer.

Yabi [533] is an internet-based workflow environment for various domain-specific tasks, and which uses a three tier system: 1) a frontend web application for the user interface, 2) a middleware layer for process management, tool configuration, analysis history tracking and user management, and 3) a resource manager that exposes data and then compute resources to the middleware. Yabi skill requirements are less than both Taverna and Galaxy, with much of the complexity of managing and configuring the processes abstracted away from the user and delegated to the systems administrator. The display of the workflow scheme is also simplified to a linear form, although branching and parallel processing of tasks is still possible. This allows complex workflows to be created without the workflow clutter from many multiple different processes running in parallel, and provides a more intuitive and easy to understand workflow. Yabi separates out roles to the user, systems administrator and software developer. This is designed to abstract out the complexities for the user and allow the systems administrator to focus on their role without needing to be involved in software development. Monitoring the progress of processes through Yabi can be followed through a web-browser. Yabi is also able to access data in a heterogeneous manner, which distinctly separates it from both Taverna and Galaxy, as the backend where data are stored can be in various remote locations, and accessed through SSH [537], GridFTP [538], SFTP , Amazon Simple Storage Service (S3) [539] and others. In addition, various compute resources can be used, which include Torque [540] and PBS Pro [541], residing on HPC, Grid or Cloud based machines. This architecture lends itself well to performing tasks across various compute and storage resources for analysis, without relying on any one single resource. This capacity is critical when the

user has large datasets accessible from a remote site, with the results outputted to a different remote site for review.

Yabi has also been implemented into a command line interface for users who are more familiar with the command line, and who do not wish to write scripts in order to run jobs on HPC systems. During the design of workflows each process is configured to know which types of input and outputs to expect, with an indication to the user when a particular input is required from the output of a previous process. In addition, workflows in Yabi can be saved and re-used, with plans to make these workflows sharable between different installation instances of Yabi.

All these tools lend themselves well towards proteomics and proteogenomics pipelines, which could be easily supplemented with pre-processing and post-processing stages. However, to really reach out to the scientific community, tools such as Yabi are needed to simplify the tasks by abstracting away the complexities, and allowing flexibility to where the tools can be deployed and accessed. This can be done even from remotely distant locations, so that scientists can focus on the science, instead of struggling with the tools and the logistics of processing large datasets.

3 METHODOLOGIES

This chapter outlines the key Materials and Methods used throughout this thesis and culminates in the synthesis of a roadmap for undertaking proteogenomics.

3.1 DATASETS

3.1.1 Proteomics and genomics datasets

The MS/MS spectra and genomic datasets from all four case studies (Chapter 4: Bacteria, Chapter 5: Grape, Chapter 6: Human and Chapter 7: Wheat) were obtained from a variety of different sources, including research collaborators and publically available datasets and generated through prediction tools, as outlined in the Materials and Methods sections for each case study. All protein datasets included a common source of contaminants downloaded from the Global Proteome Machine (GPM) (ftp://ftp.thegpm.org/fasta/cRAP/crap.fasta), comprising a broad range of contaminants including keratins, chymotrypsin and trypsin. Other contaminants were added if suspected to be in the samples, such as non-target proteins and other uncommon proteases if used in the study. These were added to the known protein datasets prior to MS/MS database searching.

Visualization of all gene models and their supporting evidence, including novel and known peptides, was done using the Integrative Genomics Viewer (IGV) [523]. For illustrative purposes for this thesis, GenomeTools [542] was used.

When visualizing the MS/MS spectra assigned to the known and novel peptides, and for illustration in this thesis, Lorikeet Spectrum Viewer (https://code.google.com/p/lorikeet/) was used.

3.1.2 RNA-seq datasets

For eukaryote studies, to account for alternatively spliced genes and to define a search space for spliced peptides, Illumina RNA-seq data was obtained from the Sequence

Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) through DNA Nexus (http://sra.dnanexus.com/), and also obtained from collaborators, the details of which are highlighted in each respective chapter.

3.2 DATA FORMATS

The initial step for any proteogenomics analysis is the conversion of data formats. File formats outputted from mass spectrometers are in a proprietary format and before any analysis can begin, they require conversion to an open source format such as mzXML, mzML and MGF, to be processed by the analysis tools.

The main workhorse of any MS-based proteomics analysis pipeline is the MS/MS database search tool; in this case it is MS-GF+ [407] (version v9949 2/10/2014). MS-GF+ can accept mzML, mzXML, MGF, MS2, PKL and DTA file formats. Due to the pre-processing steps undertaken for clustering and quality filtering, for each case study investigated it was necessary to convert the original file format, if not already formatted and whether proprietary or open format such as mzXML, to MGF.

The tool msconvert, from the Proteowizard package (http://proteowizard.sourceforge.net/), was used for conversion of the MS/MS spectral file format. The Windows 32-bit version of msconvert was used, which was the only free open source version available that includes vendor reader support. The msconvert tool is capable of running on the command line or through a graphical user interface (GUI).

3.3 MS/MS DATABASE SEARCHING

The MS/MS database search was performed by MS-GF+, which uses the following settings: 1) Depending on the source of the MS/MS spectra the protease was either

trypsin, chymotrypsin, Aspn or unspecified, 2) a precursor mass tolerance was then chosen, which varied depending on the case study and/or dataset used and whether the search was for the assessment of optimal precursor mass tolerance, mentioned later in Section 3.4.6. 3) The number of allowed modifications per peptide was set to 2, with modifications: carbamidomethylation of cysteine (C+57), oxidation of methionine (M+16) and protein N-terminal acetylation (+42). 4) The maximum peptide length was set to 30 amino acids (aa), 5) the isotope error range was set to "0,2", 6) the instrument was set to either high-res Time-of-Flight (TOF) (e.g. Quadrupole TOF (QTOF)) or low-res Linear Trap Quadrupole (LTQ) (e.g. Ion Trap) depending on the mass spectrometer used to generate the data. 7) The number of tolerable termini was set to 1, 8) the number of reported matches per MS/MS spectrum was set to 10, and 9) a reverse sequence decoy database was generated.

In Enosi (version 1.0), some MS-GF+ parameters mentioned above are always on by default and are not changeable (i.e. they are hard coded into the tool), such as the number of tolerable termini, maximum peptide length, number of reported matches per MS/MS spectrum, use of a reversed sequence decoy database and the isotope error range. In addition, there are restrictions to parameters such as the number and types of modifications and a precursor mass tolerance in Daltons (Da) is restricted to between 0.0 to 2.0 Da, as a mass window above this would increase search time and reduce the sensitivity of the search by considering many more possible precursor masses, which becomes a problem and more so in proteogenomics searches. However, due to the inaccuracies with low-accuracy mass spectrometers (e.g. LTQ), there is no well-defined optimal precursor mass tolerance that can be chosen, as the actual mass of each peptide ion may vary widely, resulting in the MS/MS spectra not being identified if the mass window is too narrow or too wide. In some situations, going above 2.0 Da, and looking at a range of precursor mass tolerances, e.g. up to 3.0 Da may be warranted to improve

upon coverage, as demonstrated in a later chapter utilizing low-accuracy MS/MS spectra. As a result, the database searches were performed using an in-house script, independent of Enosi, with each MS/MS spectral file searched against each database file. This step was also taken due to the limitation that Enosi required an SGE cluster for scheduling, which was not accessible to the dissertation author. Only access to SLURM and PBS Pro schedulers were available, as further explained later in Section 3.5.

3.4 PROCESSING AND PREPARATION OF DATASETS

3.4.1 Formatting of gene model and protein sequence datasets

Datasets from the various case studies required varying degrees of modification depending on the state of the publicly available annotations or the output from prediction and annotation tools. For the proteogenomics pipeline to be compatible with the GFF file an “mRNA” line must be present preceding the “CDS” line, while a “gene” line is not critical for processing. In cases where only “gene” and “CDS” lines exist, the word “gene” from column three, can be replaced with “mRNA”. Also the protein FASTA headers and/or Parent identifiers in the GFF file were modified to ensure that they were the same. The “Parent” identifier for mRNA was also required to be the same on the corresponding CDS and exon features. Additionally, to assist downstream annotation, where the “ID” and “Parent” identifiers were too generic or inconsistent across all features, they were modified to include the gene name and/or accession for easy identification.

In some instances, gene feature coordinates were found to be in the wrong orientation on the reverse strand (e.g. 5’ to 3’ on the forward strand and 3’ to 5’ on the reverse strand). In particular, for Enosi to determine known mapped peptide coordinates, CDS and exon coordinates must be in a 5’ to 3’ orientation on the reverse strand (i.e. the highest coordinate to lowest coordinate), and the reverse for the forward

strand. In instances where GFF files contained CDS and exons in the wrong orientation, an in-house script, utilizing GFFutils (https://github.com/daler/gffutils), was used to parse the GFF file and to output all gene features in a 5' to 3' orientation, regardless of which strand was implicated. In addition to the gene feature orientation, modifications to the GFF files were carried out with a series of Unix commands.

3.4.2 Pre-processing of RNA-seq datasets

All Illumina reads were first checked for quality using FASTQC [543], followed by trimming using Trimmomatic [544], with quality score thresholds set to an average of 20 across a 4 bp window. Minimum length was set to 20 bp and Phred scores were set to 33 (Illumina 1.8+) or 64 (Illumina 1.5+) depending on the reads. The first 15 bp of the reads were cropped to remove any Illlumina bias present, and the appropriate Illumina adapter sequences (either TruSeq2 or TruSeq3) were chosen in each case depending on the sequencing method used. All unpaired reads were merged and labelled single end (SE). All trimmed reads were then gzipped. FASTQC was then re-run on the trimmed reads to confirm the quality of the reads before alignment to the genome.

3.4.3 RNA-seq alignments

All alignments to the respective genomes in each case study were carried out using the Spliced Transcripts Alignment to a Reference (STAR) aligner (version 2.3.1x) [127], using the two-pass scheme which was conducted using an in-house bash script running on a Cray XC30 supercomputer (Magnus, from the Pawsey Supercomputing Centre https://www.pawsey.org.au/). Magnus consists of two compute cabinets, each holding 52 compute blades, with four nodes per blade and each node supporting two 8-core Intel Sandy Bridge Xeon E5-2670 processors running at 2.6 GHz, with 64 gigabytes of DDR3-1866 per compute node, at 1.6GHz. Each STAR job on the supercomputer was assigned 16 cores and 63 GB RAM and ran within a 12 hour walltime. STAR was run with 16 threads, and the parameters "genomeSAindexNbases", "genomeChrBinNbits",

"genomeSAsparseD", and "limitGenomeGenerateRAM" were set to "7", "10", "20", and "63000000000", respectively (optimised for aligning to particularly larger, highly fragmented genomes). The two-pass scheme was implemented by using the splice-junctions output from the 1st pass, as input for the 2nd pass to improve upon the sensitivity of splice junction detection. STAR was chosen by comparison with other alignment tools such as TopHat2 [126] due to its improved alignment speed, sensitivity and reporting fewer false positives [545].

All alignment results from paired end and single end reads were in Sequence Alignment/Map (SAM) format, which were then subsequently sorted by coordinate and converted to sorted Binary Alignment/Map (BAM) format files using the tool SortSam from Picard tools (http://picard.sourceforge.net/).

3.4.4 Preparation of proteogenomics splice database

All sorted BAM files from the STAR alignments were then merged together into a single BAM file using SamBamba (http://lomereiter.github.io/sambamba/), which is a SAM and BAM tool capable of quickly merging many SAM and BAM files into one larger file with relatively little memory overhead. The Picard tool MarkDuplicates was then run to filter any PCR and optical duplicates from the merged BAM files. In some cases, where the BAM files were particularly large with many sequences in the header, MarkDuplicates was run before merging the BAM files using Sambamba. The resulting BAM files, with redundant read alignments removed, were then converted to a SAM file using SAMTools [546].

Packaged with Enosi (version 1.0) is a script called BuildSpliceGraph (version 0.4.11), which takes SAM files and a genome FASTA file. Using the SAM file mentioned above, BuildSpliceGraph was run to generate a SpliceInfo file, with a maximum peptide length set to 30 aa and a minimum number of spliced reads set to 2.

Finally, a splice FASTA file was generated; representing all spliced peptides and containing genomic coordinates in the FASTA header. This splice FASTA file was then used as a splice database in the proteogenomics analysis. In instances where the genome was very large and/or fragmented, resulting in very large numbers of splices to process, each splice from the SpliceInfo file was separated out and run in a massively parallel manner on Magnus, mentioned previously in Section 3.4.3 for RNA-seq alignment.

3.4.5 Preparation of six-frame translation database

A stop-codon-to-stop-codon translation of the genome was generated in all six frames. This was undertaken using the Enosi tool, using 30 aa as the minimum Open Reading Frame (ORF) size and with the minimum file size for each file set to 500 Mb, unless otherwise specified for each specific case study.

3.4.6 Pre-processing MS/MS spectra and MS/MS database search optimization

In each case study the total set of MS/MS spectra were clustered using MS-Cluster [312], thus improving the signal to noise ratio by merging MS/MS spectra and removing redundant MS/MS spectra matching the same peptide, which improved the overall quality of the MS/MS spectra and reduced the number of aberrant MS/MS spectra attributing to false positives. This also reduced the overall dataset size, resulting in improved search times. PepNovo [302] was then used to quality filter the clustered MS/MS spectra. Due to the limited MS/MS spectral dataset sizes, the level of quality filtering was optimised for each dataset to reduce losses. Comparisons using the known proteome, were made between the original MS/MS spectra, clustered MS/MS spectra, and clustered+quality filtered MS/MS spectra across a range of quality score thresholds (PepNovo scores 0.001, 0.005, 0.01, 0.05, 0.1, and 0.2), to ascertain the most appropriate pre-processing procedure to use before undertaking proteogenomics analysis.

Often in peptide mass spectrometry users default to around 20 ppm with high accuracy Orbitrap, Fourier Transform (FT) and TOF machines, while 2.0 - 4.0 Da is often the choice for lower accuracy LTQ machines. Additionally, the accuracy of any one machine can alter over time, over different mass ranges (depending on the machine), and can also differ from day to day. In keeping with this trend and to keep the choice of precursor mass tolerance consistent while assessing quality filtering and clustering it is best to default to using a precursor mass tolerance of 20 ppm for high-accuracy MS/MS spectra and 2.0 Da for low-accuracy MS/MS spectra. Comparisons of all results between the original MS/MS spectra, clustered MS/MS spectra, and clustered+quality filtered MS/MS spectra, were then made, taking the largest number of peptide-spectrum matches (PSMs) <5% peptide false discovery rate (FDR), and the most appropriate pre-processing strategy was chosen for further proteogenomics analysis.

For most of the case studies in this thesis, a high accuracy mass spectrometer, such as the LTQ Orbitrap and QTOF was used to generate the MS/MS spectral datasets. The high accuracy MS/MS spectral datasets are well suited to applying a precursor mass tolerance optimization step. The use of high accuracy machines is frequently ignored or forgotten when performing MS/MS database searches [161] and it is often prudent to optimize MS/MS database search parameters before conducting the analysis proper [547]. A similar approach was taken with the proteogenomics analysis, by using the high sensitivity and precision of MS-GF+ in conjunction with the choice of an optimal precursor mass tolerance to reduce false positives and improve peptide discovery rates, while maximizing on the identification sensitivity of proteogenomics annotations. The choice of precursor mass tolerance also affects the sensitivity of MS/MS database searches. By increasing the error window, particularly for the larger fixed value errors in Da often used with low-accuracy mass spectrometers, such as with an LTQ, the

search space and search time is essentially increased, which consequently reduces the sensitivity of the search (up to a point dependent on the accuracy of the mass spectrometer). This fact is reflected in Enosi as outlined in Section 3.3, where Enosi purposely restricts the choice of precursor mass tolerances to between 0.0 and 2.0 Da when selecting Da as the precursor mass unit.

Assessment of the optimal precursor mass tolerance for each dataset was conducted over a number of ranges (generally 0.5 - 5.0 Da for low-accuracy and 0.5-150 ppm for high-accuracy) to deduce the optimal precursor mass tolerance for the pre-processed MS/MS spectral dataset, as before, taking the largest number of PSMs <5% peptide FDR.

MS-Cluster was used to cluster the MS/MS spectra as previously outlined. However, in cases where the number of MS/MS spectra was less than 100,000, clustering was not performed as this would prove ineffective and it would probably result in MS/MS spectral losses. PepNovo on the other hand uses a machine-learning approach based on prior MS/MS spectra on which it was trained on during its development [314], looking at peak intensities and numbers. The training data were likely derived from cell lysates, which would have a higher signal-to-noise ratio, higher peak intensities and large numbers of mass peaks compared to samples derived from 1D or 2D gel digests. Therefore, any MS/MS spectra derived from similar sources to cell lysates, instead of 1D gel digests, would be more suitable for quality filtering. PepNovo applies a quality score to each MS/MS spectrum, which is then applied to filter the MS/MS spectra. As a result, some case studies with MS/MS spectra derived from 1D and 2D gels were not quality filtered, and similarly, when there were a limited number of MS/MS spectra, no clustering was performed.

The original MGF files and also the single merged larger MGF file following clustering (and quality filtering where applicable) were split up into separate smaller files each consisting of 65,000 MS/MS spectra, using an in-house MGF splitting script on a local machine or 20,000 MS/MS spectra each, for running on a cluster with a 24 hour walltime (MS/MS spectral numbers below 1,000 were not recommended as this would lead to significantly lower accuracy FDR calculations). The original MGF files were also split, as some MS/MS spectral files can be too large for tools like MS-GF+ to process effectively or at all. Enosi was designed to run directly on a SGE cluster, where it would split up the MS/MS spectral files based on the number of compute nodes specified, but when the number of compute nodes is few or is only ever run on a single node, each MS/MS spectral file may still remain too large to process effectively.

To evaluate the different pre-processing strategies and precursor mass tolerances, an in-house script was used along with MS-GF+ (version v9949 2/10/2014). The in-house script contains parameters as explained in Section 3.3. A decoy database was created by reversing the sequences from the target database and then was combined with the target database and indexed using a suffix array. This was done automatically by MS-GF+ otherwise, depending where the search was run e.g. on a cluster, then the generation of decoy sequences and indexing of the databases were performed as separate jobs. If running multiple split MS/MS spectra across multiple split databases on a cluster, the results were then merged and a 1% PSM FDR applied. This was applied using the ComputeFDR function from MS-GF+, with the spectral E-value used as a score. Calculations of the peptide FDR and protein FDR were then determined based on the method outlined in [411]. In addition, the number of PSMs, number of non-redundant peptides, and number of non-redundant proteins were then calculated prior to and post-FDR filtering. Ultimately, further control on the FDR is employed within each case study, either prior to proteogenomics analysis using the more traditional combined

FDR strategy (see Section 3.5) or the two-stage FDR strategy (see Section 3.5.2), or during proteogenomics analysis by the event probability.

Other parameters to consider for optimization included post-translational modifications (PTMs). The inclusion of PTMs has the effect of also increasing the search space and reducing sensitivity. A more prudent approach is to leave the search for protein polymorphisms and PTMs to only proteomics searches and focus on what is most appropriate for proteogenomics in the context of improving sensitivity by maximizing the number of new annotations. As was pointed out by Tsur et al [402], for every protein there is likely at least one unmodified peptide, and so by adding further modifications to a proteogenomics search will only hamper attempts to discover new proteins and novel annotations. As such, the selection of appropriate PTMs for the MS/MS database search were kept to the default throughout this thesis: carbamidomethylation of cysteine (C+57), oxidation of methionine (M+16) and protein N-terminal acetylation (+42).

3.5 PROTEOGENOMICS PIPELINE

A proteogenomics pipeline was customized utilizing the proteogenomics tool Enosi (version 1.0) [8, 437, 548, 549]. Enosi incorporates the highly sensitive and accurate MS/MS database search tool, MS-GF+, and also has the ability to search an RNA-seq derived FASTA splice-graph database [549].

For all case studies, the appropriate pre-processing strategy and precursor mass tolerance were selected based on an evaluation using the known proteome, as outlined above in Section 3.4.6. The clustered and quality filtered MS/MS spectra were then split using an in-house MGF splitting tool into 65,000 MS/MS spectra when running on a local machine or 20,000 MS/MS spectra on a cluster, to remain within a 24 hour walltime.

The Enosi tool assigns different annotation events to the identified peptide clusters: novel genes, identified outside the boundaries of the peptide linkage distance, distal events, identified within the boundaries of the peptide linkage distance, but outside the boundaries of a gene and proximal events, identified within the boundaries of the peptide linkage distance and within the boundaries of a gene. The novel gene events consist of novel genes, while distal events consist of gene boundaries, translated UTRs and reverse strand events, and proximal events consist of frame-shifts, exon boundary, novel exon and novel splice events (when applicable for eukaryote studies).

The parameters selected for Enosi included a six-frame translation of the genome generated as outlined above in Section 3.4.5, a minimum cluster size (total peptides per cluster) set to 1, a minimum number of unique peptides per cluster set to 1 and an initial minimum event probability of 90%, before further manual filtering among the different annotation event types. Through MS-GF+, the Enosi tool automatically generates a decoy database of reversed sequences with both target and decoy databases combined and indexed using a suffix array. As part of the pipeline, the MS/MS database search parameters were set as outlined in Section 3.3, with the majority of case studies applying a two-pass search approach outlined below in Section 3.5.1. The final MS/MS database search results were then either all merged in a combined FDR strategy or split into novel and known PSMs, termed a two-stage FDR strategy [439, 474], as outlined below in Section 3.5.2. A ≤1% PSM FDR was then applied, with local FDRs for each PSM calculated. For the combined FDR strategy, following FDR filtering all identified peptides were then grouped into novel peptides and known peptides by comparison to the known proteome.

Using an in-house script, all PSMs identified from the known proteome were divided into two confidence levels: all proteins that contained a mapped peptide and all proteins that contained ≥2 peptides with 1 unique peptide. This division was done to

remove ambiguity due to the protein inference problem and resulted in a list of high confidence known protein identifications. All novel peptide identifications were then mapped to their genomic coordinates, clustered according to a peptide linkage distance with each cluster inferred as a different annotation event, as mentioned previously. All annotation events were then screened through a combination of methods; e.g. event probabilities, multiple unique peptides, sequence homology to curated proteins, peptides mapping to the known proteins being annotated, parsimony and spectral counts, which can provide an estimate of protein abundance that correlates well with transcript abundance, if high enough protein coverage from the known peptides is achieved. The novel events passing screening were then interpreted to infer gene model changes (Figure 3.1), which were then visualized in a genome browser, such as IGV, and if a eukaryote study, incorporated as hints for gene prediction using the gene prediction tool Augustus [101, 102].

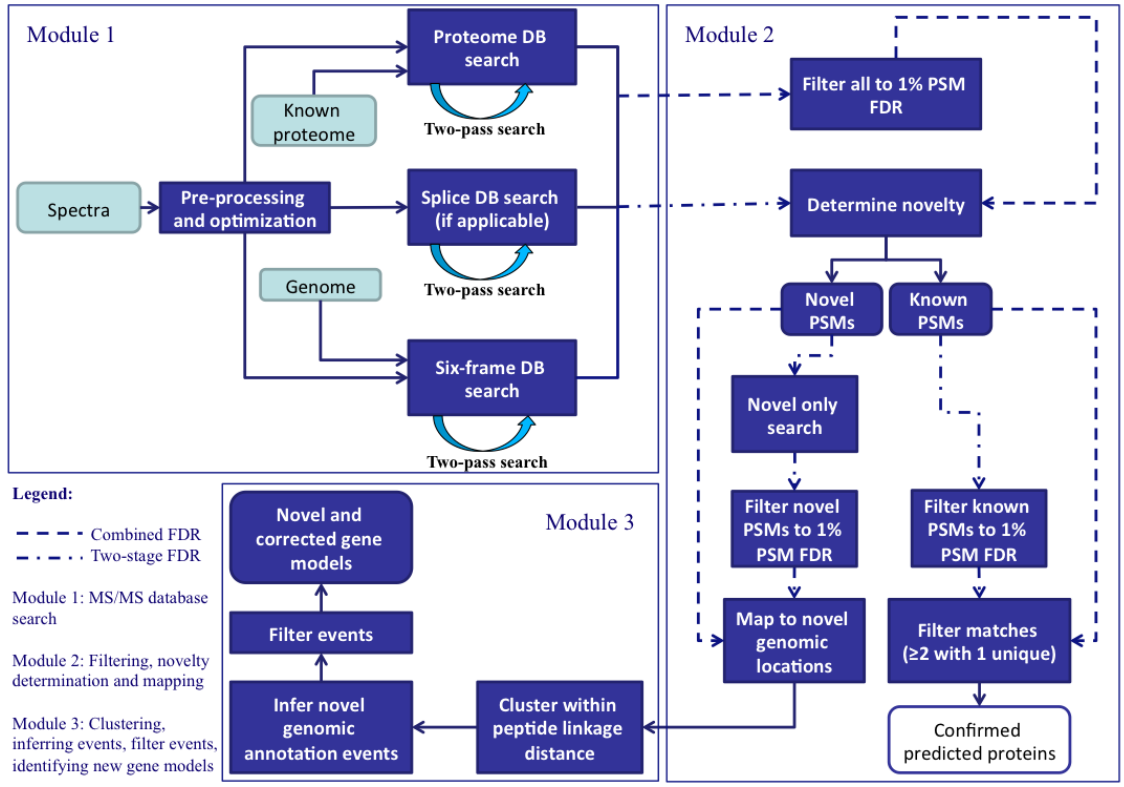


Figure 3.1 Customized proteogenomics pipeline

In module 1, a MS/MS spectral pre-processing and optimization step was carried out, following a MS/MS database search with an optional two-pass search implemented against the six-frame translation of the genome, known proteome, and where applicable, a splice graph. In module 2, FDR filtering was applied to results, in either a combined FDR strategy (dashed line) or a two-stage FDR strategy (dash dot line), on the known results and novel results separately. Identified known proteins were filtered by parsimony (≥2 peptides) with 1 unique peptide. Novel peptides were mapped to the genome. In module 3, novel peptides were clustered within the peptide linkage distance, annotation events inferred, then filtered based on event probability, number of unique peptides, sequence homology, presence/lack of multiple peptides mapped to the known proteins being annotated, and spectral counts, followed with identification of novel and/or refined gene models.

It should be noted that the current version of Enosi (version 1.0) has some inaccessible parameters hard-coded into the tool, such as the number of tolerable termini, restricted precursor mass tolerance in the range of 0.0 to 2.0 Da (with precursor mass tolerance in ppm being unrestricted), and maximum peptide length. The number of reported matches per MS/MS spectra, use of reversed sequence decoy database and isotope error range are hidden, but are always on by default, including a restriction on the allowed number and types of modifications. Also when selecting Da as the precursor mass unit, Enosi purposely restricts the use of a precursor mass tolerance to between 0.0 and 2.0 Da. As previously noted, this is to limit the search space and search time during proteogenomics searches, as precursor mass tolerances in Da much higher

than 2.0 Da would inflate the FDR and significantly reduce the sensitivity of PSM identifications, and consequently the number of identified proteogenomics annotations. However, in some cases the limit of 2.0 Da was purposely exceeded, as outlined previously in Section 3.3.

The following proteogenomics specific parameters were chosen: 1) annotation events were initially filtered based on 1 unique peptide per cluster, 2) a minimum cluster size of 1 peptide, 3) a minimum event probability of 90% and 4) the peptide linkage distance, which was based on the minimum intergenic distance for particularly small compact genomes such as prokaryotes or the size of the majority of genes (>95%) in eukaryotes. Other more general parameters included the selection of the MS/MS instrument and precursor mass tolerance, which was based on an assessment of the MS/MS spectra outlined previously in Section 3.4.6.

The selection of the peptide linkage distance is important because it impacts how annotation events are defined, more so for novel gene and distal events like gene boundary and reverse strand events, as these can be identified further afield from the peptide cluster, depending on the size of the peptide linkage distance. A larger peptide linkage distance would reduce the number of novel genes found, and increase the number of distal events, while reducing it would identify more novel genes and fewer distal events. As such, the peptide linkage distance was considered carefully. Importantly, during proteogenomics annotation any identified distal events such as gene boundary and reverse strand events could possibly be novel genes as the peptide linkage distance is a fixed variable, unlike reality where the intergenic distances across the genome can vary widely.

Once all peptide clusters were assigned as novel gene, distal events or proximal events, they were further filtered manually based on ≥2 unique peptides per cluster

and/or event probabilities ≥99.9%, which were considered as high confidence annotation events. Any other proximal events were accepted based on a relatively high event probability of ≥99.8%. In addition, all annotation events were further screened, particularly to validate single unique peptide annotation events by: 1) the presence/lack of unique and shared peptides mapping to the known proteins being annotated particularly for the proximal events, with consideration of their frame in relation to the novel peptides; 2) spectral counts; and 3) sequence homology to protein sequences in public protein repositories such as NCBI NR, NCBI RefSeq protein and/or NCBI SwissProt, to identify any annotation events already validated within protein repositories [439]; and in some cases 4) the peptide length was a consideration which would impact significance when searching protein repositories. The presence/lack of mapped unique and shared peptides to the reference known proteins being annotated were determined using a customized script, with manual checking of the frame of these peptides to see if any peptides in close proximity to the novel peptides were in a different conflicting frame, which was important for proximal events.

To improve specificity of annotation event identification and ensure that unique loci were identified, peptide clusters with at least 1 unique peptide only were accepted. The same principle was applied to improve specificity and remove ambiguity with protein identifications. This is a similar concept to using proteotypic peptides as a unique signature for a single protein in peptide mass spectrometry [550], often used in Selective Reaction Monitoring (SRM) in targeted proteomics [551]. This principle is therefore applied by using sequence homology searches of the unique and novel peptides, using BLASTP against NR, RefSeq protein and/or SwissProt to reduce the ambiguity while screening. Other information found from the sequence homology search such as the same chromosome and genomic region identified from the proteogenomics analysis, preferably with support from EST, mRNA and/or protein,

with 100% coverage and identity particularly for proximal events, was considered as conclusive evidence. However, for one early case study in a bacterial genome stringent event probabilities alone were used for novel genes, distal events and proximal events, with their selection based on cross correlation between different bacterial gene prediction tools.

In each case study the current annotation of the genome was provided as a reference. These were provided as GFF file and protein predictions as FASTA sequences, and where necessary were modified as outlined in Section 3.4.1.

For larger proteogenomics studies which required parallel computing, Enosi utilizes a Distributed Resource Management Application API (DRMAA) scheduler to coordinate each run of MS-GF+. However, throughout the course of this thesis, a DRMAA scheduler was not accessible from any of the available clusters provided by the Pawsey Supercomputing Centre (http://www.pawsey.org.au/), with PBS Pro and SLURM the only installed schedulers. To work around this issue, MS-GF+ searches of the six-frame translation, proteome and any splice graph databases were carried out using an in-house script which would mimic the initial stages of the Enosi proteogenomics pipeline, namely database indexing and the MS/MS database search. The Pawsey Supercomputing Centre's supercomputer Zeus was utilized for this task due to its 24 hour walltime and large number of relatively freely available processors due to its smaller user base and subsequent shorter queue wait time, and a submission limit cap of 96 jobs (although not all running concurrently). The Zeus supercomputer has 29 nodes, each with two 8-core Intel Xeon E5-2670 CPUs, and between 128 GB to 512 GB RAM per node. Enosi is capable of running the proteogenomics workflow from any stage in the analysis. This facility was utilised after receiving the MS/MS database search results, which were subsequently all merged with 1% PSM FDR applied (combined FDR) or merged based on the known proteome search space and novel

proteogenomics search space (two-stage FDR), with 1% PSM FDR applied to each (Figure 3.1). During FDR filtering each PSM had its local FDR calculated, which was run on a different machine and did not require parallel processing.

In instances where the merged results, using the combined FDR strategy, gave a large tab-separated value (TSV) file, upwards of 12 gigabytes (a consequence of searching a large six-frame translated genome and/or splice graph database), the files were split into smaller result files based on the MS/MS spectral files, as filtering on the FDR and calculating local FDR for each PSM on a single large result file extended processing time significantly. It was found that the processing time could be reduced to around a few days or up to a week per file if the results were merged into 2 (see Section 6.2.5) or 4 (See Section 5.2.5) TSV files, which differed depending on the dataset. Results derived from low-accuracy MS/MS spectra took less time to process than high-accuracy MS/MS spectral results, as more highly significant PSMs are usually found in high-accuracy data. However, such a strategy can result in inaccurate FDR calculations and shifts in the eventual event probabilities as the same peptide matches can be distributed across multiple different search result files leading to slight biases with some results. As a consequence, a balance between processing time versus accuracy of FDRs and event probabilities was required, limiting all the results to a few result files at most, a necessary compromise if results are to be obtained in any reasonable time frame. For results obtained from the two-stage FDR strategy, no splitting of result files was necessary, as this strategy implemented a filtering of known and novel PSMs and consequently reduced the final result file sizes. This issue indicates a need for new processing approaches or algorithms to process larger sets of results, often found from combined FDR approaches, in a faster and more accurate way without subdividing the results.

As outlined previously [548], a hybrid approach to proteogenomics is often suitable for larger more gene-sparse genomes, where a combination of event probabilities and unique peptides per cluster are used for different annotation event types. In such a case, novel genes and distal events require at least 2 unique peptides per cluster with more relaxed event probabilities, while proximal events require 1 unique peptide with more stringent event probabilities being applied and with the peptide linkage distance set to a distance represented by the majority of genes (generally >95% genes). The rationale behind this judgement is that within large intergenic spaces, such as in eukaryotic genomes, at least 2 unique peptides are more likely to be found and the possibility of finding false positives is much higher. Conversely, within the intragenic space, around annotation events such as frame-shifts and exon boundaries, more than 1 unique peptide is less likely to be found due to the confined region and false positives are less likely to occur. When MS/MS spectral datasets are very large such an approach would likely correlate well with accepting at least 2 unique peptides for novel gene and distal events, but when the MS/MS spectral datasets are small, many annotation events may simply lack a second unique and novel peptide due to low coverage. In such instances, a single unique peptide could be accepted when other evidence suggests a likely valid annotation event, such as was previously described. Within the context of the proteogenomics analysis, the peptide linkage distance defines intergenic and intragenic spaces. The linkage distance determines how large the peptide clusters can become (the intragenic space), while it also defines the distance between peptide clusters (the intergenic space). When applying this same principle to smaller, compact genomes such as bacteria, it is more prudent to set the number of unique peptides per cluster to 1 and apply a more stringent event probability to novel genes and distal events, and the same stringency or less for proximal events, while setting the peptide linkage distance to a suitably smaller intergenic distance.

3.5.1 Two-pass search approach

A two-pass search approach was applied in a number of case studies, similar to the approach demonstrated in [428], to improve the sensitivity and identification rate of PSMs and proteogenomics annotations. In the first-pass the six-frame translated genome with added protein contaminant sequences was searched against using MS-GF+, with parameters as outlined in Section 3.3 and the exception that no decoy database was added to the target database and searched. This was repeated for the known proteome database and splice graph (if applicable).

Any PSMs that matched the six-frame translated genome, known proteome or a splice graph, were identified, while any PSMs that only identified contaminant sequences were discarded. No FDR filtering was applied to any matches during this first-pass search. In an earlier bacterial study (Chapter 4) all matches were taken from the first-pass. However, in later studies (Chapters 6 and 7), since the MS/MS database searches are taking the top 10 matches, it was realized that these could also include very low significant matches. These matches were subsequently removed, keeping all PSMs below a spectral E-value of 1.0E-05. By comparison, a 1% PSM FDR roughly equals to a spectral E-value of 1.0E-10 and so this retains all significant matches, while removing all insignificant matches unlikely to be retained below the 1% PSM FDR threshold. The choice of a 1.0E-05 spectral E-value was a relatively arbitrary choice, and could be adjusted if there was supporting evidence to do so.

A new MS/MS spectral dataset for each database, and a new target sequence database were created from the first-pass search, based on the accepted PSMs and using an in-house script. The new MS/MS spectral dataset was then used to search the new target databases, along with their decoys in the second-pass search during a normal run of the proteogenomics pipeline.

3.5.2 Two-stage FDR strategy

The two-stage FDR strategy was outlined in [439], with similar strategies in [474] and [552]. Instead of applying a 1% PSM FDR across all proteogenomics results, the FDR was applied across the different search spaces. In [474] this was applied to the known proteome and a unified novel set of proteogenomics results, with MS/MS spectral identifications in the known set removed from the novel set before MS/MS database searching of the proteogenomics search space. However, this can be overly conservative, as a MS/MS spectral identification first identified in the known set does not necessarily represent the most 'correct' identification for that MS/MS spectrum. Although when the proteogenomics search space contains a variety of putative variant peptides as demonstrated in [474], this can be justifiable to reduce the potential for incorrect identifications with Single Amino acid Polymorphisms (SAPs) being misidentified as PTMs. Similarly, in [552] prior to MS/MS database searching, the sequences from the proteogenomics search space are *in silico* digested, with the peptides identified from the known search space by sub-string matching and subsequently removed to create a database of likely novel peptides. However, this does not filter out any spurious MS/MS spectra, applying the full MS/MS spectral dataset to both known and novel search spaces. This approach also assumes that all MS/MS spectra used in a search have no missed cleavages, which would limit coverage particular for non-tryptic ended peptides at the proteins N-terminus, and also limiting the digestion to a number of well-known proteases and it would also limit the analysis to only *in vitro* protease cleaved peptides, which would miss *in vivo* proteolytic peptides if they were of interest.

In this thesis a different approach was taken with PSMs identified after filtering the MS/MS spectra through a two-pass search approach, removing spurious MS/MS spectra and reducing the search space to high confident matches, where the search space

was then divided into novel and known sequences. All PSMs that identified a novel peptide had a 1% PSM FDR applied, with the same applied to a separate set of known PSMs identifying known proteins. This allowed more accurate FDR calculations for both the ‘known’ search space and the ‘novel’ search space. Some MS/MS spectra may be interpreted as novel and known, with possibly differing spectral E-values, which this method does not account for. However, this method does improve on the accuracy of the FDR, and does not make any assumptions of a MS/MS spectrum belonging to ‘known’ or ‘novel’ peptide sequences, unlike the more conservative approach [474].

For Enosi to determine novelty prior to FDR filtering, a different approach to the workflow was required. Firstly, the results from the known proteome search had a 1% PSM FDR applied. Following this, the results from the six-frame and splice graph search are merged and filtered to 100% PSM FDR to identify all possible PSMs in a format which is compatible with Enosi’s determine novelty stage. All PSMs identified at 100% PSM FDR were then split into known and novel by Enosi. Any PSMs identified as a novel PSM are then checked to ensure they fall below a 1.0E-05 spectral E-value, after which the identified novel MS/MS spectra and novel sequences were used in a novel-only MS/MS database search, followed by the application of a 1% PSM FDR. The remainder of the workflow is as normal, which uses a known 1% PSM FDR filtered set of PSMs and novel 1% PSM FDR filtered set of PSMs. This essentially reduces the size of the proteogenomics search space to only PSMs identified as novel, improving the accuracy of the FDR filtering for both the novel and known peptides.

Some MS/MS spectra can be identified in both the novel and known search spaces as a result of different MS/MS spectral interpretations with different spectral E-values in the top 10 matches. Future work could see the use of a spectral E-value comparison between the novel and known PSMs, to more accurately assign a PSM to a

novel or known peptide sequence, with any ambiguous matches with the same spectral E-values discarded.

3.5.3 Gene prediction

Once the novel annotations were filtered and reviewed, the gene prediction tool Augustus [102] (version 3.02), was used for studies involving eukaryotic genomes. Depending on the study, Augustus was either first trained (generating gene models) or ran with the default gene models, the details of which are presented within each chapter.

Once the novel annotations were filtered and reviewed, the gene predictions were improved using the novel peptides as hints. Further hints generated from other extrinsic supporting evidence were also generated for some case studies, further outlined in their respective chapters.

The hints generated from EST, RNA-seq and the current annotations were designated group name ‘E’ with associated default weighting, and the hints from any repeat regions were designated group name ‘RM’ with associated default weighting. All novel peptides were designated group name ‘M’, with the default high significance weighting to force incorporation of the hints. The hint files were then combined and used as hints during the Augustus gene prediction. Augustus was run using the parameters listed in Table 3.1.

It is worth noting that Augustus can predict more genes and proteins than there actually are due to two factors: 1) a fragmented genome can cause Augustus to predict two or more incomplete genes across contigs and scaffolds and 2) given the available evidence, due to possible incomplete coverage, predictions may be split into two or more genes and proteins. These two points need to be considered when reviewing any Augustus predictions.

Table 3.1 Applied Augustus parameters

Parameter	Value
--strand	both
--genemodel	complete
--AUGUSTUS_CONFIG_PATH	Path for the trained gene models (generated or default)
--extrinsicCfgFile	Config file with extrinsic weights for 'E', 'RM' and 'M' group names.
--singlestrand	true
--altnernatives-from-evidence	true
--gff3	on
--print_utr	on
--protein	on
--codingseq	on
--species	Set to the species name
--hintsfile	Hints file containing novel peptides (group M), any available extrinsic hints, current annotations (group E), and any repeat hints (group RM)
--allow_hinted_splicesites	atac

4 BACTERIAL PROTEOGENOMICS

4.1 INTRODUCTION

Bacteria have long been the dominant form of life on Earth, since they first evolved some 3.5 billion years ago. They have since diversified into many forms as they adapted across almost all environments across the globe. They are an essential component of our biosphere, with some species playing a role in the formation of nitrates in the soil, a task that nitrogen-fixing bacteria (rhizobium) perform in the root nodules of host plants, others contribute to disease (e.g. *Mycoplasma pneumoniae* causing pharyngitis, bronchitis, and pneumonia), and still others can contribute to a healthy digestive tract (e.g. *Lactobacillus acidophilus*).

Understanding how different bacteria function is important, so that diseases can be treated, soil for agriculture can be improved using nitrogen fixation, and our overall wellbeing improved. The biotechnology revolution over the last few decades has benefited greatly as a direct result of discoveries from bacteria, such as the use of DNA polymerase from thermophilic bacteria used in the DNA polymerase chain reaction (PCR) [553], a staple technology used in the amplification of DNA for sequencing. In addition, a new technology discovered from bacteria has emerged, called clustered regularly interspaced short palindromic repeats (CRISPR) [554], based on the innate immune response of bacteria to recognize viruses, which claims to be able to directly edit DNA *in vivo*, potentially allowing for the correction of genetic diseases.

To better understand the underlying mechanisms that underpin CRISPR, nitrogen-fixation, pathogenicity causing disease and others, within bacteria, genome sequencing and annotation efforts are required. The throughput and quality of bacterial genome sequencing has improved significantly over the years. For example, Illumina sequencing technology can now sequence 100 bacterial sequences simultaneously [555].

However, the throughput of genome annotation strategies has not been able to match, and bacterial genome annotation strategies often involve the use of *ab initio* gene prediction tools and sequence similarity, which are prone to errors and inconsistencies, as was introduced in Section 2.2.2. Therefore, a new methodology, capable of maintaining high throughput and quality of genome annotation is needed.

This study applies a new approach towards genome annotation, termed proteogenomics, and also demonstrates the power that repurposing legacy proteomics data has for improving the genome annotation of the nitrogen-fixing bacteria *Bradyrhizobium diazoefficiens*.

Agriculture requires a constant supply of nitrates to the soil to allow crops to grow and achieve high yields. This often requires artificial means using fertilizers. Unfortunately, fertilizer run-off into water sources such as estuaries causes problems with toxic algal blooms resulting in anoxic rivers leading to a large die off of fish populations. By understanding the molecular mechanisms underpinning the processes for nitrogen-fixation in bacteria such as *Bradyrhizobium diazoefficiens*, it is highly likely that a new strain could be created with higher rates of nitrogen-fixation and compatibility with a wider range of host plants, allowing for a more viable, economical and natural alternative to fertilisers.

4.1.1 Outline of this study

The overall aim of this study was to undertake the genome annotation of *Bradyrhizobium diazoefficiens*, applying a proteogenomics approach using the Enosi tool [82] and to compare with the findings from another study which used Genosuite [502].

B. diazoefficiens is an important agricultural nitrogen-fixing bacterium. The genome was first sequenced by Kaneko et al. [556] in 2002, with gene predictions

carried out by Glimmer [557] and sequence similarity. It was found that *B. diazoefficiens* had a genome size of 9.1 Mbp and a GC content of 64%.

The Enosi tool was initially built with a focus on eukaryotic genomes, and the latest version incorporates MS-GF+, which applies a combinatorics approach to scoring and assignment of significance to peptide-spectrum matches (PSMs). As part of this study, a proteogenomics pipeline was employed for *B. diazoefficiens,* which has not been previously demonstrated.

A customized proteogenomics pipeline incorporated a pre-processing and optimization step, using the known proteome to evaluate the MS/MS spectral dataset and improve on the PSM identification rate. In addition, a two-pass search approach was applied which demonstrated how it could further improve on sensitivity. Two additional MS/MS spectral datasets (PRIDE Accessions 10112 and 10113) were also included, and with comparisons made between multiple annotations from both methods.

4.2 MATERIALS AND METHODS

4.2.1 Proteomics and genomics datasets

The *Bradyrhizobium diazoefficiens* USDA 110 (previously *B. japonicum*) genome sequence and gene models [556] in GFF and proteome FASTA file formats were downloaded from NCBI BioProject 57599 and further gene models were predicted using Prodigal [558] and RAST [134]. Gene models were modified with an in-house script to be compatible with the proteogenomics pipeline (Appendix File 4.1).

Peptide mass spectrometry datasets, totaling 757,807 MS/MS spectra, were downloaded from PRIDE [559] as MGF files, using accessions 10099-10104 and 10112-10116. Initially generated in proteomics studies [4, 5], samples were run on a 1D gel, subsequently digested with trypsin and run through an LTQ Orbitrap. The MS1 was run on the Orbitrap and MS2 was run on the LTQ.

Contaminant proteins were used, as outlined in Section 3.1.1. In addition, three sources of plant protein contaminants were also used [4, 5]; Soybean (*Glycine max*), Cowpea (*Vigna unguiculata*) and Siratro (*Macroptilium atropurpureum*), obtained from Phytozome v9.0 (http://www.phytozome.net/) and UniProt (http://www.uniprot.org/). Plant protein databases were merged and redundancy was removed using CD-HIT [560] based on 100% homology. These databases were then appended to the reference NCBI protein predictions before being used in the MS/MS database search to identify any contamination.

4.2.2 MS/MS database searching

The MS/MS database search was performed by MS-GF+, as outlined in Section 3.3. In this case study, trypsin was used as the protease, instrument was set to low-res LTQ (Ion Trap), and the precursor mass tolerance used was set to 6.0 ppm, as determined from a preliminary MS/MS spectral dataset assessment, detailed below in Section 4.2.3.

4.2.3 Dataset processing

The NCBI predicted protein sequence FASTA file and GFF file required formatting into a compatible format for proteogenomics analysis, as outlined in Section 3.4.1. The total of all 757,807 MS/MS spectra obtained for this study were first assessed by searching against the known proteome, examining the effects of using MS-Cluster to cluster the MS/MS spectra, PepNovo to quality filter the MS/MS spectra, and an assessment of optimal precursor mass tolerances, as outlined in Section 3.4.6. Since all the MS/MS spectra were derived from an LTQ Orbitrap mass spectrometer and of high-accuracy, this must be reflected in the search parameters. Thus assessment of the optimal precursor mass tolerance involved a range of tolerances: 0.5, 1.0 up to 10.0 in 1.0 ppm increments, and then up to 20.0 ppm in 5.0 ppm increments.

4.2.4 Proteogenomics pipeline

The proteogenomics pipeline was used as outlined in Section 3.5, utilizing the Enosi tool (version 1.0) and a combined false discovery rate (FDR) strategy. In addition, a two-pass search approach was used, similar to the approach demonstrated in [428], and which is outlined in Section 3.5.1. Essentially the first-pass MS/MS database search was performed to identify matching sequences without a decoy database and no FDR filtering, which were then used to define the database for the second-pass MS/MS database search, which included a reversed decoy database of the target sequences and a 1% PSM FDR filtering. The two-pass search approach was used to assess whether any improvements could be observed to the sensitivity of the MS/MS database search of the six-frame translated bacterial genome. After a preliminary assessment of the MS/MS spectra, as outlined above in Section 4.2.3, the choice of clustering the MS/MS spectra, but not quality filtering using PepNovo was decided, resulting in 285,344 clustered MS/MS spectra and a 6.0 ppm precursor mass tolerance.

The total 285,344 MS/MS spectra were split into separate MGF files of 65,000 MS/MS spectra each using an in-house MGF splitting tool, before running each MS/MS spectral file through MS-GF+ on a single Ubuntu machine with 40 processors and 64 GB RAM against the known proteome and a six-frame translation of the genome. The two-pass search approach was also applied, according to Section 3.5.1. However, in this case study all matches were used to define the search space for the 2nd pass, and not those identified with a ≤1.0E-05 spectral E-value, as was demonstrated in later studies, which came to light at a later date after this study had been conducted. The combined FDR strategy and two-pass search approach are illustrated in Figure 3.1.

The choice of parameters for the proteogenomics pipeline, as outlined in Section 3.5, included an initial minimum event probability for novel genes, distal events and proximal events of 90%, a peptide linkage distance of 60 bp representing the optimal

minimum intergenic space within an operon (see Section 4.3), a minimum cluster size of 1 and a minimum of 1 unique peptide per cluster.

Further event probability thresholds were applied to the different annotation event types based on cross-validation between all of the three gene models (NCBI, Prodigal and RAST), with higher stringencies applied to novel gene and distal events. The final event probability thresholds applied were 99.5% for novel genes, 99.1% for distal events and 99.1% for proximal events. As outlined in Section 3.5, the known proteins were identified based on all mapped peptides and also those that contained ≥2 peptides with 1 unique peptide.

4.3 RESULTS AND DISCUSSION

The present study outlined the benefits of the customized proteogenomics pipeline and: 1) brought awareness to how different -omics platforms can be integrated, e.g. in this case genomics and proteomics; 2) demonstrated the differences in sensitivity and specificity between Enosi and Genosuite; 3) demonstrated the use of a two-pass search approach (Section 4.2.4) to database searching; 4) identified 155 novel genomics annotations for *B. diazoefficiens* and; 5) more broadly identified the power of repurposing legacy proteomics data for use in genome annotation.

4.3.1 Evaluation of pre-processing MS/MS spectra

Prior to running the proteogenomics pipeline, an evaluation of MS/MS spectra pre-processing and parameter optimization was conducted (Appendix File 4.2). All 757,807 MS/MS spectra were clustered by a factor of 2.66. It was found that clustering reduced the peptide FDR after an initial 1% PSM FDR filtering from 5.37% to 1.58%, and reduced the protein FDR from 31.40% to 9.69% (Appendix figure 4.1A-C). As can be seen in Appendix Figure 4.1A, the number of total MS/MS spectra lost after quality filtering ranged from 10% at the lowest end to 80% at the most stringent cut-off.

Applying scores between 0.05 – 0.1, as recommended by PepNovo (detailed in the Help File bundled with the tool), resulted in 50% of the MS/MS spectra being lost. These MS/MS spectral losses showed a significant drop in the number of unique peptides discovered after FDR filtering (Appendix Figure 4.1D), while showing only a gradual drop in the number of PSMs (Appendix Figure 4.1E), with the peptide FDR and protein FDR only improving negligibly with quality filtering (Appendix Figure 4.1B-C). Losses in MS/MS spectra, resulting in the reduction in the number of unique peptides reported without any significant reductions in false positive rates, would indicate the dataset was not improving but was losing valuable MS/MS spectra for novel proteogenomics discoveries.

The losses observed were probably attributable to a combination of the methodology employed by PepNovo and the source of the proteomics data, which were derived from a 1D gel followed by trypsin digestion and LC-MS/MS, as explained in Section 3.4.6.

These results indicated a clustered MS/MS spectral dataset with the absence of PepNovo quality filtering was most suitable for further proteogenomics analysis as this resulted in the highest gains in peptide discovery with minimal losses of MS/MS spectra, while keeping the peptide FDR below 2%. In addition, the search tool used in this analysis, MS-GF+, uses a different approach to quality scoring prior to beginning the search, based on log-likelihood ratios [292], and was able to remove poor MS/MS spectra in the study with no apparent impact to the search results.

4.3.2 MS/MS database search parameter optimization

As outlined in Section 3.4.6, high-accuracy MS/MS spectra are well suited to precursor mass tolerance optimization, as tighter tolerances often improve upon the sensitivity of PSM identifications. This holds true for this study, which uses high-accuracy MS/MS

spectra, generated from an LTQ Orbitrap mass spectrometer.

A clustered set of MS/MS spectra was used to assess the precursor mass tolerances over a range, as outlined previously in Section 4.2.3. From this analysis it was determined that 6.0 ppm was the optimal precursor mass tolerance to use (Appendix Figure 4.2). After a ≤1% PSM FDR was applied the maximum number of PSMs obtainable was 53,349 at 6.0 ppm, while the peptide FDR was 1.69% (Appendix File 4.2).

4.3.3 Effects of preliminary analysis on proteogenomics results

To determine the effectiveness of pre-processing the MS/MS spectra by clustering/not clustering and optimizing the precursor mass tolerance, the different parameters and pre-processing steps were compared. It was found that the total run time of the unclustered MS/MS spectra took more than 10x longer than the clustered dataset and there were 337 more annotations with ≥90% event probability. When these novel annotation events were filtered to the same event probability stringency, as outlined in Section 4.2.4, the majority of the annotation events were removed leaving 186 annotation events, 31 more than when clustering and optimizing the precursor mass tolerance. Although these results indicate a gain of 31 annotation events by not clustering, they most likely include greater numbers of false positives, as was noted from the known proteome searches with 5.36% peptide FDR and 31.4% protein FDR when using unclustered MS/MS spectra, compared to 1.58% peptide FDR and 9.69% protein FDR when clustered. The peptide FDR and protein FDR would likely be further inflated by including the six-frame translation, and therefore the confidence applied to these annotation events is relatively low.

Comparing precursor mass tolerances revealed that there were 23 more annotation events with a ≥90% event probability using 6.0 ppm, than by using 20.0

ppm. However, there were only 16 more annotation events found with a 6.0 ppm precursor mass tolerance when more stringent event probability thresholds were applied. Therefore, there is a significant advantage to clustering in combination with the application of a precursor mass tolerance optimization step with the known proteome. Clustering has the advantage of removing many spurious annotation events and reducing search and post-processing times. In addition, optimizing the precursor mass tolerance improved the number of annotations reported and reduced the error window for peptide and protein identifications by removing the potential for incorrect matches. A redundant set of MS/MS spectra totaling 757,807 was then clustered by a factor of 3 and used in the proteogenomics pipeline.

4.3.4 Proteogenomics pipeline

There are several key variables in the proteogenomics pipeline to define the number and type of accepted annotation events, including: the minimum cluster size (peptides per cluster), the minimum number of unique peptides per cluster, the maximum peptide linkage distance and the event probability.

The peptide linkage distance is the most difficult variable to define in bacterial genomes. It is usually calculated based on the size of ≥95% of genes, with the same value defining the distance between a cluster and a neighbouring gene to include in annotation event inference. This would make little sense in prokaryotic genomes due to the compactness of the genome, with many genes close together within operons and with a large proportion of gene overlap. Therefore the maximum peptide linkage distance of 60 bp was chosen because: 1) prokaryotic genomes are often compact with overlapping genes; 2) gene overlaps >60 bp may be considered as misannotations [561, 562]; 3) the maximum distance between genes within an operon is considered to be around 50-60 bp [502]; and 4) high GC bacterial genomes often have longer ORFs [563], increasing the likelihood of false-positive PSMs across the length of the ORF,

which could inadvertently be grouped into the peptide cluster if the peptide linkage distance was sufficiently large enough. However, a peptide linkage distance of 60 bp may also inadvertently classify novel genes as distal events, such as reverse strand annotations, due to a number of truly novel genes on the reverse strand overlapping a known gene. As a result, careful consideration is needed when reviewing novel gene and distal event annotations.

Due to the compact nature of the genome and small peptide linkage distance, one unique peptide per cluster was used to assign annotation events, particular for novel genes and distal events (as opposed to ≥2 unique peptides outlined in [82]), and also due to the lower chance of two or more unique peptides appearing within the smaller peptide cluster. To avoid potential false positives resulting from the inclusion of only one unique peptide, stringent event probabilities were applied, particularly for novel genes. The event probability stringency could be adjusted due to the small genome size based on calculated annotation event FDRs [82], but tools to determine this adjustment were not currently available, therefore to ensure better specificity more stringent event probabilities across all annotation events were applied.

4.3.5 Proteogenomics analysis

The proteogenomics analysis identified 155 novel annotation events among 145 genes from NCBI, with 250 annotation events among 234 genes from the RAST annotation and 88 annotation events among 83 genes from the Prodigal predictions (Table 4.1 and Appendix Files 4.3 and 4.4).

From these results, it can be seen that the proteogenomics evidence agreed more with the Prodigal predictions and demonstrated the least agreement with the RAST annotations, with the NCBI annotations residing in-between these two extremes.

Table 4.1 Summary of bacterial proteogenomics annotations

The results of the proteogenomics pipeline (NCBI reference compared to RAST and Prodigal predictions) are compared with the Genosuite analysis [502].

Proteogenomics tool	**Enosi**			**Genosuite**
MS/MS Search tool(s)	**MSGF+**			**MassWizz/OMSSA/ X!Tandem/InsPecT**
Reference annotation	NCBI	RAST	Prodigal	NCBI
Total NCBI reference genes	8,317	NA	NA	8,317
Total 'known/predicted' protein-coding genes	8,317	8,715	8,498	8,317
Raw MS/MS search 'known/predicted' protein matches ≤1% PSM FDR	3,123	3,134	3,188	NA
Proteogenomics mapping: Total 'known/predicted' proteins ≤1% PSM FDR	3,550	3,557	3,617	2,591
Proteogenomics mapping: Total 'known/predicted' proteins ≤1% PSM FDR (≥2 peptides with 1 unique)	2,194	2,182	2,233	NA
Total identified 'novel' peptides ≤1% PSM FDR	330	538	210	221
Raw MS/MS search 'known' peptides ≤1% PSM FDR	15,103	15,061	15,289	NA
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR	24,975	24,729	25,088	24,194
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR (≥2 peptides with 1 unique)	15,579	15,389	15,684	NA
Frame-shifts	9	47	8	2
Exon boundaries	22	21	19	48
Gene boundaries	19	54	17	36
Reverse strands	45	33	12	21
Novel genes	60	95	32	36
Total annotation events	155	250	88	107
Total genes affected	145	234	83	107
Total novel peptides in affected genes	259	450	144	283

Findings from [502] were based on ≥2 unique peptides or single peptides with ≥5 significant PSMs ≤1% FDR. Due to differences in semantics, the following was assumed:

- Any identified translation initiation start (TIS) sites in [502] would be considered exon boundaries in this study.
- There was no description of gene boundaries from [502], and are possibly described as novel genes.

NA: Not available

Note: Some values for the number of identified proteins have been revised since publication [6].

4.3.6 Novel gene annotations

There were a total of 155 novel annotation events from the NCBI predictions, such as with gene bll2355 and bll2356 (*prp1*) where both were annotated as reverse strand events as the novel peptides were found within the 60 bp peptide linkage distance of the genes. Eight unique peptides, supported by 16 MS/MS spectra (Appendix Figure 4.8),

were found to be overlapping gene bll2355 and then aligning across the gene boundary into an intergenic region, with an event probability of 100%. There were also no peptides that mapped within these NCBI known genes, providing no conflicting evidence for the reverse strand annotation. Where this annotation event was identified as a reverse strand, the previous study [502], RAST and Prodigal predictions concluded differently, identifying a novel gene annotation on the opposite strand. Both RAST and Prodigal did not predict the same gene; instead they predicted genes that overlap bll2355 on the opposite strand covering the entire region where the 8 novel peptides mapped (Figure 4.1). Searching proteins bll2355 and bll2356 against NCBI NR with BLASTP revealed no significant matches to bll2355 except to itself and the close relative *B. japonicum* USDA 6, both characterised as hypothetical proteins. However, bll2356 matched, both to itself and to several dozen other close relatives, all of which are characterised as a metallophosphoesterase protein. Therefore bll2355 should be removed or annotated as putative, with a new gene model annotated on the opposite strand in line with the Prodigal and RAST gene models (Figure 4.1), with no annotation changes to bll2356. As a result from this annotation, it is clear that some reverse strand annotations may be novel genes, making the potential true number of novel genes as high as 125 (Table 4.1). In conclusion, there is a need to determine how well the peptide evidence can fit into heuristically determined gene models. This highlights the need for bacterial gene prediction tools such as Prodigal and Glimmer [557] to accept peptide level evidence as hints, in a similar way to tools like Augustus [102], in order to remove much of the manual annotation and improve throughput.

Figure 4.1 Reverse strand or novel gene annotation

Genome view of potential reverse strand annotation of bll2355 that was probably a novel gene annotation on the opposite strand given the supporting evidence.

Another novel gene discovered in [502] between blr2145 (*CYP114*) and blr2146 was identified in this study as a gene boundary event for blr2146. Compared to the results in [502], only three of the five peptides identified were also identified in this study (Appendix Figure 4.3), with three MS/MS spectra supporting them (Appendix Figure 4.10). The other two peptides have probably fallen outside the FDR threshold. The annotation of a novel gene was however supported by RAST and Prodigal gene predictions (Appendix Figure 4.3). Additionally, known peptides mapped to blr2145 and blr2146, which have been characterized as cytochrome P450 hydroxylase and a dehydrogenase, respectively, by RAST, and which was in agreement with searches against NR with BLASTP. Further proteogenomics evidence, with a smaller peptide linkage distance would likely annotate this event as a novel gene. This highlights the difficulty in assigning a peptide linkage distance for prokaryotes, as the actual intergenic distance in this particular case between the Prodigal and RAST predictions and blr2146 is only a few base pairs, and with only 38 bp between the most 3' novel peptide and blr2146.

A good example of a high confidence novel gene annotation was a novel gene located at positions spanning 539,180 to 539,438 containing 7 unique peptides outside of the known NCBI predictions with an event probability of 100% (Appendix Figure 4.4), and supported by eleven MS/MS spectra (Appendix Figure 4.11). This annotation was also supported by the RAST annotation and Prodigal predictions. Interestingly, RAST annotated the gene as matching a bll3711 protein, which happened to be a paralog gene in *B. diazoefficiens* much further downstream at positions 4,102,260 to 4,102,625. This was confirmed after searching the RAST prediction against NR with BLASTP, which matched bll3711 with an E-value of 1E-81 and query coverage and percentage identity of 100%. This novel gene annotation was also confirmed in [502].

4.3.7 Sequencing error: A discovery from exon boundary/frame-shift annotation

An example of when annotation events can indicate a potential underlying problem with the genome sequence is between two different genes, blr0352 and blr0353. Here the unique peptide "IAANPDDVNALYR", supported by three MS/MS spectra (Appendix Figure 4.9), mapped to a region overlapping both genes. This suggested an exon boundary for blr0353 and a frame-shift for blr0352. Both genes contain mapped peptides, indicating both are being expressed (Figure 4.2A). Both genes were searched against NR using BLASTP, with blr0352 shown to contain a Peptidase C14 domain to which each known peptide has been mapped, while blr0353 contains a tetratricopeptide repeat domain (TPR) that the novel peptide partially overlapped. The Prodigal predictions mirror the NCBI annotation, while the RAST annotation indicates one large gene and protein product with what appeared to be a change in frame in the middle of the gene, and was annotated in agreement with the BLASTP result. The RAST protein prediction which spanned both gene regions was searched against NR using BLASTP and revealed a number of top matches with ≥90% identity to homologous proteins in *Japonicum* USDA 122, *Japonicum* S23321 and SEMIA 5079. Multiple sequence

alignments of these proteins with the NCBI blr0352, prodigal and RAST prediction, using Muscle [564, 565], indicated a greater agreement with the Prodigal predictions and the NCBI annotation, except for a region towards the 3'end where a change in frame occurred (Figure 4.2B). The genomic nucleotide regions for these homologous genes were also aligned using Muscle and identified what appears to be a possible sequencing error with the insertion of a guanine (G) at position 370,898, outlining a string of guanines (Figure 4.2C) shown previously to induce sequencing errors where there are strings of G-C or A-T pairs in the genome [566].

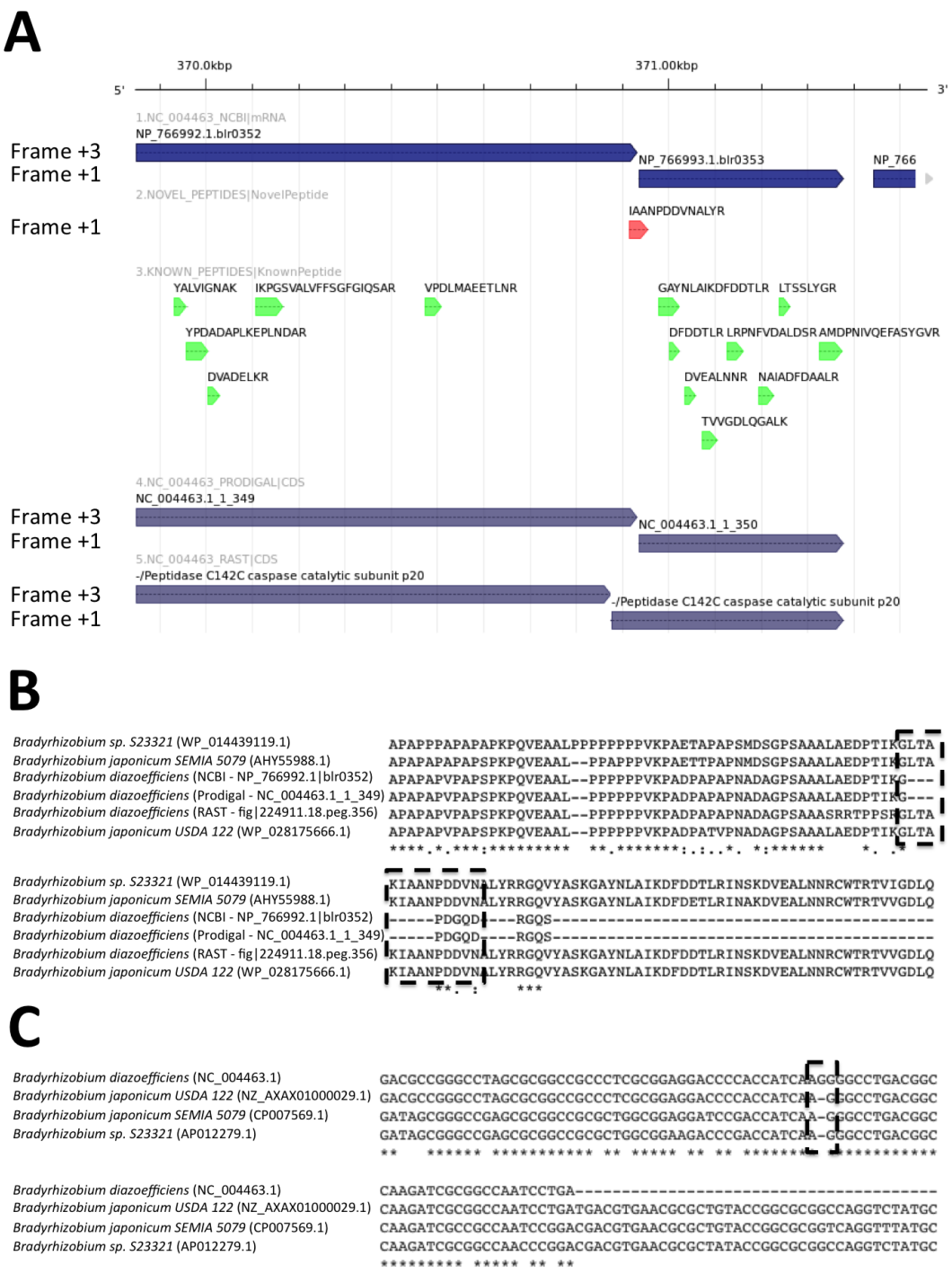


Figure 4.2 Exon boundary and frame-shift annotation or sequencing error

(A) Genome view of potential frame-shift of blr0352 and exon boundary of blr0353. (B) Multiple protein sequence alignment of blr0352, RAST, Prodigal and protein homologs, with a sequence discrepancy resulting from a possible sequencing error (dashed lines) (C) Multiple nucleotide sequence alignment of blr0352 in *diazoefficiens* USDA 110, *japonicum* USDA 122, S23321 and SEMIA 507. A "G" insertion at position 370,898 (dashed lines).

Although the unique peptide did not overlap directly with the possible sequencing error, it highlighted a problem with the sequence in this region. Further sequencing, PCR or alignment of peptide/protein, mRNA, cDNA or EST sequences to

this region are required to confirm the error, as this single nucleotide insertion may actually be one of numerous differences between *B. diazoefficiens* and other closely related species. Correction of this sequencing error would re-assign the unique and novel peptide as a known peptide. The sequencing error was likely a contribution from the presence of the TPR region down-stream to sequencing difficulties of high GC content in bacterial genomes.

4.3.8 Gene boundary annotations

Another annotation was with genes bll0795 and bll0794 (*PhoH*), where both were annotated as gene boundary events. A unique peptide mapped in between bll0795 and bll0794, with an event probability of 99.8% (Appendix Figure 4.6), which was supported by one MS/MS spectrum (Appendix Figure 4.14). Only bll0794 contained mapped peptides that were in the same frame as the novel peptide, and not the same frame as bll0795, so a gene extension to bll0794 and not bll0795 was suggested. Interestingly, although this unique peptide has high confidence, it did not agree with the RAST annotation or Prodigal predictions (Appendix Figure 4.6). Also the study from [502] gave the impression that this unique peptide belongs to gene bll0794, even though it clearly falls outside of the gene boundary and in the intergenic region. Searching bll0794 against NR with BLASTP revealed the protein matched numerous phosphate starvation-inducible proteins (PHOH), within other related species, which was also confirmed from the RAST annotation.

4.3.9 Exon boundary annotations

An annotation from [502] suggesting an alternative translation initiation start (TIS) site was also in agreement with this study, which found an exon boundary event for bll2019 (*NolA*). The unique peptide “IGELAEATGVTVR” was detected overlapping *NolA* at the 3’-end, with an event probability of 99.8% (Appendix Figure 4.5A) and was supported by three MS/MS spectra (Appendix Figure 4.12). The *NolA* gene also

contained known peptides within the same frame as the novel peptide. Searching the NOLA protein against NR with BLASTP revealed that the protein matched numerous other nodulation proteins in other related species. In contrast, the novel peptide was identified from the RAST annotations as a novel gene, as the *NolA* gene was not predicted by RAST, while Prodigal was able to predict the full length of the *NolA* gene (Appendix Figure 4.5A). Another exon boundary event or TIS from [502] in agreement with this study was for gene bll2380. The study from [502] reported four novel peptides overlapping and upstream of bll2380. This study confirmed this as an exon boundary event (Appendix Figure 4.5B) with an event probability of 100% and was supported by fourteen MS/MS spectra (Appendix Figure 4.13), and also supported by the RAST annotation and Prodigal predictions. A number of known peptides also mapped to bll2380 in the same frame as the novel peptides (Appendix Figure 4.5B). Searching bll2380 against NR with BLASTP revealed that the protein matched numerous glycosyltransferase proteins in related species, confirming the same RAST annotation.

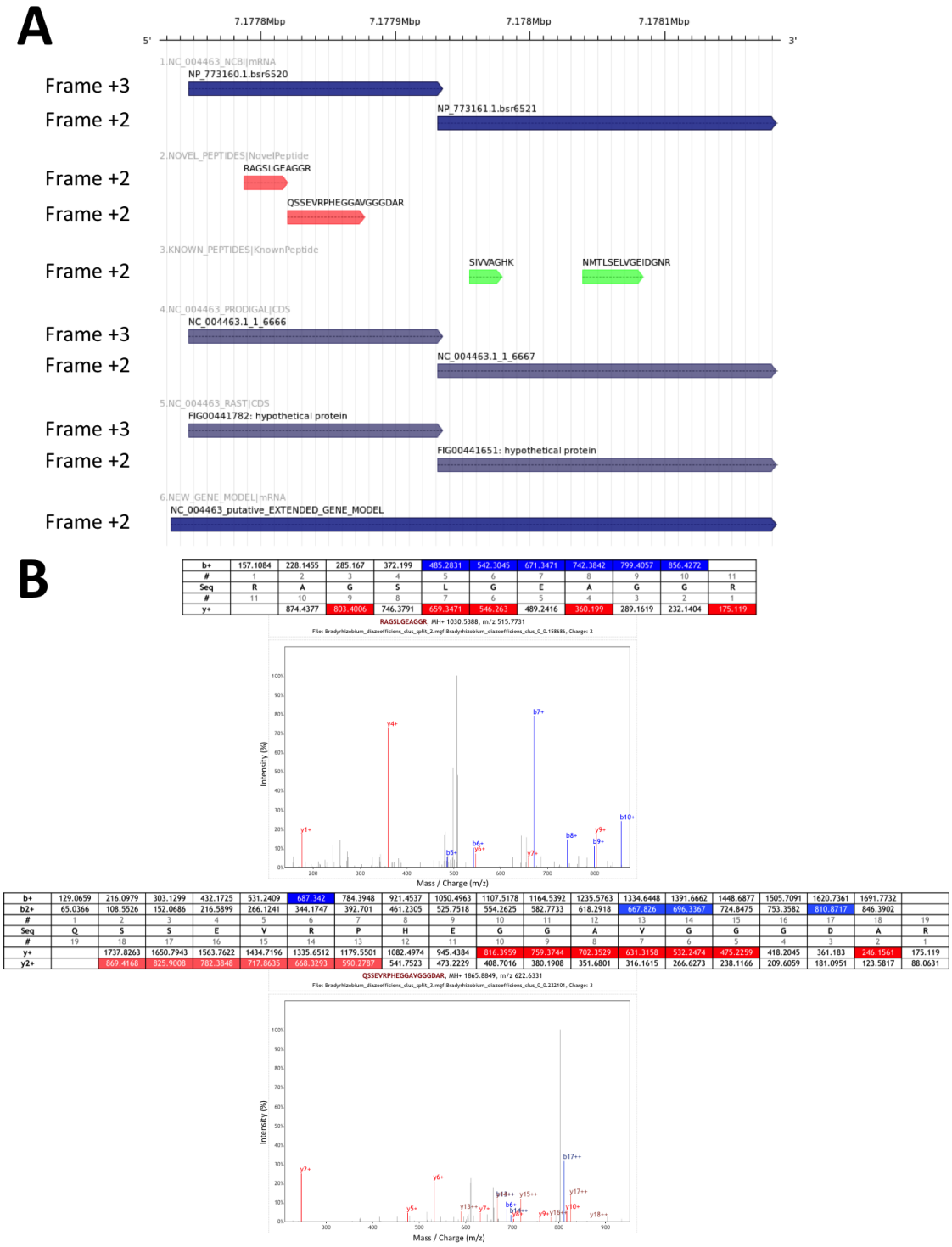
4.3.10 Frame-shift annotation

An example of a high confidence frame-shift annotation was with gene bsr6520. The gene bsr6520 and the gene bsr6521 downstream appeared to have swapped names according to NCBI Gene (http://www.ncbi.nlm.nih.gov/gene/1053434) and the NCBI protein database, which is in contrast to the NCBI annotation and GFF file presented in NCBI BioProject 57599. Further references throughout this chapter use the naming presented in the NCBI annotation and GFF file during the proteogenomics analysis to preclude any confusion.

The bsr6520 gene has 2 exclusively unique and novel peptides “RAGSLGEAGGR” and “QSSEVRPHEGGAVGGGDAR” (Figure 4.3A with two supporting annotated MS/MS spectra in Figure 4.3B) mapping to a different frame, both contributing to an event probability of 99.99%. The gene bsr6521 down-stream was also

in the same frame as the unique peptides. Searching both bsr6520 and bsr6521 against NR with BLASTP revealed bsr6520 contained a match to a hypothetical protein with a Domain of unknown function (DUF4169), while gene bsr6521 matched to arylsulphate sulphotransferase with a Ribbon-helix-helix domain (RHH_4). A previous study [502] identified that both unique peptides belonged to bsr6521 and also to bsr6520, so there may have be some nomenclature confusion, as mentioned earlier. In addition, the results from [502] did not show a report of any frame-shift. This annotation suggests changing the frame of bsr6520, while bsr6521 should be extended to position 7,177,733 where a methionine initiation start probably resides, fusing with bsr6520, as the new frame of bsr6520 contained no stop-codons until the 3' end of bsr6521 (Figure 4.3A).

Further evidence is needed in the form of peptide/protein, mRNA, cDNA or EST sequences bridging the divide between the genes before any definite conclusions can be drawn. Neither the RAST annotations nor the Prodigal predictions agreed with the proteogenomics annotation, instead choosing the original frame from NCBI, which highlights the caution which needs to be applied when using gene prediction tools such as Glimmer, used for the original NCBI annotation of *B. diazoefficiens*. In addition, from the BLASTP results the incorrectly annotated bsr6520 gene can be seen matching to a number of homologues in other *Bradyrhizobium* species, which include species *Japonicum* USDA 124, *Japonicum* USDA 4, S23321, WSM1743, and URHA0013, all containing the DUF4169 domain and all with a percentage identity ≥90%, suggesting this annotation could be applied to these other genomes using an ortho-proteogenomics approach [79].



b+	157.1084	228.1455	285.167	372.199	485.2831	542.3045	671.3471	742.3842	799.4057	856.4272	
#	1	2	3	4	5	6	7	8	9	10	11
Seq	R	A	G	S	L	G	E	A	G	G	R
#	11	10	9	8	7	6	5	4	3	2	1
y+		874.4377	803.4006	746.3791	659.3471	546.263	489.2416	360.199	289.1619	232.1404	175.119

b+	129.0659	216.0979	303.1299	432.1725	531.2409	687.342	784.3948	921.4537	1050.4963	1107.5178	1164.5392	1235.5763	1334.6448	1391.6662	1448.6877	1505.7091	1620.7361	1691.7732	
b2+	65.0366	108.5526	152.0686	216.5899	266.1241	344.1747	392.701	461.2305	525.7518	554.2625	582.7733	618.2918	667.826	696.3367	724.8475	753.3582	810.8717	846.3902	
#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Seq	Q	S	S	E	V	R	P	H	E	G	G	A	V	G	G	G	D	A	R
#	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
y+		1737.8263	1650.7943	1563.7622	1434.7196	1335.6512	1179.5501	1082.4974	945.4384	816.3959	759.3744	702.3529	631.3158	532.2474	475.2259	418.2045	361.183	246.1561	175.119
y2+		869.4168	825.9008	782.3848	717.8635	668.3293	590.2787	541.7523	473.2229	408.7016	380.1908	351.6801	316.1615	266.6273	238.1166	209.6059	181.0951	123.5817	88.0631

Figure 4.3 High confidence frame-shift annotation

(A) Genome view of potential frame-shift annotation of bsr6520. (B) Annotated MS/MS spectra supporting the novel peptides suggesting the frame change of bsr6520.

4.3.11 N-terminal acetylated peptides

An example of a conflicting annotation from [502] which indicated a potentially over-predicted gene was with known NCBI gene blr0594 (*trxA*). N-terminal acetylated peptide "TIIDQGNGAAGPAAADLIK" was identified from the study in [502],

indicating a GTG instead of an ATG at its N-terminal end, which is known to code for the initiator Methionine in high GC genomes and so suggested an alternative TIS site and hence indicated that the gene was over-predicted. By searching amongst all peptides mapping to known genes it was found that the unique peptide contained no acetylation of the N-terminal end (Appendix Figure 4.7 and supporting MS/MS spectrum in Appendix Figure 4.15), even though numerous other peptides with N-terminal acetylation were found amongst the known peptides. According to MS-GF+ the PSM had a spectral E-value of 4.05E-13, an E-value of 1.67E-5 and a Q-value of 0.0, indicating a significant identification. The effect of clustering the MS/MS spectra was considered as a cause of the missed modification, but it was soon ruled out when unclustered MS/MS spectra were searched against the known proteome using MS-GF+ (results not shown). Therefore, the cause could be attributed to subtle differences in sensitivity in the detection of post-translational modifications between the different search tools. In addition, both Prodigal and RAST agreed with the alternative TIS site, as the 5' end of each predicted gene aligned with the coordinates of the unique peptide (Appendix Figure 4.7), suggesting agreement with the results from [502]. To follow this up all N-terminal acetylated peptides within the NCBI known proteins were explored, revealing that each protein they mapped to indicated an agreement with the TIS of each gene.

Of the 3,550 NCBI known proteins containing 24,975 mapped peptides, 36 different proteins contained N-terminal acetylated peptides (Appendix File 4.4). While out of 2,194 high confidence NCBI known proteins, containing 15,579 mapped peptides and at least 2 peptides with 1 unique peptide, there were 20 different proteins containing N-terminal acetylated peptides. Thus confirming, as previously indicated in [502], that N-terminal acetylation may be more widely utilized in bacteria than previously thought. The novel peptides were then screened for N-terminal acetylated peptides, and it was

found that there were no N-terminal acetylated peptides, except for unique peptide "MPMNVPSIAASNMLGMRR" with a low event probability of 93.1%, which was subsequently removed from further analysis. The peptide was found mapped within gene bll3136 (*fdhF*) on the reverse strand, which according to the NCBI protein database is already well characterized as a formate dehydrogenase alpha subunit, adding confidence that this was likely a false positive peptide.

4.3.12 NCBI vs RAST vs Prodigal annotations

As a preliminary step before using NCBI, RAST and Prodigal predictions for proteogenomics analysis, a comparison was conducted using Mummer [567] to determine which protein predictions were in greatest agreement with one another. It was found that using NCBI as the reference dataset Prodigal matched to 7,698 genes (90.6%) and RAST matched to 7,585 genes (87%), implying that more Prodigal genes were in agreement with the NCBI annotations even though there were many more RAST annotations. This indicates that RAST contains a number of false positives or genes predicted in the wrong frame compared to the other annotations. This is further supported from the proteogenomics analysis, which found many more novel annotation events with RAST (250) than with Prodigal (88), while NCBI (156) was in-between.

It was found that the RAST annotation tool incorrectly predicted the CDS phase for a number of genes in the GFF output file. Discrepancies are present between the predicted protein sequence (not shown) and the gene model for the prediction in the GFF file. For example, CDS was in phase 1, when on visual inspection of the predicted protein sequence, it should have been 0. This resulted in many incorrect peptide coordinates for the known genes amongst the RAST predictions. Consequently, the error highlighted the need for highly curated reference datasets before proceeding with any proteogenomics annotation, as by definition the mapping to a reference proteome is only as good as the reference annotation.

4.3.13 Impact of search space

A total of 3,123 of 8,317 proteins annotated by NCBI were identified during MS/MS database searches using MS-GF+, while the total number of proteins mapped by proteogenomics was 3,550. Of these, 2,194 high confidence proteins had ≥2 peptides with 1 unique peptide. Prodigal predictions showed the largest agreement from the proteogenomics analysis, with a total of 3,617 proteins. Of these, 2,233 high confidence proteins had ≥2 peptides with 1 unique peptide (Table 4.1). A total of 4,456 proteins were identified when the same search was conducted against only the known NCBI proteome. Comparisons between proteomics- and proteogenomics-only searches revealed a loss of 1,333 proteins out of 4,456, i.e. a loss of 30%, with similar losses confirmed in other studies [81, 324, 439]. Based on these results, it is likely that many novel annotations derived from the six-frame translation have also been missed due to the loss in sensitivity.

In an attempt to improve on the number of annotations reported, a two-pass search approach similar to the approach reported in [428], was applied for the NCBI annotations (Appendix Files 4.3 and 4.4). Above an event probability of 90%, from the raw MS/MS database search, resulted in the identification of 7 additional novel annotation events and 5 known proteins (total 3,128), however no changes were observed for all known proteins and high confidence proteins once they were mapped by proteogenomics. For the novel annotation events, once filtered to the same stringency applied previously, no changes in the number of novel annotation events could be seen. Adjusting the stringency of these thresholds would include the additional novel annotations at the cost of potentially increasing the false positive rate. A means to determine the FDR at the annotation event level or support annotations through orthogonal evidence is needed before the thresholds could confidently be lowered to include further annotation events.

It seems probable that eukaryotic genomes, where the proportion of protein-coding genes occupy a relatively smaller fraction of the genome and where there are relatively fewer sense-antisense gene overlaps, would likely benefit significantly from the two-pass search approach. In addition, by combining the accuracy and sensitivity of Enosi/MS-GF+ with the two-pass search approach in situations where the genome size is large, such as the human genome (~3 Gbp) or even larger in the hexaploid wheat genome (~17 Gbp), the likely increase in sensitivity of identifying novel annotation events would be worthwhile, and should be incorporated as standard, or at the very least as an option, in all proteogenomics pipelines.

4.4 SUMMARY

The present study has highlighted the advantages of proteogenomics, the power of repurposing legacy proteomics data and has brought awareness to how different –omics platforms can be integrated. Primarily, the study has made a significant contribution to the genomic annotation of *Bradyrhizobium diazoefficiens*, identifying 259 novel peptides contributing to 155 novel annotation events, consisting of 9 frame-shifts, 22 exon boundaries, 19 gene boundaries, 45 reverse strands and 60 novel gene events in a total of 145 genes. Through the identification of these annotation events a possible sequencing error was flagged and further validation is required to resolve some false positive annotation events. Some of the lessons learnt from this study include: 1) the problems identified when using a fixed peptide linkage distance; 2) the high proportion of false positive annotation events with overlapping genes reported by Enosi; 3) the relative ineffectiveness of a two-pass search approach in bacterial genomes with a high proportion of overlapping genes; and 4) the loss in sensitivity when applying a combined FDR strategy.

4.5 CONCLUSIONS

While Enosi benefited greatly from improved sensitivity, improvements were still required to deal with overlapping bacterial genes. In contrast, Genosuite was capable of distinguishing such features and provided a higher specificity, but a lower sensitivity than Enosi. Clustering and the selection of appropriate precursor mass tolerances improved efficiency in proteogenomics searches, while the problems with the reduction in sensitivity due to six-frame searches resulted in a 30% loss of known proteins when using a combined FDR strategy. This was partially overcome using a two-pass search approach, however the search space still proved to be an obstacle for bacterial proteogenomics. This was probably due to the high level of overlapping genes and the relative proportion of coding to non-coding genes compared to larger genomes such as human and wheat, where the non-coding portion remained relatively small by comparison, and possibly as a result of including very low significant matches from the first pass in the second pass. Overall, the methods employed in this study provided a means to better understand the field of proteogenomics, thereby identify current gaps in understanding and facilitate additional future improvements in the field.

4.6 ACKNOWLEDGEMENTS

Chapter 4, was published in the journal Proteomics, in a proteogenomics special edition, as “High-throughput parallel proteogenomics: A bacterial case study” [6]. The dissertation author was the primary author of this paper. The dissertation author designed the proteogenomics workflow, ran the analysis and wrote the paper. The dissertation author would like to thank the Centre for Comparative Genomics for their assistance and use of their compute resources.

5 GRAPE PROTEOGENOMICS

5.1 INTRODUCTION

The grapevine industry has successful global market access and large economic support worldwide. The genus *Vitis* is important to the wine industry and as a perennial fruit, part of the staple diet in the Mediterranean where there is reduced prevalence of heart disease [568]. The putative causative agents reducing the prevalence of heart disease may well be derived from grapes, with a number of key candidates being resveratrol, quercitin and ellagic acid [569], with resveratrol attracting extensive media attention in recent years as a potential life-extending drug [570], and which further adds to the *Vitis* market value and opens up the potential for many unexplored medical benefits. The broad spectrum of commercial applications of grapevines challenges the industry to improve yields, quality, resistance to diseases and abiotic stress conditions across the globe. One particularly important *Vitis* species is *Vitis vinifera*, which has recently been subjected to sequencing e.g. sequencing of the heterozygous variety Pinot Noir [571], and a 93% homozygous Pinot Noir, from genotype PN40024 [572]. The sequencing and assembly of the latter variety has since been improved from 8X coverage to 12X coverage resulting in a 487.1 Mbp assembled genome. Genomic annotation of the 8X and 12X was also undertaken, with the 8X gene prediction being performed using GAZE [110], published along with its sequencing [572], while the 12X sequence and assembly has since resulted in three different iterative improvements in annotation. The first named 12Xv0 was performed using GAZE. The second named 12Xv1 resulted from the combination of 12Xv0 and gene predictions by the tool JIGSAW [108], undertaken at CRIBI in Padova, Italy [573]. The third improvement named 12Xv2 (since updated to 12Xv2.1) was undertaken recently, using assembled transcripts from RNA-seq, *ab initio* predictions, proteins, and ESTs [7].

5.1.1 Outline of this study

The aim of this study was to apply proteogenomics to further improve on genomic annotation of the grape genome and to compare the complexity of performing proteogenomics annotation in larger plant genomes, in relation to smaller bacterial genomes, as was demonstrated in Chapter 4. In addition, the benefits and shortcomings of current proteogenomics strategies were outlined. The latest grape Pinot Noir 12X genome assembly and 12Xv2.1 genome annotation were obtained from CRIBI, while the proteomics data were in the form of 177,174 MS/MS spectra derived from Cabernet Sauvignon grape berry skins, used in an earlier proteogenomics study by the dissertation author in Chapman et al [8]. In that study an earlier version of the proteogenomics pipeline was applied, with 29 annotation events found including; 1 frame-shift, 3 translated UTRs, 1 exon boundary, 6 novel exons, 9 gene boundaries, 3 reverse strands and 6 novel gene events. The present study expands on that work by using an improved proteogenomics pipeline and an additional 2,701,718 MS/MS spectra derived from Cabernet Sauvignon shoot tips, used in a large proteomics study on the effects of water deficiency [9]. In addition, RNA-seq data derived from *Vitis vinifera* Corvina cultivar [10] and a large RNA-seq study looking at multiple cultivars (unpublished), was used for the identification of splice regions.

5.2 MATERIALS AND METHODS

5.2.1 Proteomics and genomics datasets

The latest assembled grape Pinot Noir 12X genome [572] with the genotype identifier PN40024, and the 12Xv2.1 genome annotation and protein predictions [7] were downloaded for use from the CRIBI web site (http://genomes.cribi.unipd.it/DATA/) (Appendix File 5.1).

The MS/MS spectra were derived from finely ground shoot tips of Cabernet Sauvignon, across 3 biological replicates on 4 different days during water deficit [9].

The samples were digested with trypsin, aided by Lys-C digestion, and were run on a LTQ XL mass spectrometer (Thermo), with fractionation performed by HPLC and a series of gas-phase fractionation (GPF) steps to further aid in separation. A total of 2,701,718 MS/MS spectra were downloaded from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [574], using the identifier PXD000123. An additional 177,174 MS/MS spectra from the previous proteogenomics study by the dissertation author in Chapman et al [8] were also used, derived from Cabernet Sauvignon berry skins which were finely ground, extracted, digested and fractionated in the same way as the shoot tips, with the samples running on an LTQ Velos Pro (Thermo) mass spectrometer.

As outlined in Section 3.1.1, a source of contaminants was appended to protein sequence predictions in the 12Xv2.1 annotation before being used in the MS/MS database search to identify any contamination.

5.2.2 RNA-seq datasets

Illumina RNA-seq datasets were obtained from two sources. One source was from a recent study [10], where the transcriptome of *V. vinifera* cultivar Corvina was sequenced, looking at three different developmental time points, i.e. post-fruit set (PFS), mid-ripening (MR), and mid-withering (MW), at approximately 2 months post-harvest and from different tissues, organs and development time-points. The RNA-seq reads were downloaded from the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) through DNA Nexus (http://sra.dnanexus.com/), using identifier SRA055265. The second source of RNA-seq data, which is currently unpublished, consisted of grape skins from multiple different cultivars Chardonnay, Cabernet Sauvignon, Merlot, Pinot Noir, Semillon, Cabernet Franc and Sauvignon Blanc, that had been subjected to different water deficits, time-points and sugar levels. This data was obtained from collaborator Ryan Ghan from the University of Nevada

and who was a co-author to the previous grape proteogenomics study in Chapman et al [8].

5.2.3 MS/MS database searching

The MS/MS database search was performed by MS-GF+, as outlined in Section 3.3. In this case study trypsin was used as the protease, the instrument was set to low-res LTQ (Ion Trap), and the precursor mass tolerance was 2.0 Da and 3.0 Da for two different MS/MS database searches, as determined from a preliminary MS/MS spectral dataset assessment, detailed below in Section 5.2.4.

5.2.4 Dataset processing

The 12Xv2.1 protein sequence FASTA file and GFF file required formatting into a compatible format for proteogenomics analysis, as outlined in Section 3.4.1. Many of the CDS phase in the 12Xv2.1 annotation were incorrect. The errors were retained for the proteogenomics analysis, to highlight corrections to the annotation via proteogenomics. However, GenomeTools [542] was used to correct the CDS phase errors to be used later as hints for gene prediction, and for visualization, as outlined in Section 3.1.1 (Appendix File 5.1).

The total 2,878,892 MS/MS spectra obtained for this study were first assessed by searching against the known proteome, examining the effects of using MS-Cluster to cluster the MS/MS spectra, PepNovo to quality filter the MS/MS spectra, and with an assessment of optimal precursor mass tolerances, as outlined in Section 3.4.6. Since all the MS/MS spectra were of lower accuracy, derived from an LTQ mass spectrometer, this factor had to be reflected in the search parameters. Therefore, assessment of the optimal precursor mass tolerance involved a range of tolerances, 0.5 up to 5.0 Da, in 0.5 Da increments (Appendix File 5.2).

All RNA-seq data were pre-processed for quality and aligned to the 12X grape

genome as detailed in Section 3.4.3. The resulting alignment BAM files were then merged and used to generate a splice graph FASTA database. A six-frame translation of the genome was also generated. The methods used for both splice graph and six-frame translation generation are outlined in Section 3.4.

5.2.5 Proteogenomics pipeline

The proteogenomics pipeline was used as outlined in Section 3.5, utilizing the Enosi tool (version 1.0). This study began before a two-pass search approach with two-stage false discovery rate (FDR) strategy was considered, and so a combined FDR strategy was applied to all MS/MS search results using a 1% peptide-spectrum match (PSM) FDR. However, due to the low-accuracy of the MS/MS spectral dataset, two runs were performed using 2.0 Da and 3.0 Da precursor mass tolerances to improve on the number of PSMs identified. The choice of these two different precursor mass tolerances were based on the preliminary assessment of the MS/MS spectra, detailed previously in Section 5.2.4 and later discussed in Section 5.3 (Appendix File 5.2).

A clustered MS/MS spectral dataset, quality filtered to a PepNovo score of 0.01, gave 1,594,076 MS/MS spectra. The MS/MS spectra were then split into 20,000 MS/MS spectra each using an in-house MGF splitting tool, before running each MS/MS spectral file through MS-GF+ on a cluster against the known proteome, two six-frame translation files (minimum 500 MB each), and a splice graph FASTA file for each of the 2.0 Da and 3.0 Da proteogenomics runs. Each set of results were then merged across 4 tab-separated value (TSV) files, and further processed through the proteogenomics pipeline. It was necessary to merge the results across four different TSV files instead of a single TSV file, due to limitations in processing large result files from the combined FDR strategy, as was outlined in Section 3.5.

The choice of parameters for the proteogenomics pipeline, as outlined in Section 3.5, included an initial minimum event probability for novel genes, distal events and proximal events of 90%, a peptide linkage distance of 18,000 bp representing >95% of gene sizes in the current annotation, a minimum cluster size of 1 (total peptides per cluster) and a minimum of 1 unique peptide per cluster. Following this, for all annotation events which did not have ≥2 unique peptides and/or ≥99.9% event probability, screening was performed based on the number of assigned PSMs, sequence homology to sequences in NCBI NR, NCBI RefSeq protein and NCBI SwissProt, with an emphasis placed on matches to the same chromosome and genomic region, supported by EST, mRNA and/or protein evidence, and with all proximal events requiring at least 100% query coverage and identity. Mapped known peptides matching the protein being annotated, and their frame in relation to the novel peptides, were also considered, particularly for proximal events (Appendix File 5.3).

The results from the proteogenomics analysis at 2.0 Da and 3.0 Da were aggregated for the known proteome by providing all FDR filtered results together and starting a normal run of the proteogenomics pipeline. For the novel identifications, the aggregation of results was undertaken directly through the Enosi tools aggregation function, taking the novel peptide locations, known peptides and their locations from the 2.0 Da and 3.0 Da results.

5.2.6 Improving gene predictions

Once the novel annotations were filtered and reviewed, the gene prediction tool Augustus [102], was used to improve the overall gene models of the 12Xv2.1 annotation. Augustus was first trained using the automated Augustus web server [575], by providing ESTs, cDNA (including FL-cDNA) and mRNA downloaded from the NCBI nucleotide database and a repeat masked *V. vinifera* genome, using RepeatMasker [576] and Tandem Repeat Finder [577]. Once the Augustus *V. vinifera*

gene model parameters were generated, hints in the form of extrinsic evidence to improve the gene models were generated. These included ESTs, cDNA, and mRNA, generated using the BLAST-like Alignment Tool (BLAT) [122], while intron hints were generated using the RNA-seq BAM file (used previously for generation of the splice graph) with the Augustus script bam2hints, and repeat masked hints were generated from RepeatMasker during previous training. In addition, the current 12Xv2.1 annotation was used as hints (with the CDS phase corrected using GenomeTools), along with the hints from novel peptides reviewed previously from the accepted novel annotations. The Augustus gene prediction tool was then run, using parameters as outlined in Section 3.5.3.

5.3 RESULTS AND DISCUSSION

The present study outlined improvements to the 12Xv2.1 annotation of *V. Vinifera*, demonstrating the benefits of proteogenomics by identifying 341 (103 exclusively) novel annotation events, and in particular showed how the use of legacy data from other studies can value-add and improve on the gene models. The study was a good example of the importance of sharing proteomics and RNA-seq data, to be utilized beyond the initial scope of the generation of such data, and provided awareness of how different -omics platforms, such as genomics, proteomics and transcriptomics can be integrated. The study also outlined different strategies towards proteogenomics and the benefits of considering and evaluating each parameter and annotation, instead of blindly applying thresholds with no reflection of their suitability for the study in question. This study further reflects on some of the strategies taken, and considers further improvements.

5.3.1 Evaluation of pre-processing MS/MS spectra

Prior to running the proteogenomics pipeline, the MS/MS spectra were evaluated for the optimal pre-processing strategy and precursor mass tolerance (Appendix File 5.2). All

2,878,892 MS/MS spectra were clustered by a factor of 1.50. It was found that clustering reduced the peptide FDR after an initial 1% PSM FDR filtering from 11.70% peptide FDR to 4.70% peptide FDR and reduced the protein FDR from 39.30% to 17.90% (Appendix Figure 5.1A-C). As can be seen in Appendix Figure 5.1A, the number of total MS/MS spectra lost after quality filtering ranged from 3.3% at the lowest end to 64% at the most stringent cut-off. Applying scores between 0.05 – 0.1, as recommended by PepNovo (detailed in the Help File bundled with the tool), resulted in around 33% to 45% of the MS/MS spectra being lost. However, at a score of 0.01 only 17% of MS/MS spectra were lost, while maintaining a peptide FDR of ~3%, as well as retaining many PSMs and unique peptides which were lost with higher scores than 0.01 (Appendix Figure 5.1D-E). Any losses in MS/MS spectra, resulting in a reduction in the number of PSMs and the number of unique peptides reported, without any significant reductions in false positive rates, would indicate that the dataset was not improving but was losing valuable MS/MS spectra for novel proteogenomics discoveries.

Taken together, the results for this particular MS/MS spectra indicate a clustered MS/MS spectral dataset, with a PepNovo quality filtering score cut-off of 0.01 was most suitable for further proteogenomics analysis, as this resulted in the best balance between keeping the largest number of PSMs and filtering out poor quality MS/MS spectra, while keeping the peptide FDR to around 3%.

5.3.2 MS/MS database search parameter optimization

Only low-accuracy MS/MS spectra were available for this study, which limited its potential for identifying the highest number of PSMs due to a higher false positive rate as a consequence of the larger error window. A clustered set of MS/MS spectra, which was then quality filtered to a PepNovo score of 0.01, was used to assess the precursor mass tolerances (Appendix File 5.2). In this study, in contrast to Chapter 4 (where a well-defined optimal precursor mass tolerance could be chosen), due to the larger mass

window, the sensitivity of PSM identifications increased the further the mass error window was increased, and the peptide FDR after an initial 1% PSM FDR can be seen to plateau off after 2.0 to 3.0 Da, remaining under 4.5% (Appendix Figure 5.2). The optimal precursor mass tolerance here was not necessarily the highest, which could theoretically keep going above the 5.0 Da range looked at in the assessment. However, larger precursor mass tolerances would negatively impact the sensitivity of the database search as more potential PSMs would be considered in the search, further negatively impacting search time and false positive rates. As a result, two different precursor mass tolerances were chosen, which gave the best balance between the sensitivity of PSM identifications and peptide FDR: 2.0 Da and 3.0 Da, which gave 57,968 PSMs at 3.3% peptide FDR and 81,278 PSMs at 3.9% peptide FDR, respectively.

Using both precursor mass tolerances for proteogenomics analysis by aggregating the results together improved the overall identification rate. While the 2.0 Da precursor mass tolerance identified a slightly different set of novel peptides as a result of the smaller search space, identifications which were missed due to the restricted mass window were then identified using the 3.0 Da precursor mass window, which identified some of the same novel peptides but also included others not previously identified (results not shown). In reality, using a much larger range of precursor mass tolerances could be selected e.g. 1.0 Da to 4.0 Da, which would identify more PSMs than 2.0 Da and 3.0 Da alone. However, to keep the post-analysis simple (by keeping the number of time-consuming MS/MS database searches to a minimum), and to avoid inflating the FDR further than absolutely necessary, these two were chosen which most agreed with the assessment (Appendix Figure 5.2B).

5.3.3 Proteogenomics pipeline

A proteogenomics pipeline was customised using Enosi with MS-GF+, as outlined in Section 3.5 and illustrated in Figure 3.1, utilizing only the combined FDR strategy with

no two-pass search approach, instead improving upon the sensitivity and identification rate by performing two proteogenomics runs with two different precursor mass tolerances and aggregating the results (Appendix Files 5.3 and 5.4).

A number of key variables required consideration in this study, just as they were considered in Chapter 4. However, in contrast, the peptide linkage distance was easier to determine in this study, as the genome was eukaryotic containing much larger intergenic distances. The peptide linkage distance was chosen based on the size of ≥95% of genes in the current genome annotation, which was found to be 18,000 bp.

For the number of unique peptides per cluster a minimum of 1 unique peptide per cluster was applied, with a minimum peptide cluster size of 1. Following this step annotation events were accepted based on a number of criteria outlined in Section 5.2.5. Such an approach, while laborious in screening large numbers of PSMs, was able to identify more novel annotations than simply applying a stringent event probability and allowed for annotation events to be validated against proteins previously not considered in the original annotation. Based on this approach, the final minimum event probabilities were 99.80% for novel genes and 98.374% for distal events and proximal events, with a number of other single peptide annotation events removed during screening. For example, some single unique and novel peptides, particularly those with lower event probabilities, did not match anything significant within the grape family in NR, RefSeq protein or SwissProt databases while others only matched bacteria, plant mitochondria or chloroplasts, and some proximal events did not have 100% coverage and identity with some matches. For a number of proximal and translated UTR events in particular, due to their close proximity to the known annotation, there was an overall correlation observed between the probability of the annotation event and the number of known peptides also mapping to the protein being annotated. A similar observation was made with the number of spectral counts and other supporting evidence as the event

probabilities increased (Appendix File 5.3). Many of the higher event probability annotation events were later found to incorporate into new gene predictions, however, this was not consistently the case, as further discussed in Section 5.3.4.

Many of the novel peptides found to match chloroplasts and mitochondria were often accompanied by higher numbers of PSMs compared to the majority of other novel peptides. This would be expected due to the larger numbers of such proteins relative to the rest of the plant cell, as chloroplasts and mitochondria are numerous, particularly in rapidly growing grape shoot tips used in this study. To avoid such contamination, a cell component isolation step during sampling could have been applied, but as this dataset was legacy data and not originally purposed for proteogenomics, this was not a path that could be pursued. The source of the proteomics data can therefore be another source of false positives. In this case, due to the source of the genome being the cultivar Pinot Noir, with the proteomic data derived from Cabernet Sauvignon and RNA-seq data covering various different cultivars, variations between the peptide sequences and target genomic and RNA-seq sequences could occur. This could lead to potential misinterpretations of the variation as a post-translation modification (PTM) or identification of novel peptides in multiple locations where there are none, which would negatively impact the final event probability. A means to limit the false positive rate could be to limit the MS/MS database search to essential PTMs, limit the search space or alternatively by adding known variant peptides to the target database such as that demonstrated with splice graphs in [474]. However, the possibility of using variant sequences in proteogenomics was not known until late in the study and was consequently not explored.

Future studies utilising data generated solely for proteogenomics could generate large amounts of data specifically for one variety, and thus control for these types of false positives, with the inclusion of RNA-seq data from other grape varieties,

conditions and time points to identify variants using the methods demonstrated in [474]. Additionally, a source of false positives could be the lack of sequence coverage of the genome, with large regions remaining unsequenced. This possibility may be the case with the 12X genome assembly in this study, which is fragmented and consists of 19 chromosomes, a number of which also have random fragment counterparts and also a large unassigned chromosome (ChrUn), which can be seen in a relatively recent comparative analysis between the 8X and 12X assemblies [578]. A fragmented genome can also result in a number of misidentified PSMs, as outlined in Section 2.4.1, which highlighted a number of work-around solutions including: the use of *de novo* sequencing tools, searching the interpreted MS/MS spectra against homologous sequences, using mutation tolerant search tools, or modifications to approaches such as template proteogenomics which align stretches of interpreted MS/MS spectra against relatively short homologous sequences for construction of whole proteins in the absence of a target genome sequence. However, such approaches were not explored in this study, as they would require more algorithmic development for use in proteogenomics for whole genomic re-annotation, and as such, have yet to be implemented in tools such as Enosi.

In addition to using the sequence homology approach for the screening of 1 unique and novel peptide annotation events, the peptide length was also considered; with short peptides (<10aa) only considered if the match was to the identical chromosome and region identified from the proteogenomics analysis. As in Chapter 4, the event FDR could not be calculated to determine which event probability provided an acceptable FDR at the annotation event level. Therefore, screening each novel annotation event in the manner described above provided another way in which to discriminate true from false positives. How effective this was at the annotation event level could not, however, be determined.

5.3.4 Proteogenomics analysis

The proteogenomics analysis was performed based on the acceptance of novel annotations with preference given to high event probabilities and ≥2 unique peptides per cluster accepted. Other annotation events like those with lower event probabilities and only 1 unique peptide per cluster, were screened by considering the event probability, spectral counts, mapping of unique and shared peptides to proteins being annotated with their frame in relation to the novel peptides considered and sequence homology to known sequences in the same chromosome and region, containing orthogonal supporting evidence. To account for an inflated number of gene boundary and reverse strand events for each peptide cluster, due to the use of a fixed peptide linkage distance, as was explained in Section 3.5, an exclusive number of annotation events for gene boundary and reverse strand events, and their associated genes and proteins was also determined to indicate the numbers without the effect of a fixed peptide linkage distance, as shown in parenthesis in Table 5.1.

This series of screening led to final event probabilities of 99.80% for novel genes and 98.374% for distal events and proximal events, which followed with the identification of a total of 133 novel peptides and 341 novel annotation events (103 exclusively) among 216 genes (67 exclusively) from the 12Xv2.1 annotation (Table 5.1, Appendix Files 5.3 and 5.4).

This study showed a large improvement over the 29 annotation events identified during the preliminary study [8] mentioned in Section 5.1.1. The novel annotations along with the 12Xv2.1 reference annotation were then used as hints for Augustus gene prediction. A total of 84,948 genes and 93,754 proteins (≥66 aa in length) were predicted (Appendix File 5.5), and of these, 57 predicted proteins had 110 novel peptides incorporated (Table 5.1), of which 94 novel peptides were unique and identified in 54 of the 57 predicted proteins (Appendix File 5.6). The number of protein-

coding genes and proteins predicted by Augustus was far higher than the original reference 12Xv2.1 predictions (Table 5.1). In addition to new predictions previously not identified, these high numbers could also be attributed to two factors, as was outlined in Section 3.5.3.

The novel peptides incorporated into the predictions ranged in event probabilities from 98.374% to 100%, which was within the same range as all resulting filtered novel annotation events. However, the majority of the distribution of novel annotation events that contained peptides, which were incorporated into the Augustus gene predictions, belonged to the higher event probabilities. There were also some novel peptides, from high event probability annotation events, which could not be incorporated into predictions. This was possibly due to the interpretation of real MS/MS spectra derived from contaminants, such as from chloroplasts and mitochondria or misidentified variant peptides mistaken as containing PTMs. The event probabilities could not discriminate against these types of false positives and could only provide a probability of the whole annotation event being correct based on the product and quality of all the MS/MS spectra in the annotation event.

The exclusion of many novel peptides from Augustus gene predictions were also observed in the Arabidopsis proteogenomics analysis from [81], however this was not discussed in that study. The dissertation author received confirmation that this was the case with the Arabidopsis proteogenomics study (S. Payne, personal communication, September 29, 2012), and so this should warrant being addressed in the Enosi tool in later versions. Because if some accepted novel peptides could not be included into the predictions, it is highly probable that some novel peptides that were included should not have been, leading to false predictions. This could perhaps be addressed with the consideration of other evidence, additional parameters and threshold stringencies before defining the final peptide clusters and inferring annotation events, to avoid including

what may be spurious novel peptides, instead of mainly relying on parsimony of unique peptides within annotation events and their event probabilities.

The number of novel peptides incorporated into the predictions was quite high at 110 novel peptides (83%). Of the 110 incorporated novel peptides, 15 were exclusively derived from the splice graph, while 7 were identified in both the six-frame translation and splice graph, with the remaining 88 novel peptides identified exclusively in the six-frame translation.

A BLASTP search was performed, by searching all 93,754 Augustus-predicted proteins (Table 5.1) against the 12Xv2.1 proteins, taking the top match with E-value ≤1E-10. Any sequences that did not match were considered novel predictions, sequences that had a query coverage ≥95% with at least 1 mismatch were considered to be the same prediction as the reference protein, and the remaining matches were considered to be modified predictions, either due to Augustus predicting different gene models or modified as a direct result from the supporting evidence. From this analysis there were 42,257 non-paralogous novel protein predictions, 32,837 modified predictions and 18,660 predictions considered to be essentially the same as the reference.

Searching all 57 protein predictions that had the novel peptide evidence incorporated, against the 12Xv2.1 proteins, taking the top match with E-value ≤1E-10, identified 49 protein predictions likely to be modified predictions, leaving 3 protein predictions, that found no match and were considered as non-paralogous novel protein predictions (Table 5.1).

Based on the annotation events incorporated into the Augustus gene predictions, the minimum event probabilities which led to a new Augustus gene prediction were: gene boundary, translated UTR, reverse strand and exon boundary event 98.374%,

frame-shift event 99.80%, and novel exon event 99.193%.

Table 5.1 Summary of grape proteogenomics annotations

The results of the proteogenomics analysis of grape 12Xv2.1 annotation.

Total 12Xv2.1 genes	31,845
Total 'known' protein-coding genes	31,654*
Total 'known' proteins	55,373*
Raw MS/MS search 'known' protein matches ≤1% PSM FDR	2,773
Proteogenomics mapping: Total 'known' proteins ≤1% PSM FDR	7,536
Proteogenomics mapping: Total 'known' proteins ≤1% PSM FDR (≥2 peptides with 1 unique)	1,117
Total identified 'novel' peptides ≤1% PSM FDR	325
Raw MS/MS search 'known' peptides ≤1% PSM FDR	7,886
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR	11,779
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR (≥2 peptides with 1 unique)	5,048
Frame-shifts	5
Translated UTRs	37
Exon boundaries	16
Novel splices	1
Novel exons	9
Gene boundaries	160 (24)
Reverse strands	112 (10)
Novel genes	1
Total annotation events	341 (103)
Total genes affected	216 (67)
Total proteins affected	326 (101)
Total novel peptides in affected genes/proteins	133
Total Augustus protein-coding gene predictions	84,948
Total Augustus protein predictions	93,754
Total Augustus gene predictions with incorporated novel peptides	55
Total Augustus protein predictions with incorporated novel peptides	57
Total novel peptides incorporated into Augustus protein predictions	110
Improved protein predictions with incorporated novel peptides	49
Novel non-paralogous protein predictions with incorporated novel peptides	3

* The original consisted of 31,845 protein-coding genes coding for 55,564 proteins. A total of 191 proteins which were <20 aa in length were removed from the analysis.

Note: Numbers in parenthesis represent the exclusive numbers. The inflationary effect of a large peptide linkage distance on gene boundaries and reverse strands was removed by assigning a peptide cluster as either a proximal or distal event, not both, with preference placed on proximal events.

5.3.5 Novel gene annotations

A single novel gene annotation event was discovered (Table 5.1), located on chromosome 14, and spanning positions 14,715,245 to 14,715,284 with an event probability of 99.80%. This novel annotation event consisted of a single unique peptide, with 2 PSMs assigned (Figure 5.1, and with two supporting annotated MS/MS spectra

in Appendix Figure 5.3). Searching the novel peptide against the grape family in NR revealed a significant match to a predicted glutathione S-transferase protein (XP_002266106.1 with E-value = 3E-07) of 263 aa in size, which was also located on chromosome 14, within the same genomic region, with mRNAs, ESTs, and proteins supporting the prediction.

However, the novel peptide could not be incorporated into the Augustus gene prediction, although two genes in close proximity were predicted on different strands (Figure 5.1). While a BLASTP search of gene g54248 against the grape family in NR found no significant matches, gene g54247 in the same frame as the novel peptide, further upstream, found a significant match to a hypothetical protein (CAN74624.1 with E-value = 0.0). Although as with the BLASTP search, exon and exon_part hints (Figure 5.1) suggest that there should be an identified gene in this region. Further evidence is likely needed to add support to the novel gene event before the weighting of the evidence is sufficient to predict a new gene in this region, in line with the supporting evidence found in NR.

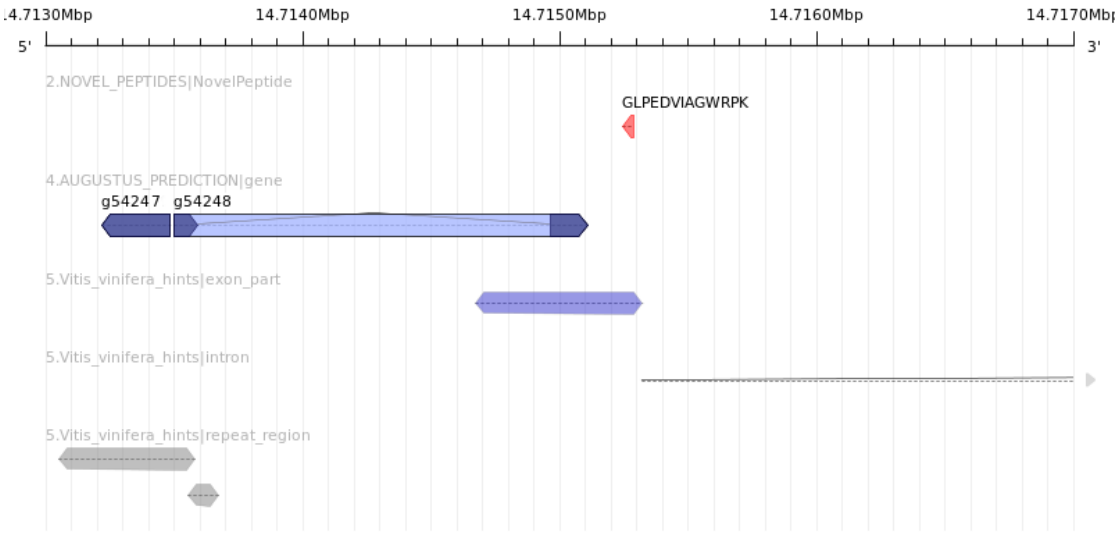


Figure 5.1 Novel gene annotation

The novel gene event inferred from a novel and unique peptide flanks Augustus gene predictions, but is not incorporated into any new predictions. However, exon_part (EST and cDNA evidence) hints used for prediction overlap the novel peptide region before an intron region inferred from the RNA-seq evidence. In the last track a repeat region is highlighted.

Based on the peptide linkage distance no novel genes were found with ≥2 unique peptides, highlighting the importance of screening all annotations to identify valid single unique peptide annotation events. With better depth and breadth of sampling across many more grape tissues, the identification of annotation events with ≥2 unique peptides would likely improve. Regardless, careful screening of all annotations, at least with single unique peptides, was still a viable approach to identify potentially missed annotations, even if sufficient depth and breadth of sampling was available.

One of the reasons why only one novel gene event was identified was because the proteogenomics approach used categorizes peptide clusters as novel genes when they reside outside the peptide linkage distance. This is also similarly true for gene boundary events and reverse strand events, which rely on the same peptide linkage distance for the assignment of annotation events across the genome, and can result in many overlapping annotation events for a single peptide cluster even if they are unlikely. A true annotation event assignment is likely to be interpreted correctly for peptide clusters in close proximity or directly overlapping the genes, as they cannot be refuted, unlike annotation events inferred on genes with peptide clusters possibly many thousands of base-pairs away. This is a consequence of applying a general rule (gene sizes >95% in size), across all genomic features and as a result there are likely to be many more novel genes misidentified as gene boundary events and reverse strand events. A good example of where this can occur was with a reverse strand event located on chromosome 11, at positions 10,628,861 to 10,628,908, with an event probability of 99.999%, consisting of 3 novel and unique peptides with 8 PSMs assigned, but on inspection appeared more likely to be a novel gene event. In addition Augustus predicted a new gene in this location using the novel peptides and exon_part hints (Figure 5.2, and with 8 supporting annotated MS/MS spectra in Appendix Figure 5.4).

Performing a BLASTP search against the grape family in NR revealed all three

novel peptides matched 40S ribosomal protein S8 (XP_010646836.1 with E-value ranges: 1E-07 – 2E-08) with 100% query coverage and identity, and which had EST and mRNA evidence supporting it. The new Augustus gene prediction matched significantly but with poor coverage of 29% and identity of 51% to a hypothetical protein (CAN81604.1 with E-value 5E-21), which likely indicated a truly novel identification.

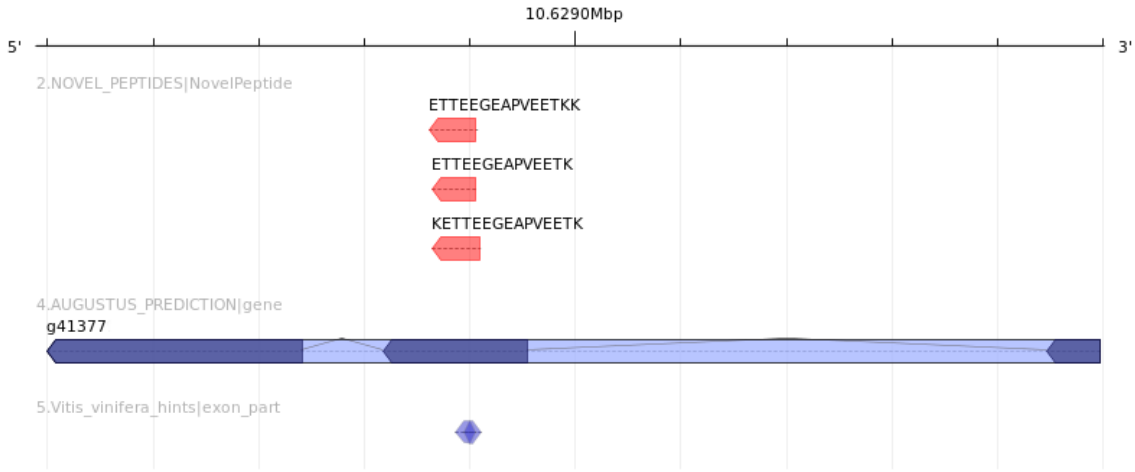


Figure 5.2 Novel gene annotation and prediction misidentified as a reverse strand event

The novel gene annotation was identified through what was inferred as a reverse strand event, due to the peptide linkage distance including a nearby gene. The novel peptides were incorporated into a new Augustus gene prediction, which were also supported by exon_part (EST and cDNA evidence) hints.

5.3.6 Gene boundary annotations

There were 160 (24 exclusive) gene boundary events identified (Table 5.1). However, this could have been artificially inflated as the peptide linkage distance was applied evenly across all peptide clusters identified in the genome, as a number of genes from the reference annotation could be seen in close proximity to each other, well within the 18,000 bp peptide linkage distance.

An example of a gene boundary event is on chromosome 18, spanning positions 4,109,211 to 4,112,683, with gene VIT_218s0001g04980, consisting of two protein isoforms with an event probability of 100% and with 13 PSMs identifying 4 unique peptides. In addition, an Augustus gene prediction using the novel peptides and reference annotations as hints was able to improve the gene model, incorporating the

novel peptides, previous annotations, exon and intron hints, into the new prediction (Figure 5.3, and with 13 supporting annotated MS/MS spectra in Appendix Figure 5.5).

Performing a BLASTP search against the grape family in NR revealed that all novel peptides matched acetyl-CoA carboxylase 1-like protein (XP_002285808.2 with E-value range: 5E-07 – 3E-16), with 100% query coverage and identity. The two reference protein-coding transcripts also matched acetyl-CoA carboxylase 1-like protein (XP_002285808.2 both with E-values = 0.0); protein-coding transcript 1 with 100% query coverage and protein-coding transcript 2 with 98% query coverage, both with 100% identity. The new Augustus gene prediction also matched acetyl-CoA carboxylase 1-like protein (XP_002285808.2 with E-value = 0.0), with 100% query coverage and identity, which showed that the original prediction was under-predicted, requiring a further extension of the gene towards the 5’ region on the reverse strand.

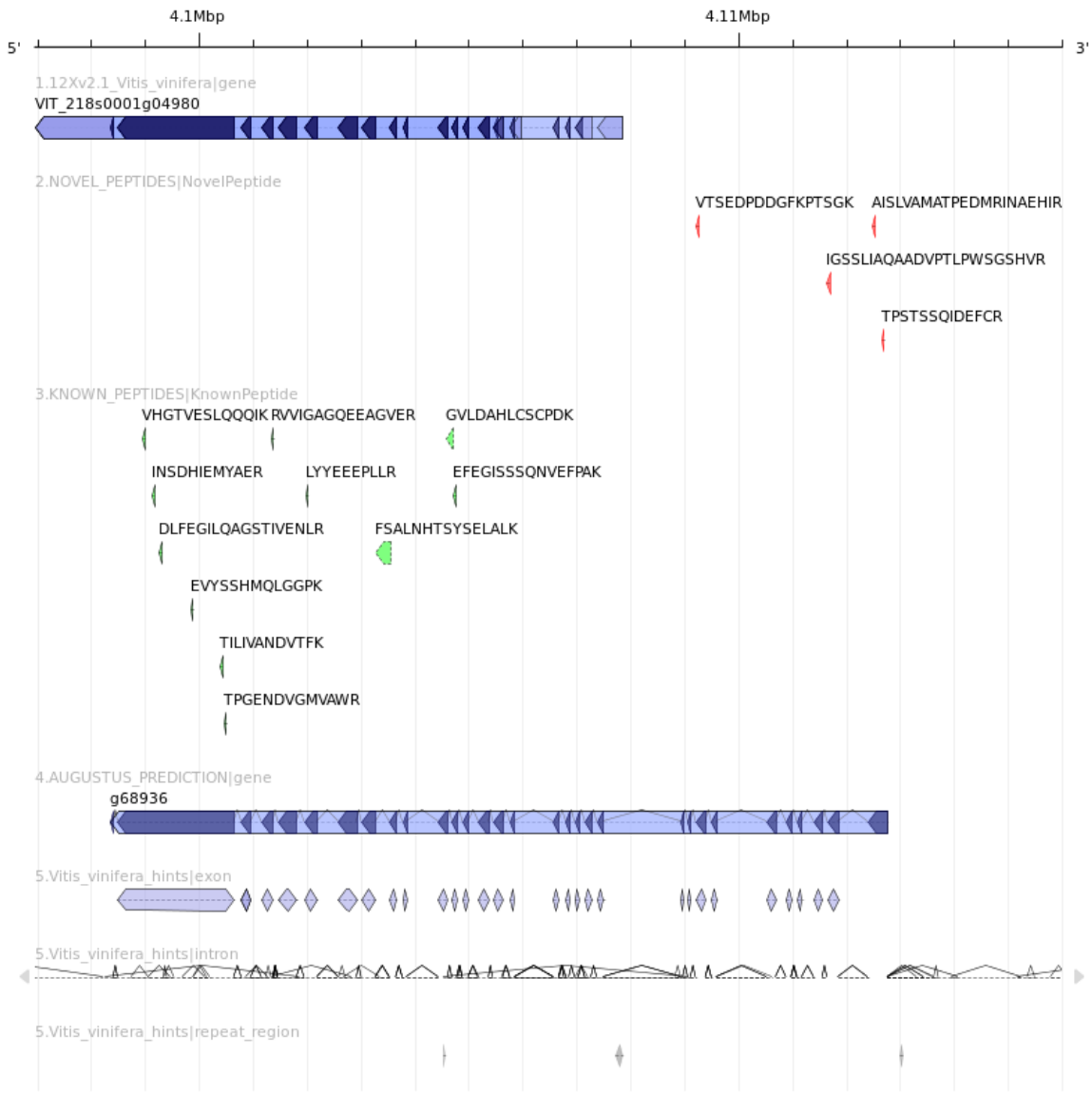


Figure 5.3 Gene boundary annotation

The gene boundary event inferred from the novel and unique peptides closely flanked reference gene VIT_218s0001g04980 from the 12Xv2.1 annotation. The novel peptides, reference gene VIT_218s0001g04980, exon (EST and cDNA evidence) and intron (RNA-seq evidence) hints were incorporated into the Augustus gene prediction. A group of peptides were also found mapped to gene VIT_218s0001g04980 indicating its expression and adding confidence to this proteogenomic annotation.

In the gene boundary annotation event, genes VIT_218s0001g05020 and VIT_218s0001g05030 were identified further upstream (based on the peptide linkage distance), but the novel peptides were in closer proximity to gene VIT_218s0001g04980, which was also the only gene amongst them with known mapped peptides. In addition, this peptide cluster also identified a reverse strand annotation event for gene VIT_218s0001g04970 further upstream, but this finding was also unlikely to be a 'real' annotation and was only inferred from the large peptide

linkage distance.

As outlined previously in Section 5.3.5, a fixed peptide linkage distance across the entire genome is a generalisation of the distribution of genes, inadvertently grouping genes into annotation events that do not belong and/or grouping peptides into peptide clusters that belong to separate annotation events.

Numerous other gene boundary events were also discovered with high event probabilities and multiple unique peptides, but these findings were also identified as translated UTR and reverse strand events, and appeared to agree more with these annotation event types than with the gene boundary event type. As such, to identify the true number of the different types of annotation events, manual screening of all annotation events was needed to see which annotation event was more likely the true case, by identifying which peptide clusters from the gene boundary and reverse strand events were exclusively identified as only those annotation events (Table 5.1), and were not also identified as other annotation events such as proximal events. This can help resolve some of the ambiguity when trying to interpret what the location of a peptide cluster could actually mean in relation to the surrounding genes instead of automatically inferring the annotation event based on its proximity to genes and the peptide linkage distance. However, a step of this nature is at odds with one of the aims of proteogenomics, i.e. to fast-track genome annotation without any loss in the quality of the annotation.

One way of improving the genomic annotation is to simply ignore classifying annotation events and focus on the identified novel peptides as hints towards gene prediction in tools such as Augustus, and to filter out any possible false gene predictions using orthologous evidence to group genes into low and high confidence sets. However, a method of this nature becomes more time-consuming, introduces error and avoids

assigning annotation events entirely, and relies heavily on the abilities of the gene prediction tool. A better approach to assigning annotation events at the proteogenomics level would be to re-define the way in which annotation events are categorized, based on additional evidence, such as known mapped peptides and their frame in correlation to the nearby novel peptides within the same or overlapping ORF.

An additional approach could be to determine the peptide linkage distance dynamically for each peptide cluster, based on the average size of genes in the immediate region in combination with other evidence such as ESTs, RNA-seq and/or the distribution of all mapped peptides and applied to machine-learning approaches, before annotation event inference, to identify the most likely genes involved in each annotation event. For example, a reverse strand annotation event would more likely to be defined for a gene when the peptide cluster directly overlaps it and has no neighbouring genes, particularly none with mapped known peptides. However, if the neighbouring genes have mapped peptides and the genes are in relatively close proximity, then the annotation could be defined as a gene boundary event overlapping the gene on the opposite strand. Such problems need to be met in a dynamic way, as the peptide linkage distance does not discriminate between entities in relatively close proximity, and those much further up or downstream of the peptide cluster.

5.3.7 Reverse strand annotation event leads to a new gene prediction

There were 112 (10 exclusive) reverse strand annotation events identified (Table 5.1), but the majority of these annotation events were also identified as gene boundary and translated UTR events due to the peptide linkage distance. However in 8 cases the peptide cluster directly overlapped a gene on the opposite strand. An example of one of these 8 reverse strand annotations was gene VIT_211s0065g00070 with its single protein-coding transcript, which was also identified as a gene boundary annotation for gene VIT_211s0065g00060 much further upstream. This annotation was identified on

chromosome 11, spanning positions 13,140,580 to 13,147,463, with an event probability of 99.960% with 2 PSMs identifying 2 unique and novel peptides. Additionally, Augustus predicted a novel gene from one of the novel peptides (Figure 5.4, and with 2 supporting annotated MS/MS spectra in Appendix Figure 5.6).

Performing a BLASTP search against the grape family in NR revealed that novel peptide “ELSNYMQPIHIHTIYLPILLLIKMEK“ matched a hypothetical protein (CAN73713.1 with E-value = 0.030), with 80% query coverage and 56% identity, and novel peptide “KVEENLNNEAR” matched a protein transport protein SEC16A homolog (XP_010646525.1 with E-value = 7E-04), with 90% query coverage and 100% identity. The reference protein-coding transcript from gene VIT_211s0065g00070 matched unnamed protein product (CBI24042.3 with E-value = 0.0), with 100% query coverage and identity. The new Augustus gene prediction matched hypothetical protein (CAN82660.1 with E-value = 2E-35), with 57% query coverage and 85% identity, indicating novel gene identification on the reverse strand. Although the reference protein matched significantly in NR, for unknown reasons it did not lead to a new equivalent prediction through Augustus. Further supporting evidence is needed before this annotation event and prediction could be accepted with confidence, as the only evidence supporting the novel prediction was a single unique peptide.

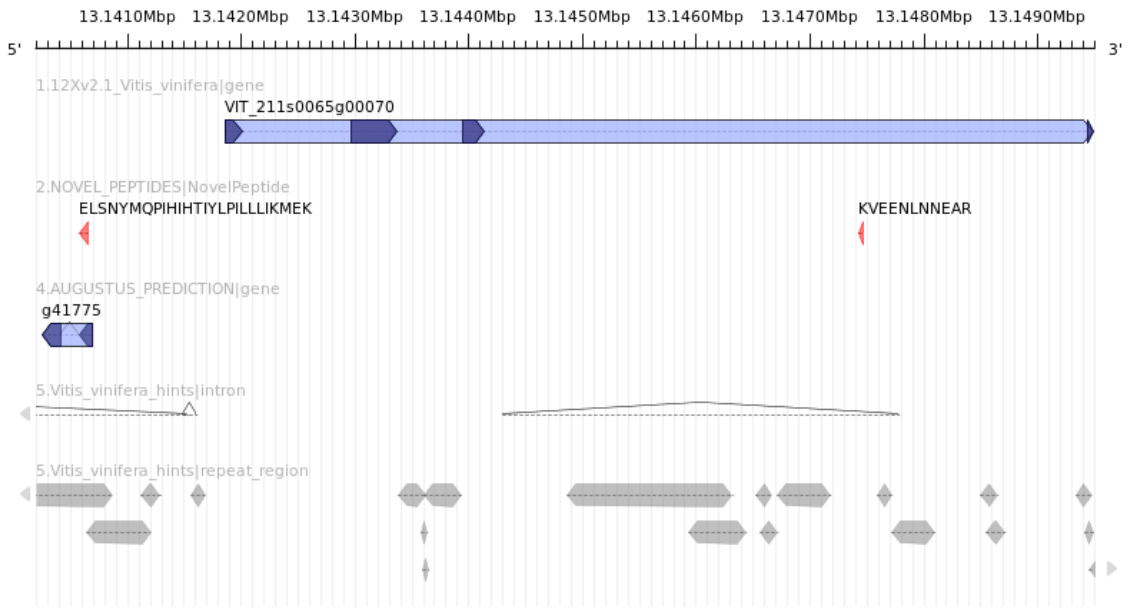


Figure 5.4 Novel gene prediction via a reverse strand event

A reverse strand event inferred from the novel and unique peptides. The novel peptide "KVEENLNNEAR" and reference gene VIT_211s0065g00070 used as a hint for Augustus gene prediction did not result in a new gene. However, peptide "ELSNYMQPIHIHTIYLPILLLIKMEK" contributed to a new Augustus gene prediction, flanking gene VIT_211s0065g00070. The intron (RNA-seq evidence) hints in this region did not contribute to any new Augustus gene predictions. The region is also covered in a number of repeat regions.

5.3.8 Translated UTR annotation

There were 37 translated UTR annotations identified (Table 5.1), of which a large proportion were also identified as gene boundary and reverse strand annotations due to the peptide linkage distance including other genes in close proximity. However, the translated UTR annotations by their definition are limited to only directly overlapping a UTR gene region. Hence, the reason for just 40 such annotations, compared to the other distal events such as reverse strands (112) and gene boundaries (159), which were annotated purely on the inclusion of a gene within the peptide linkage distance of a peptide cluster. An example was with gene VIT_208s0058g01030, protein-coding transcript 1, with the peptide cluster also indicating a gene boundary with genes VIT_208s0058g01020, VIT_208s0058g01030 and VIT_208s0058g01040, as well as a reverse strand annotation with gene VIT_208s0058g01010, due to the peptide linkage distance and close proximity of these genes around gene VIT_208s0058g01030. This annotation was identified on chromosome 8, spanning positions 10,371,636 to

10,373,023, with an event probability of 100% and with 6 PSMs identifying 4 novel and unique peptides. Using the novel peptides, reference annotation and intron hints evidence for gene prediction, Augustus predicted two genes (Figure 5.5, and with 6 supporting annotated MS/MS spectra in Appendix Figure 5.7).

Performing a BLASTP search against the grape family in NR revealed that all novel peptides matched Prosaposin protein (XP_002268581.1 with E-value range: 0.009 – 1E-11), with 100% query coverage and identity, the reference protein matched unnamed protein product (CBI18061.3 with E-value = 0.0), with 100% query coverage and identity; described as containing a Saposin-like type B domain. The two new Augustus gene predictions, g29772 and g29776, also matched unnamed protein product (CBI18062.3 for g29772 and CBI18061.3 for g29776, both with E-values = 0.0), with 72% query coverage for g29772 and 87% query coverage for g29776, both with 100% identity; and both described as containing a Saposin-like type B domain. The two new gene predictions matching the same protein, have likely been split into two prediction based on the evidence. In addition, the intron evidence (Figure 5.5) indicates splicing across the whole length of the region from gene VIT_208s0058g01030, which also crosses the regions of both new predictions g29772 and g29776, thus providing further evidence that the two new gene predictions belong in a single prediction as opposed to two.

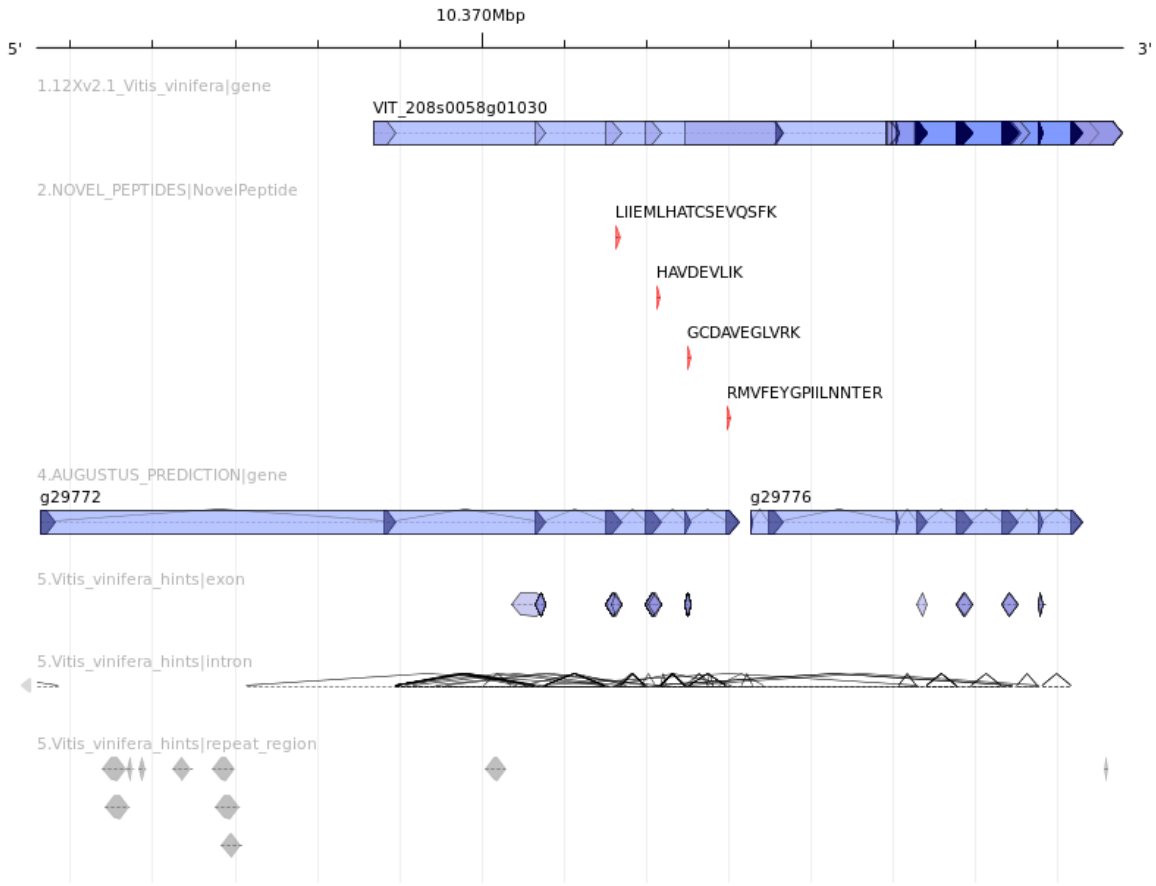


Figure 5.5 Translated UTR annotation

A translated UTR event inferred from the novel and unique peptides. The novel peptides, reference gene VIT_208s0058g01030 and exon (EST and cDNA evidence) and intron (RNA-seq evidence) hints were incorporated into new Augustus gene predictions, however the predictions were split into two separate predictions both matching the same protein. A number of repeat regions were identified upstream of the 5' end of gene VIT_208s0058g01030.

Another interesting translated UTR annotation was for gene VIT_207s0031g03000, with the peptide cluster also part of a larger peptide cluster indicating a reverse strand annotation for genes VIT_207s0031g02980, VIT_207s0031g02990, and VIT_207s0031g03010 further upstream and downstream, based on the peptide linkage distance. The gene VIT_207s0031g03000 and its single protein-coding transcript also contained a different peptide cluster, indicating an exon boundary annotation, which is discussed later in Section 5.3.10. The translated UTR annotation for gene VIT_207s0031g03000 was identified on chromosome 7, spanning positions 19,731,734 to 19,731,934 with an event probability of 100%, with 90 PSMs identifying 4 unique and 5 shared novel peptides. Additionally, using the novel

peptides, reference annotations, intron and exon hints, Augustus predicted a new gene (Figure 5.6 (translated UTR peptide cluster outlined in dash dotted line), and with a sample of 9 of 90 supporting annotated MS/MS spectra in Appendix Figure 5.8).

Performing a BLASTP search against the grape family in NR revealed that all novel peptides matched the ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCo) large subunit (AFG24212.1 with E-value range: 4E-05 - 8E-14) with 100% query coverage and identity. The reference protein matched unnamed protein product (CBI21646.3 with E-value = 2E-62), with 100% query coverage and identity; described as containing a RuBisCo large chain domain. The new Augustus gene prediction matched hypothetical protein (CAN63541.1 with E-value = 5E-166) with 77% query coverage and 99% identity; also described as containing a RuBisCo domain.

All evidence indicated that the identified and refined protein-coding gene was a RuBisCo large subunit, which is widely known as a dominant protein in the plant kingdom found more abundantly in leaves, a major source of peptides in this study. However, the protein, in particular the large subunit, was found exclusively in chloroplasts. False identifications such as these would be expected to match randomly across the genome by chance alone, but they are distributed across the length of this gene. The single protein-coding transcript from gene VIT_207s0031g03000 also matched RuBisCo large subunit, as does the new Augustus gene model. This indicated that either this was the first likely known case of an actual RuBisCo gene encoded on chromosomes in the nucleus, which seems improbable, or there is some over-assembly of the reference genome sequence, with chloroplast reads being incorporated into the chromosome 7 assembly. Further evidence as to the origin of this gene and the overall region in which the gene is found is needed before any further conclusions can be drawn. A good indication that over-assembly is the likely cause can be seen from the large number of chromosome fragments and the large unassigned chromosome

(ChrUn), as pointed out previously in Section 5.3.3, in regards to sources of false positives. Highly fragmented genomes with a large number of unassigned chromosomes can often indicate an underlying problem with the assembly, either indicating unresolved repeat regions at the ends of the scaffolds or co-assembly with nuclear chromosome reads and contaminant reads (e.g. incorporating reads from the chloroplast) leading to over-assembly; both of which would significantly hamper attempts to join scaffolds, leaving the assembly in a disjointed state and fragmented.

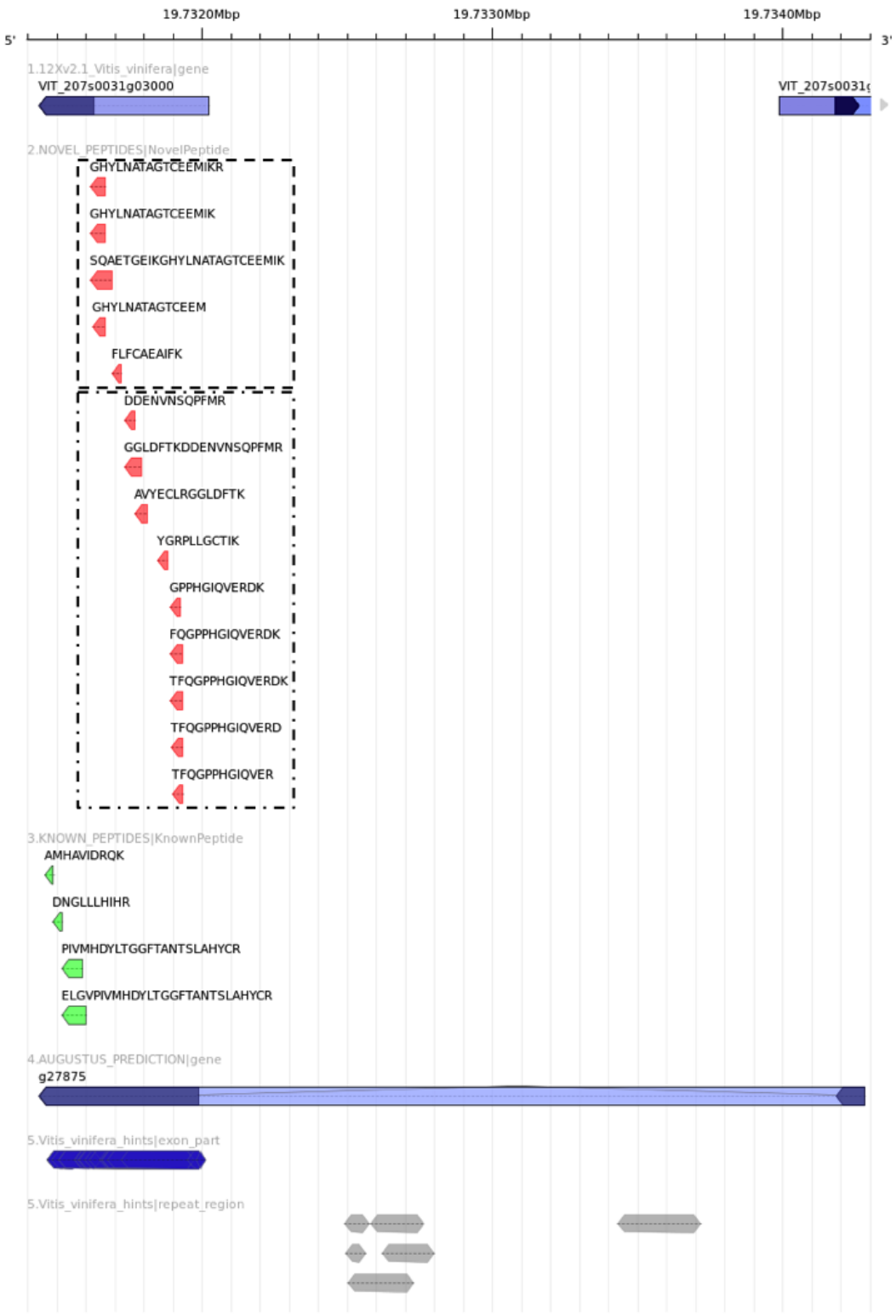


Figure 5.6 Translated UTR and exon boundary annotation

An exon boundary (dashed line) and translated UTR (dash dotted line) inferred from the novel and unique peptides. The novel peptides, reference gene VIT_207s0031g03000, exon_part (EST and cDNA evidence) hints were incorporated into the Augustus gene prediction. A number of repeat regions were identified upstream of the 5' end of gene VIT_207s0031g03000. The new Augustus gene prediction was also predicted to be spliced, with GT-AG (U2 spliceosome) donor-acceptor sites, however no other evidence indicated splicing across this region.

5.3.9 Novel splice annotation

One novel splice annotation was identified (Table 5.1) for gene VIT_201s0011g05110 and its single protein-coding transcript. The novel peptide was also identified as a gene boundary annotation for a number of other genes in close proximity, including genes VIT_201s0011g05082, VIT_201s0011g05090, VIT_201s0011g05100, VIT_201s0011g05110, VIT_201s0011g05120, VIT_201s0011g05130, and VIT_201s0011g05140. The novel splice annotation event was identified on chromosome 1, spanning positions 4,748,591 to 4,761,837, with an event probability of 99.396% and 1 PSM identifying a single novel and unique peptide. Additionally, using the novel spliced peptide to identify a new or revised Augustus gene prediction did not lead to any predictions (Figure 5.7, and with 1 supporting annotated MS/MS spectrum in Appendix Figure 5.9).

Performing a BLASTP search against the grape family in NR revealed the novel peptide matched hypothetical protein (CAN83049.1 with E-value = 0.012), with 75% query coverage and 100% identity. The reference protein also matched the same hypothetical protein (CAN83049.1 with E-value = 2E-165), with 100% query coverage and identity, described as containing a hydrophobic ligand binding site domain of a major pollen allergen. The novel peptide showed only 75% coverage as the N-terminal end of the peptide resided within the protein, and the C-terminal resided further downstream within an intergenic space. However, there was no evidence to support the inferred intron from the proteogenomics analysis, which indicated a GT-GC donor-acceptor site (U2 spliceosome) at positions 4,748,617 to 4,761,824. To see if this putative splice site could be identified via a different strategy, the genomic region spanning the exon where the N-terminal end of the novel peptide resided was extracted up to 20 bp beyond the C-terminal mapped end of the novel peptide, and was given to the NetGene2 splice site prediction web server using *Arabidopsis thaliana* as a model

[579]. No donor or acceptor sites matching those predicted by the proteogenomics analysis could be found (Appendix File 5.7). Although the intron hints contained a number of introns across the region none spanned the entire length of the region where the proteogenomics evidence indicated, which suggested that the novel spliced peptide was probably a false positive.

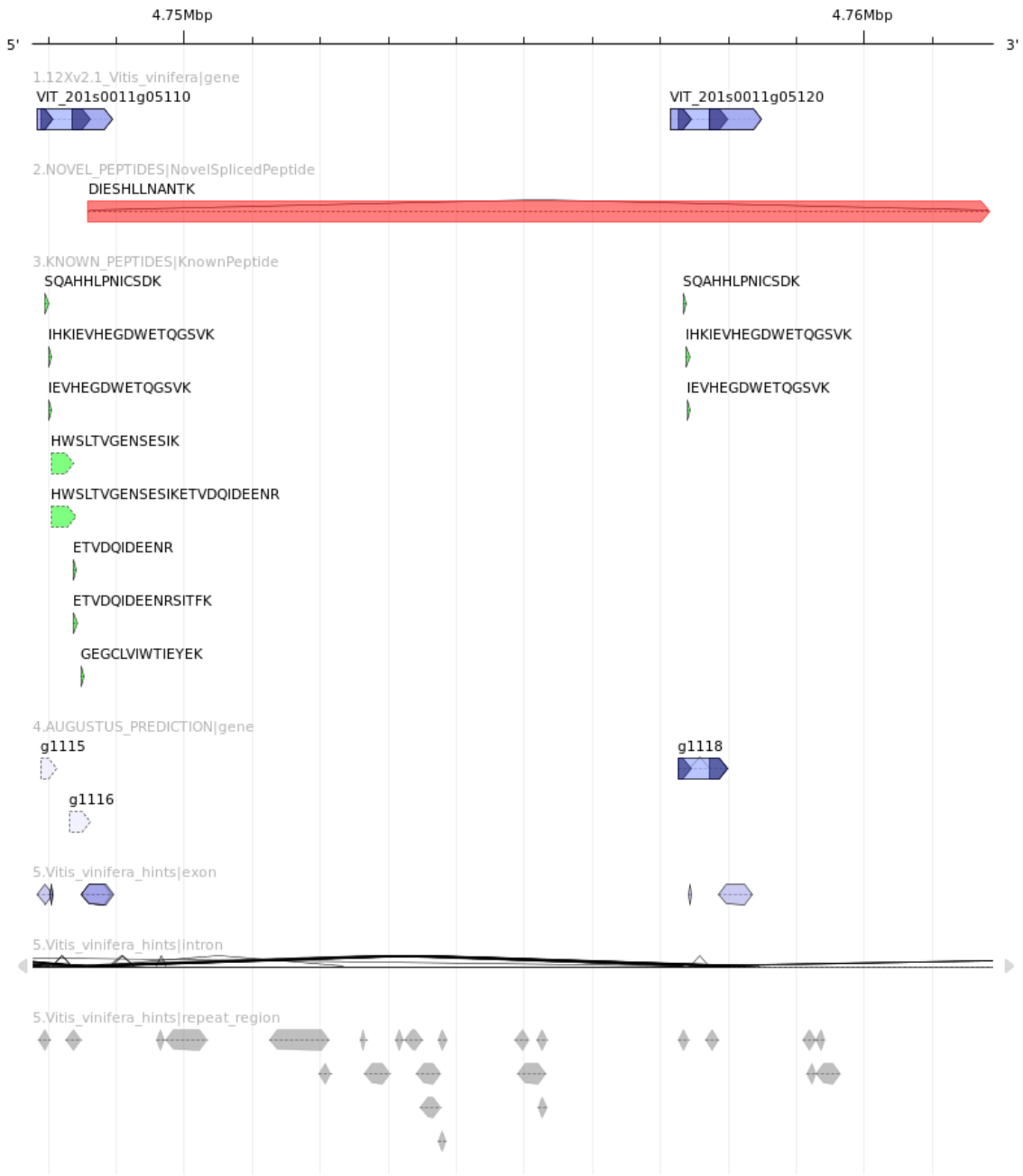


Figure 5.7 Novel splice annotation

A novel splice annotation inferred from the novel and unique peptide. The novel peptide could not be incorporated into the Augustus gene prediction. Only the genes in the region from the 12Xv2.1 annotation led to Augustus gene predictions that were supported by mapped known peptides. There were also no introns identified among the intron hints from the RNA-seq evidence that spans the entire range inferred by the novel spliced peptide. There are also a number of repeat regions dotted across the length of the region.

5.3.10 Exon boundary annotation

There were 16 exon boundary annotations identified (Table 5.1), a number of which were also identified as gene boundary and reverse strand annotations due to the peptide linkage distance including other genes in close proximity, as well as a few novel exons

identified from different protein isoforms from the same gene. An example of an exon boundary annotation was with gene VIT_207s0031g03000, mentioned previously containing a translated UTR annotation, with the peptide cluster also part of a larger peptide cluster indicating a reverse strand annotation for genes VIT_207s0031g02980, VIT_207s0031g02990, and VIT_207s0031g03010 further upstream and downstream. The exon boundary annotation for gene VIT_207s0031g03000 and its single protein-coding transcript was identified on chromosome 7, spanning positions 19,731,614 to 19,731,721 with an event probability of 100%, and 59 PSMs identifying five novel and unique peptides. Using the novel peptides and the current annotation as hints, Augustus predicted a new gene model (Figure 5.6 (exon boundary peptide cluster outlined in dashed line), and with a sample of 5 of 59 supporting annotated MS/MS spectra in Appendix Figure 5.10).

Performing a BLASTP search against the grape family in NR revealed that all the novel peptides matched unnamed protein product (CBI21646.3 with E-value range: 2E-04 - 8E-20), with 100% query coverage and identity, described as containing a RuBisCo large chain domain, which is in agreement with the findings from the translated UTR event described previously for the same gene. The reference protein and Augustus gene prediction matched unnamed protein product (CBI21646.3) and hypothetical protein (CAN63541.1), respectively, as described previously in Section 5.3.8 for a translated UTR annotation.

Another interesting exon boundary annotation was for gene VIT_206s0004g04440 and its single protein-coding transcript, with the peptide cluster also indicating a gene boundary annotation for genes VIT_206s0004g04410, VIT_206s0004g04430 and VIT_206s0004g04460, as well as a reverse strand annotation for genes VIT_206s0004g04420, VIT_206s0004g04450 and VIT_206s0004g04470, due to the peptide linkage distance and close proximity of these

genes around gene VIT_206s0004g04440. This annotation was identified on chromosome 6 spanning positions 5,401,503 to 5,401,583, with an event probability of 99.999% and 12 PSMs identifying 2 novel and unique peptides. Using the novel peptides and current annotation as hints, Augustus predicted a new gene model (Figure 5.8, and with 12 supporting annotated MS/MS spectra in Appendix Figure 5.11).

Performing a BLASTP search against the grape family in NR revealed that the novel peptides matched osmotin-like protein (XP_002281193.1 with E-value range: 2E-19 – 2E-22), with 100% query coverage and identity. The reference protein also matched osmotin-like protein (XP_002281193.1 with E-value = 0.0), with 91% query coverage and 100% identity and the Augustus gene prediction also matched osmotin-like protein (XP_002281193.1 with E-value = 0.0), with 94% query coverage and 100% identity. The revised gene prediction showed an improvement on the original gene prediction of 3% coverage, extending the original exon 1 and merging it with exon 2. In addition, the identified osmotin-like protein was also supported by EST and mRNA evidence on chromosome 6 within the same genomic coordinates as this annotation event.

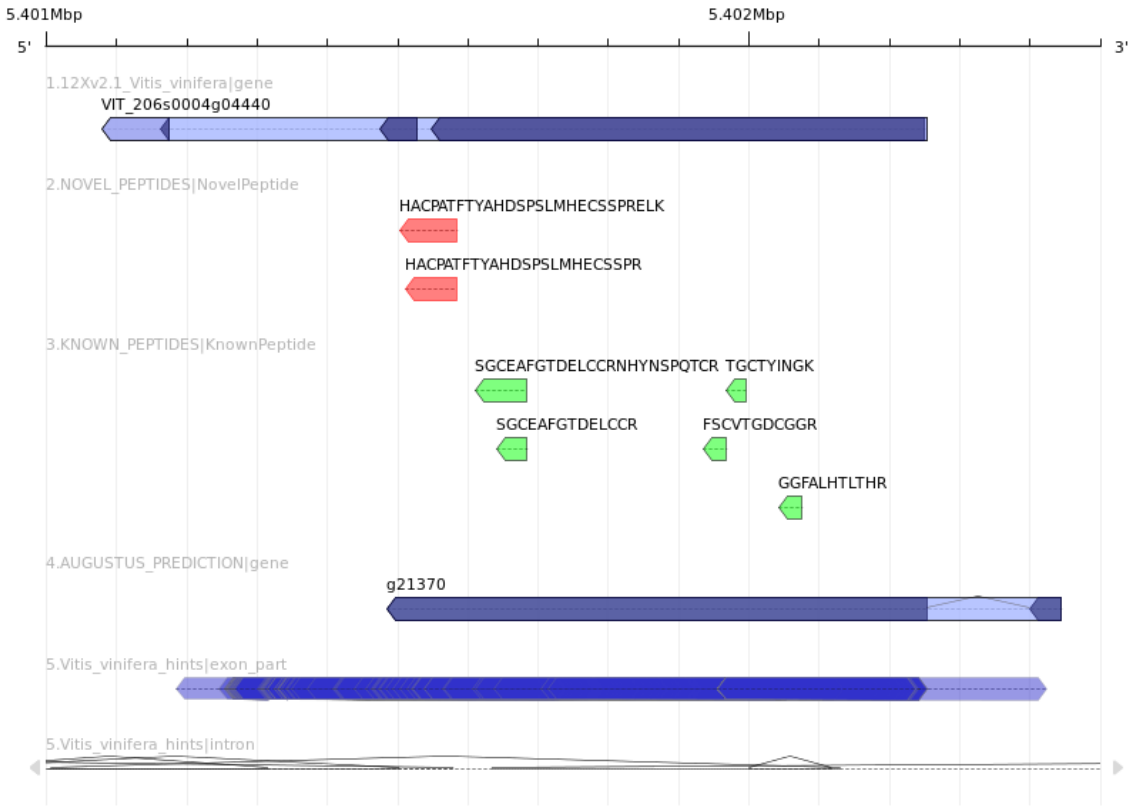


Figure 5.8 Exon boundary annotation

An exon boundary event inferred from the novel and unique peptides. The novel peptides, reference gene VIT_206s0004g04440, exon_part (EST and cDNA evidence) hints were incorporated into the Augustus gene prediction. The new Augustus gene prediction was also predicted with a spliced region containing GT-AG (U2 spliceosome) donor-acceptor sites. However, no intron (RNA-seq) evidence indicated splicing across the region indicated by the prediction.

5.3.11 Frame-shift annotation

There were 5 frame-shift annotations identified (Table 5.1), a number of which were also identified as gene boundary and reverse strand annotations due to the peptide linkage distance including other genes in close proximity. An example of a frame-shift annotation was with gene VIT_200s1339g00010 and its single protein-coding transcript, with the peptide cluster also indicating a gene boundary annotation for gene VIT_200s1343g00010 and a reverse strand annotation for genes VIT_200s1338g00010 and VIT_200s1338g00020, due to the peptide linkage distance. This annotation was identified on chromosome Un, spanning positions 38,615,528 to 38,615,653, with an event probability of 99.998% and 3 PSMs identifying 2 novel and unique peptides. Frame-shift events can also be identified due to incorrect CDS phase. Numerous CDS phase were found to be incorrect throughout the 12Xv2.1 annotation and were not

solely limited to those identified in the proteogenomics analysis. All CDS phase in the genome annotation were subsequently corrected using GenomeTools [542]. The CDS phase corrected reference annotation was then used as hints along with the identified novel frame-shift peptides using Augustus to improve the predictions and create a new gene model (Figure 5.9, and with 3 supporting annotated MS/MS spectra in Appendix Figure 5.12).

Performing a BLASTP search against the grape family in NR revealed the novel peptides matched a hypothetical protein (CAN63109.1 with E-value ranges: 2E-07 – 5E-21), with 100% query coverage and identity, described as containing a Ribonuclease T2 domain. The reference protein matched a hypothetical protein (CAN63794.1 with E-value = 3.3), with 65% query coverage and 50% identity, described as containing a retrotransposon gag protein domain. The Augustus gene prediction matched the same hypothetical protein as the novel peptides (CAN63109.1 with E-value = 0.0), with 97% query coverage and 98% identity. The BLASTP evidence indicated that the original reference protein had a poor match to a hypothetical protein, while the revised Augustus gene prediction found a significant match to a hypothetical protein, with a good protein alignment that included the novel peptides. In addition, the novel peptides and Augustus gene prediction both matched exactly the same protein, described as containing a Ribonuclease T2 domain implicated in plant leaf senescence, which correlates well with the source of the proteomics data, mainly being derived from plant leaf shoot tips. This proteogenomics evidence has led to a significant overall improvement to the annotation of this gene.

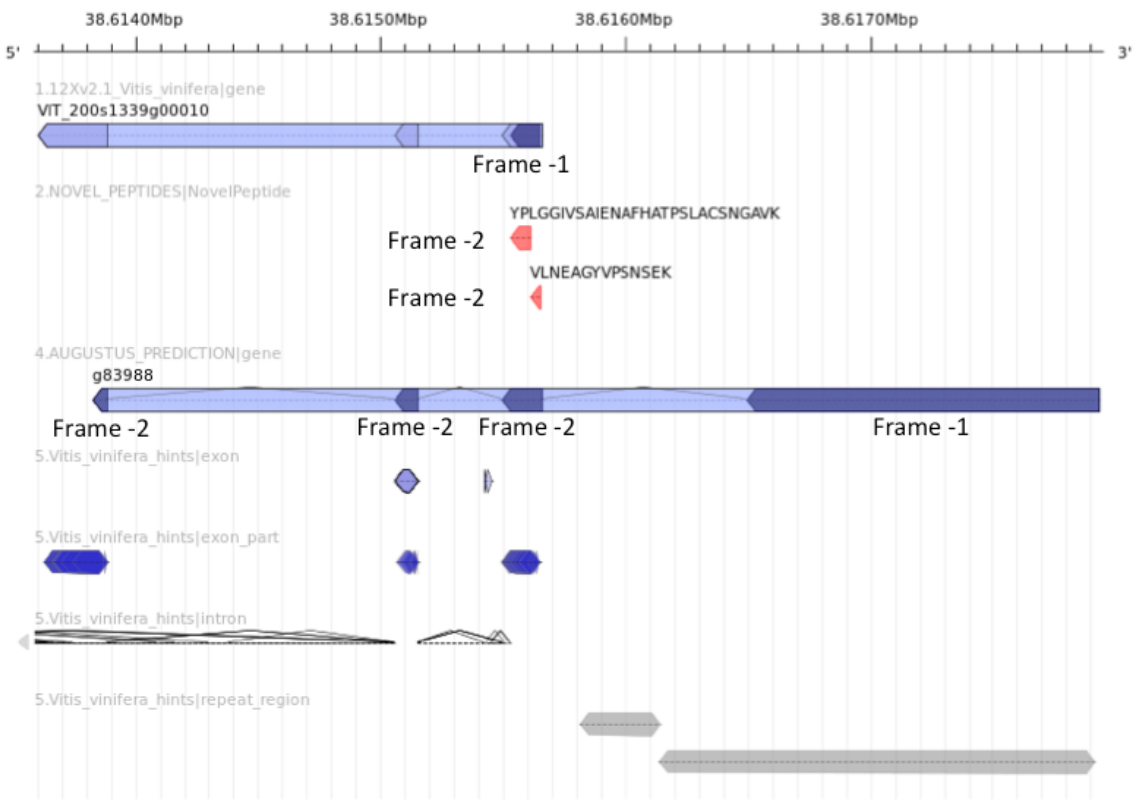


Figure 5.9 Frame-shift annotation

A frame-shift event inferred from the novel and unique peptides. The novel peptides, gene VIT_200s1339g00010, exon and exon_part (EST and cDNA) and intron (RNA-seq) evidence were used as hints for the Augustus gene prediction. All hints were incorporated into the prediction except for the hints from gene VIT_200s1339g00010, which had a CDS region in a different frame (frame -1) than that of the novel peptides (frame -2). A region further downstream contained a number of repeats.

5.3.12 Novel exon annotation

There were 9 novel exon annotations identified (Table 5.1), with a number also identified as gene boundary, reverse strand and translated UTR annotations, due to the peptide linkage distance including other genes in close proximity. An example of a novel exon annotation was with gene VIT_217s0000g02480, the same peptide cluster was also categorised as a gene boundary annotation for genes VIT_217s0000g02470, VIT_217s0000g02490 and VIT_217s0000g02500 due to the peptide linkage distance. This annotation was identified on chromosome 17, spanning positions 2,264,427 to 2,264,591, with an event probability of 99.999% and 15 PSMs identifying 3 novel and unique peptides. The novel peptides and the reference annotation were then used as hints for Augustus gene prediction, resulting in the incorporation of the novel peptides into a new gene model with complete removal of the intron (Figure 5.10, and with 15

supporting annotated MS/MS spectra in Appendix Figure 5.13).

Performing a BLASTP search against the grape family in NR revealed the novel peptides matched calcium-binding allergen Ole e 8 (XP_010663678.1 with E-value ranges: 2E-07 – 2E-15), with 100% query coverage and identity. The reference protein matched unnamed protein product (CBI15562.3 with E-value = 2E-97), with 100% query coverage and identity, described as containing an EF-hand calcium-binding domain. The Augustus gene prediction matched calcium-binding allergen Ole e 8 (XP_010663678.1 with E-value = 2E-176), with 100% query coverage and identity. Overall, the new Augustus gene prediction had a better match to the calcium-binding protein when the novel peptides were included and the intron was removed.

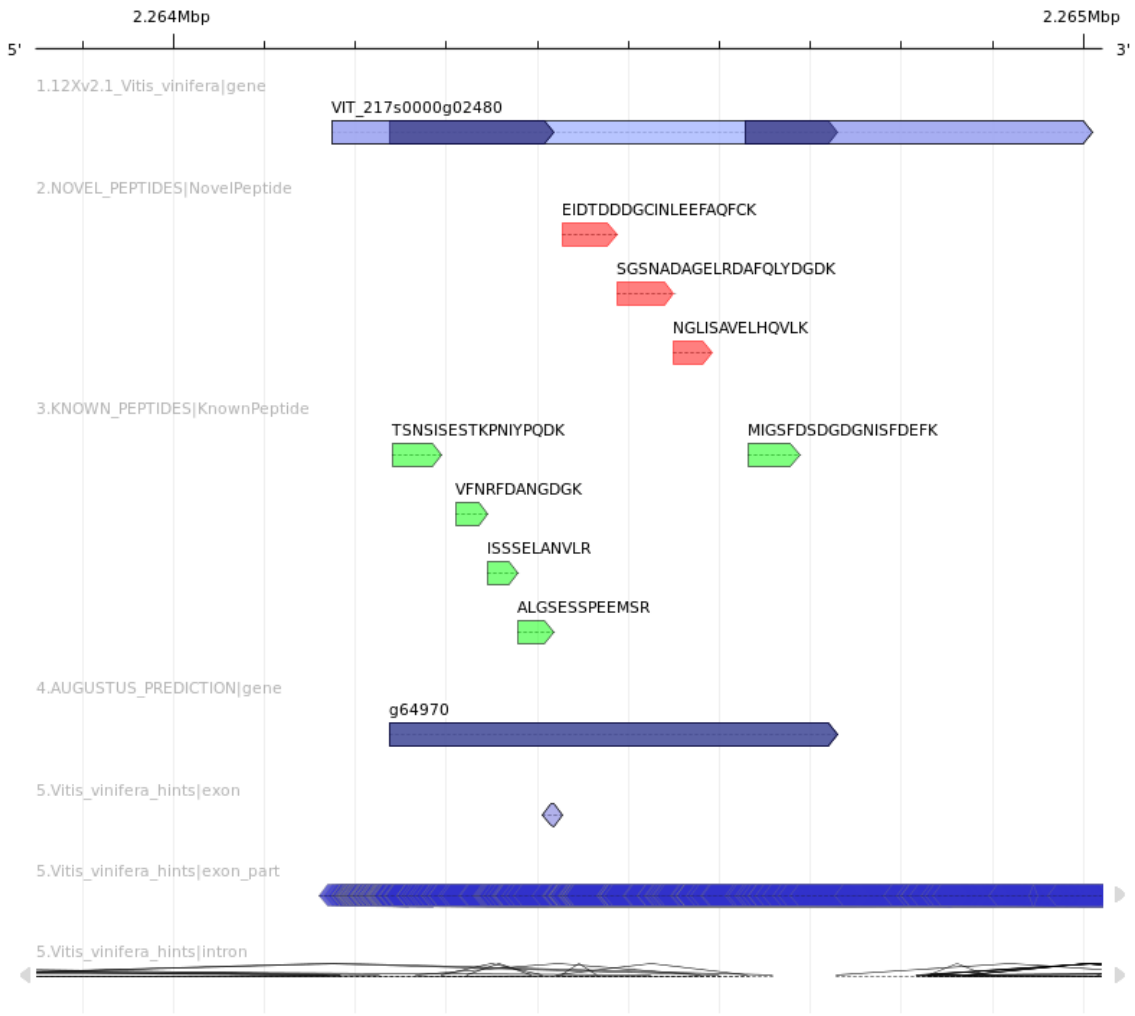


Figure 5.10 Novel exon annotation

A novel exon event inferred from the novel and unique peptides. The novel peptides, gene VIT_217s0000g02480, exon and exon_part (EST and cDNA) evidence were incorporated into the Augustus gene prediction, bridging the two exon/CDS regions in the original prediction into one exon/CDS region.

5.3.13 N-terminal acetylated peptides

Protein N-terminal acetylation contributes to many functional changes in proteins, from signalling, regulation of protein-protein interactions, and transportation of proteins to their target, such as embedding in membranes [580]. The identification of any N-terminal acetylated peptides not at the N-terminal ends of a known protein could potentially indicate an over-predicted gene requiring re-annotation, but could also indicate an alternative protein isoform with a different translation initiation start (TIS) site.

No N-terminal acetylated peptides were identified from the 133 novel peptides

(Table 5.1), however, a total of 192 N-terminal acetylated peptides were identified among the known 12Xv2.1 predicted proteins and 80 were identified from 77 high confidence proteins (>2 peptides with at least 1 unique). Of the total proteins, 5 shared N-terminal acetylated peptides were identified as conflicting with the 12Xv2.1 annotation (Appendix File 5.8), thus indicating a possible alternative TIS site, while none were identified from the high confidence proteins.

The first of these peptides was shared peptide "LQGEAVQDWREIVTYFSYPLR" identified in two different locations: 1) on chromosome 4, at positions 19,930,049 to 19,930,316, on gene VIT_204s0023g03370 with a single protein-coding transcript; and 2) on chromosome 18, at positions 12,303,451 to 12,304,114 on gene VIT_218s0001g14310 with a single protein-coding transcript. The suggestion of an alternative TIS site was not supported by the Augustus gene prediction, showing only one protein prediction in agreement with the 12Xv2.1 annotation, which was true for both predictions on chromosome 4 and 18. An example of the mapped peptides including the N-terminal acetylated peptide for chromosome 4 can be seen in Figure 5.11 (Supported with 1 annotated MS/MS spectrum in Appendix Figure 5.14). The N-terminal acetylated peptide was also found to be a spliced peptide in both identified locations and the spectral E-value for this identification was ~7.0E-14, indicating a significant spectral interpretation. Both proteins were found to share extensive sequence homology by Muscle alignment (data not shown).

Performing a BLASTP search against the grape family in NR, the single protein-coding transcript from gene VIT_204s0023g03370 was found to match naringenin,2-oxoglutarate 3-dioxygenase (NP_001268034.1 with E-value = 0.0) on chromosome 4, and the single protein-coding transcript from gene VIT_218s0001g14310 was found to match flavanone 3-dioxygenase (XP_002275563.1 with E-value = 0.0) on chromosome 18. Both proteins had their own unique peptides, but both also shared common peptides

including the N-terminal acetylated peptide. In addition, both peptides also contained unmodified shared peptide “GGFIVSSHLQGEAVQDWREIVTYFSYPLR”, of which the N-terminal acetylated peptide was a sub-string, indicating that the N-terminal acetylated peptide could have been a representative of an alternative isoform. However, the inferred translation initiation codon for both peptide locations on chromosome 4 and 18 is CAT, which does not appear to be a known alternative translation initiation codon in plants [581-584]. In addition, there were no GT-AG, GC-AG (U2 spliceosome) or AT-AC (U12 spliceosome) donor-acceptor sites preceding the peptide to use for splicing for an initiation codon further upstream. It is possible that other non-AUG TIS sites not commonly known of could explain the presence of the N-terminal acetylated peptide. Further evidence would be needed to confirm this potential alternative TIS site and/or protein isoform, particularly given that the peptide was also identified from multiple locations.

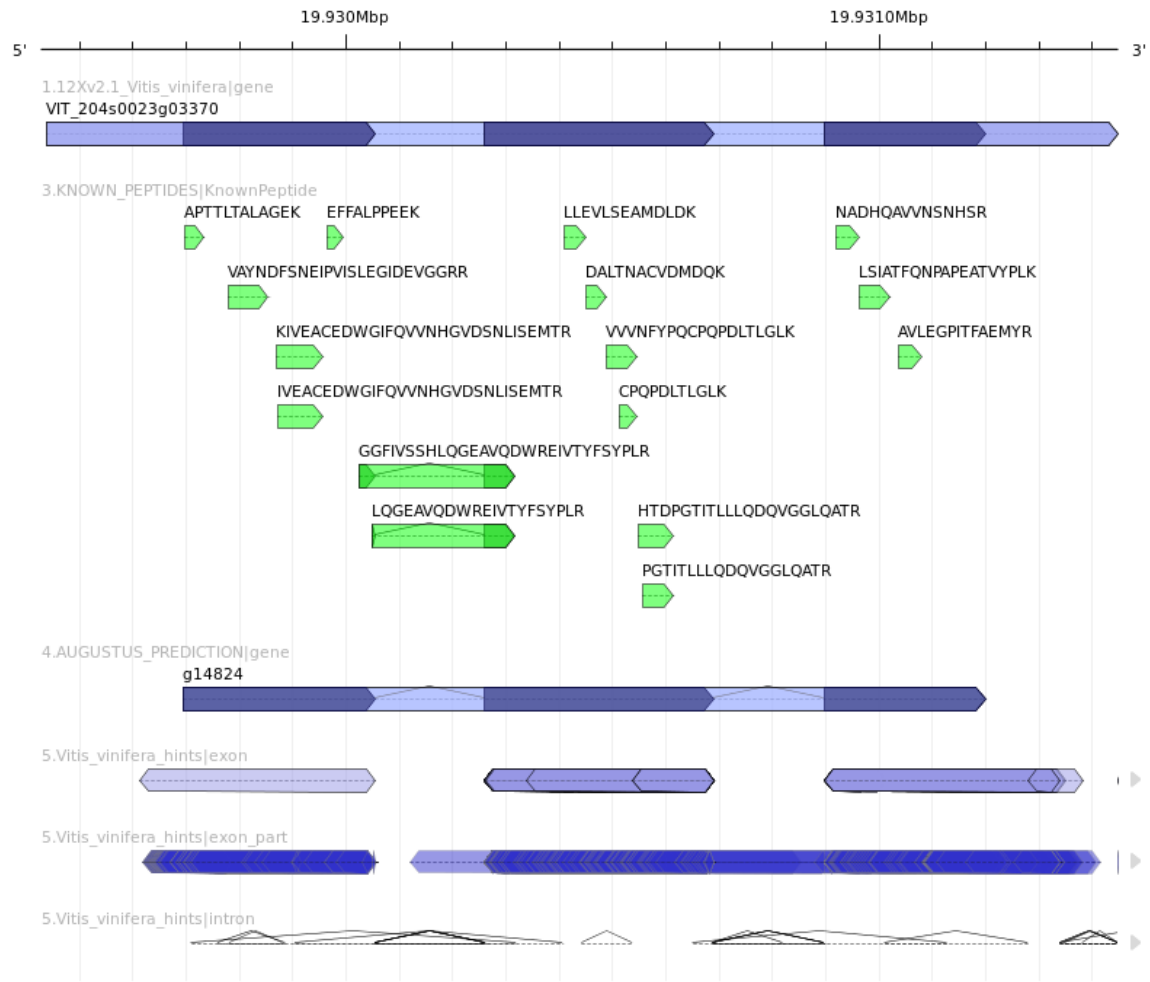


Figure 5.11 Known N-terminal acetylated peptide on chromosome 4

The known N-terminal acetylated peptide "LQGEAVQDWREIVTYFSYPLR", was shown to be a spliced peptide. However, there was no other evidence to indicate translation begins with this peptide and Augustus did not predict isoforms in agreement with it. The same was true for the peptide on chromosome 18, with similarly mapped peptides and gene prediction.

The second of these N-terminal acetylated peptides in conflict with the 12Xv2.1 annotation was shared peptide "MDTFLFTSESVNEGHPDK", identified in three different locations: 1) on chromosome 6, at positions 17,979,019 to 17,979,072 on gene VIT_206s0061g00500 with a single protein isoform; 2) on chromosome 8, at positions 18,919,477 to 18,919,530 on gene VIT_208s0007g05000 with two protein isoforms; and 3) on chromosome 14, at positions 451,155 to 451,208 on gene VIT_214s0060g00480 with four protein isoforms. Only the peptide identified on chromosome 6 at positions 17,979,019 to 17,979,072 was, however, in disagreement with the 12Xv2.1 annotations. Also, the suggested alternative TIS site was not supported by the Augustus gene prediction (Figure 5.12, and with 7 supporting

annotated MS/MS spectra in Appendix Figure 5.15). The spectral E-values for this identification across the different locations range from 1.0E-12 to 1.0E-15, indicating significant MS/MS spectral interpretations. All mapped proteins share a number of peptides, including the N-terminal acetylated peptide, and they also share extensive sequence homology, as indicated by Muscle alignment (data not shown).

Performing a BLASTP search against the grape family in NR, the single protein-coding transcript from gene VIT_206s0061g00500 was found to match S-adenosylmethionine synthase 1 isoform X1 (XP_002273336.1 with E-value = 0.0) on chromosome 6. Protein-coding transcripts from gene VIT_208s0007g05000 matched S-adenosylmethionine synthase 3 (XP_003632745.1 with E-value = 0.0) on chromosome 8. The four protein-coding transcripts from gene VIT_214s0060g00480 matched S-adenosylmethionine synthase 4 (XP_010659744.1 with E-value = 0.0) on chromosome 14. The inferred translation initiation codon is GTG, and is identified as a non-AUG translation initiation codon in *Arabidopsis thaliana* [583] *Rhabdopleura compacta* [585], Ascidian mitochondria [586] and chloroplasts [587]. There were also no splicing acceptor sites prior to the peptide, only an AG beginning 4 bp upstream within an already identified intron. Due to the validation of the GTG codon being identifiable as a known alternative translation initiation codon in plants, this peptide may indicate an alternate TIS and/or isoform. However, due to the presence of the peptide in multiple similar proteins in other locations, this annotation will remain ambiguous until further evidence can be presented.

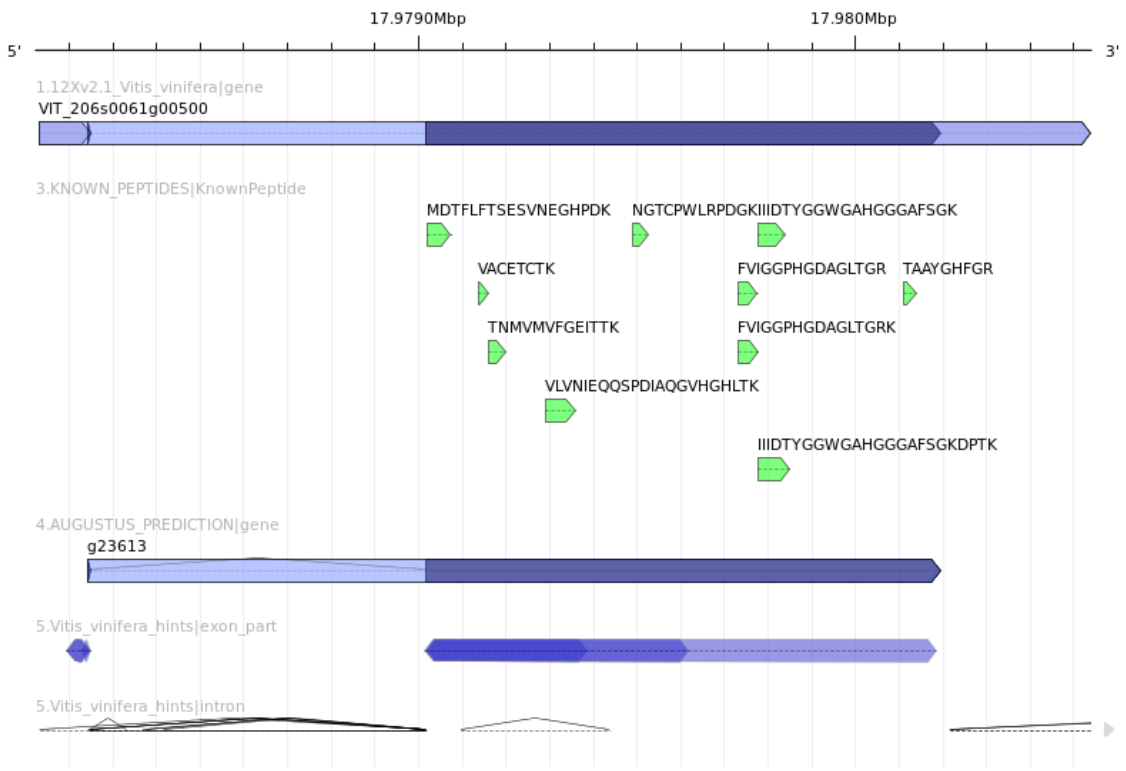


Figure 5.12 Known N-terminal acetylated peptide on chromosome 6

The known N-terminal acetylated peptide "MDTFLFTSESVNEGHPDK" resides 2 bp into the second exon, following a spliced region and the N-terminal end correlates with a known alternative translation initiation codon (GTG). However, Augustus did not predict any alternative protein-coding transcripts in agreement with it.

The third of these N-terminal acetylated peptides in conflict with the 12Xv2.1 annotation was shared peptide "FMIQTGDPLGDGTGGQSIWGREFEDEFHK", identified on chromosome 6, at positions 19,528,020 to 19,528,106 on gene VIT_206s0061g01590 with three protein isoforms. Also, the suggested alternative TIS site was not supported by the Augustus gene prediction (Figure 5.13, and with 1 supporting annotated MS/MS spectrum in Appendix Figure 5.16). The spectral E-value for this identification is ~9.0E-13, indicating a significant MS/MS spectral interpretation.

Performing a BLASTP search against the grape family in NR the protein-coding transcript 1 from gene VIT_206s0061g01590 was found to match unnamed protein product (CBI17058.3 with E-value = 0.0), transcript 2 was found to match peptidyl-prolyl cis-trans isomerase CYP71 (XP_010651713.1 with E-value = 0.0) on

chromosome 6, and transcript 3 was found to match a hypothetical protein (CAN66191.1 with E-value = 0.0). No other known peptides were found to map to gene VIT_206s0061g01590 and the N-terminal acetylated peptide was only identified from 1 PSM. The inferred translation initiation codon is GGC, however it does not appear to be a known alternative translation initiation codon in plants [581-584]. There was also no splicing acceptor site, intron region directly prior to the peptide, or intron region supported by the hints, only an AG motif beginning 4 bp upstream. There was only supporting exon evidence indicating further sites of translation upstream. Although the evidence in support of this possible alternative TIS site/protein isoform is lacking, it is possible that other non-AUG TIS sites not commonly known of could explain its presence. Therefore, more evidence is required, such as in the form of more supporting peptides mapped to the gene and preferably with at least 1 unique peptide identified from more than 1 PSM, to remove any ambiguity with the identification as well as any further EST, cDNA and RNA-seq evidence.

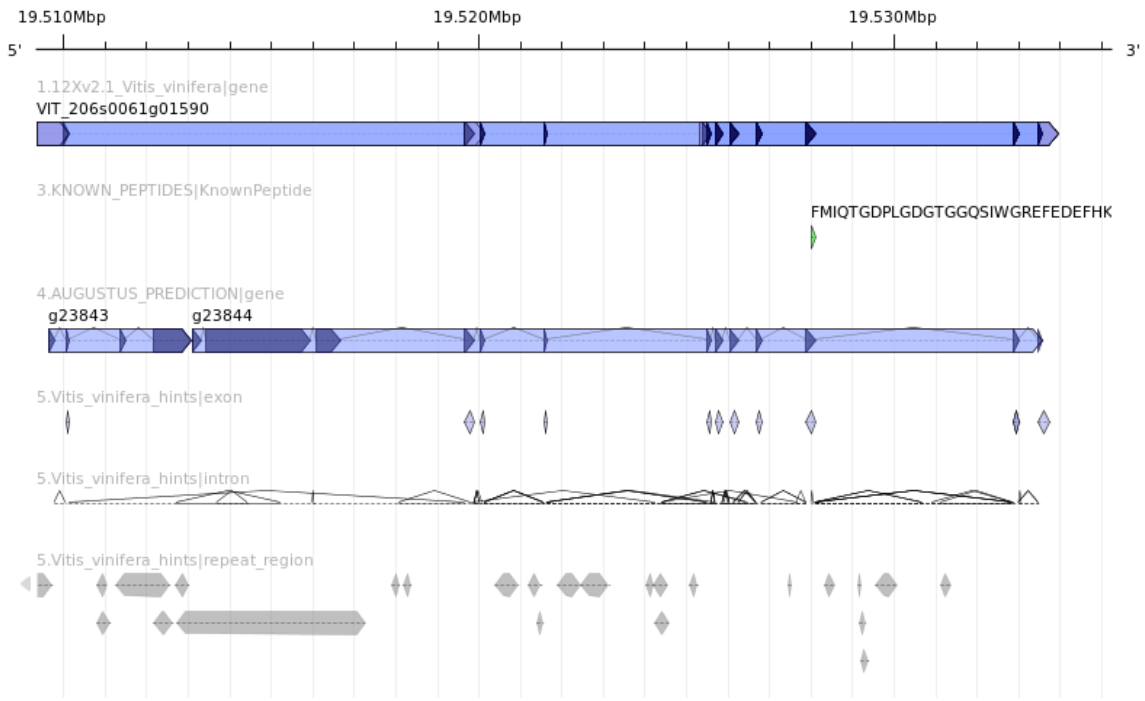


Figure 5.13 Known N-terminal acetylated peptide on chromosome 6

The known N-terminal acetylated peptide "FMIQTGDPLGDGTGGQSIWGREFEDEFHK" was located in the middle of one of the exons. However, there was no other evidence to indicate translation begins with this peptide, no known translation initiation codon or splice acceptor sites prior to the peptide, and Augustus did not predict any protein-coding transcripts in agreement.

The fourth of these N-terminal acetylated peptides in conflict with the 12Xv2.1 annotation was shared peptide "ANAASGMAVHDDCK", identified in three different locations: 1) on chromosome 6, at positions 5,236,863 to 5,236,904 on gene VIT_206s0004g04280, protein-coding transcript 2; 2) on chromosome 8, at positions 14,331,406 to 14,331,447 on gene VIT_208s0040g03360, protein-coding transcripts 3 and 4; and 3) on chromosome 13, at positions 2,558,362 to 2,558,403 on gene VIT_213s0019g00550, protein-coding transcripts 1, 2 and 3. However, only the peptide identified on chromosome 8 at positions 14,331,406 to 14,331,447 was in disagreement with the 12Xv2.1 annotations. Also, the suggested alternative TIS site was not supported by the Augustus gene prediction (Figure 5.14, and with a sample of 5 of 23 supporting annotated MS/MS spectra in Appendix Figure 5.17). The spectral E-values for this identification across the different locations range from 2.0E-12 to 5.0E-20, indicating significant MS/MS spectral interpretations. All mapped proteins share a

number of peptides, including the N-terminal acetylated peptide, but none were unique which made it impossible to determine if any one or all of them were being expressed due to the protein inference problem. In addition, the proteins shared extensive sequence homology, as indicated by Muscle alignment (data not shown).

Performing a BLASTP search against the grape family in NR, protein-coding transcript 2 for gene VIT_206s0004g04280 matched actin-depolymerizing factor 2 (XP_002284292.1 with E-value = 3.0E-100), on chromosome 6. The protein-coding transcripts 3 and 4 from gene VIT_208s0040g03360 both matched actin-depolymerizing factor 1-like (XP_002273958.2 with E-value = 6.0E-138 and 4.0E-99, respectively) on chromosome 8. Protein-coding transcripts 1, 2 and 3 from gene VIT_213s0019g00550 all matched actin-depolymerizing factor 2-like isoform X1 (XP_002284029.1 with E-value = 5.0E-100) on chromosome 13.

Preceding the N-terminal acetylated peptide on chromosome 8 at positions 14,331,406 to 14,331,447 the inferred initiation codon was CAG, but it does not appear to be a known alternative translation initiation codon in plants [581-584]. However, it is possible that other non-AUG TIS sites not commonly known of could explain its presence. The presence of the peptide could also be explained through the identification of AUG translation initiation codons found through splicing. There is a GT-AG or GC-AG (U2 spliceosome) donor-acceptor site within an already identified intron prior to the peptide, with the acceptor AG at position 14,331,404, and further upstream at position 13,330,577 for GC, and position 13,330,051 for GT. Although the evidence suggested that this could lead to an alternative initiation codon since the peptide is backed up by 23 PSMs, each with significant spectral E-values and potential donor acceptor sites for 5’ Methionine capping, the peptide could also simply be identified due to the expression of the other proteins making the N-terminal acetylated peptide the most abundant form. As any one protein could not be identified as being expressed over the others due to the

protein inference problem, further evidence is needed before claiming it as an alternative TIS site/protein isoform.

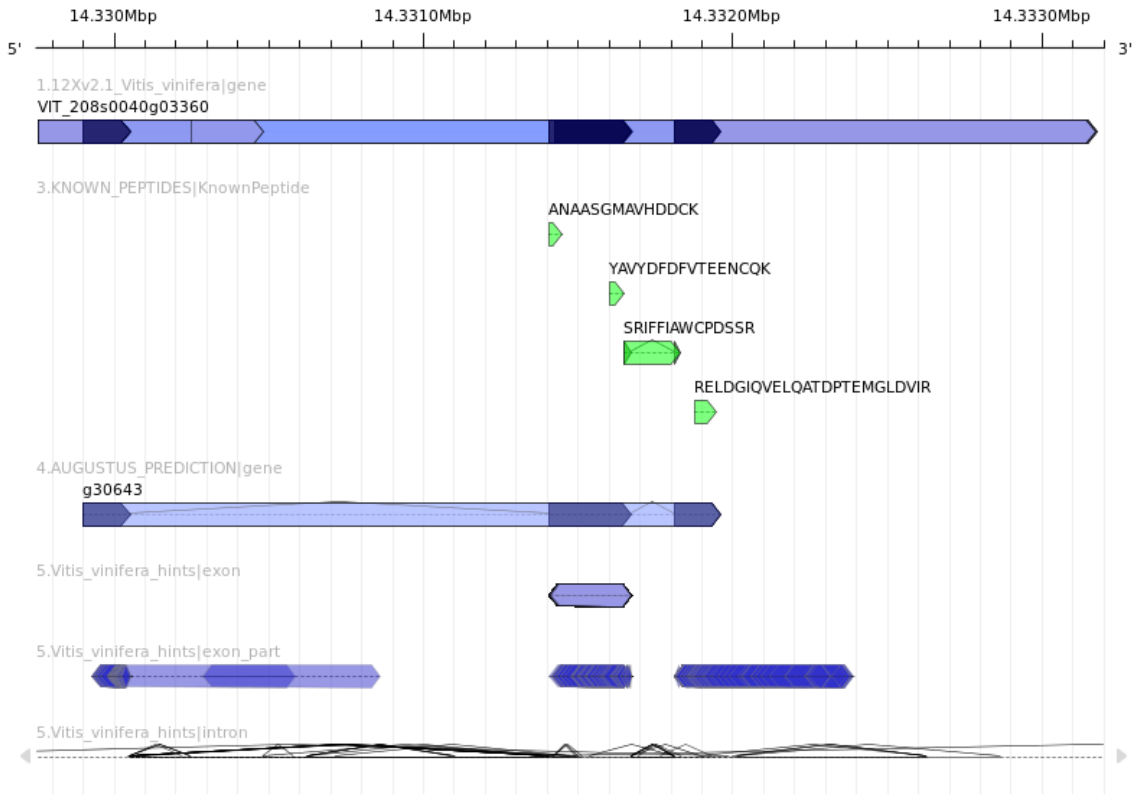


Figure 5.14 Known N-terminal acetylated peptide on chromosome 8

The known N-terminal acetylated peptide "ANAASGMAVHDDCK" was located at the start of the second CDS in 2 out of the 4 protein-coding transcripts. There were no known initiation codons prior to the peptide, however there was an AG acceptor site and two possible GT or GC donor sites further upstream. In addition, these peptides were shared with other similar proteins in the genome, which had annotations in agreement with the N-terminal acetylated peptide. Augustus did not predict any isoforms for this gene given the exon, exon_part and intron extrinsic hints. Taken together, this indicated that the inferred alternative isoform and TIS site for this gene was ambiguous, requiring further evidence.

The fifth and final N-terminal acetylated peptide in conflict with the 12Xv2.1 annotation was shared peptide "ALPNQQTVDYPSFK" identified in four different locations: 1) on chromosome 4, at positions 4,517,741 to 4,517,782, on gene VIT_204s0008g05020, protein-coding transcript 1; 2) on chromosome 4, at positions 4,522,707 to 4,522,748 on gene VIT_204s0008g05030, protein-coding transcript 1; 3) on chromosome 9, at positions 5,754,839 to 5,754,880 on gene VIT_209s0002g05940, with a single protein-coding transcript; 4) on chromosome 11, at positions 4,088,884 to 4,088,925 on gene VIT_211s0016g04780, with a single protein-coding transcript. Only

the peptide identified on chromosomes 9 and 11 at positions 5,754,839 to 5,754,880 and 4,088,884 to 4,088,925, respectively, were in disagreement with the 12Xv2.1 annotation. The suggested alternative TIS site was not supported by the Augustus gene predictions, showing only one isoform prediction for each gene region in agreement with the 12Xv2.1 annotation. An example of the mapped peptides for chromosome 9, including the N-terminal acetylated peptide, can be seen in Figure 5.15 (Supported with 2 annotated MS/MS spectra in Appendix Figure 5.18). The spectral E-values for this identification ranged from 6.0E-13 to 6.0E-14, indicating a significant MS/MS spectral interpretation. All mapped proteins shared a number of peptides, including the N-terminal acetylated peptide, but none were unique, making it impossible to determine if any one or all of them were being expressed due to the protein inference problem. The proteins additionally shared extensive sequence homology, as indicated by Muscle alignment (data not shown).

Performing a BLASTP search against the grape family in NR, protein-coding transcript 1 from gene VIT_204s0008g05020 found a match to GTP-binding nuclear protein Ran-3-like (XP_002284967.1 with E-value = 3.0E-167) on chromosome 4. Protein-coding transcript 1 from gene VIT_204s0008g05030 matched GTP-binding nuclear protein Ran-3-like (XP_002284971.1 with E-value = 2.0E-167) on chromosome 4. Protein-coding transcript 1 from gene VIT_209s0002g05940 matched GTP-binding nuclear protein Ran-3-like (XP_002285307.2 with E-value = 0.0) on chromosome 9, and protein-coding transcript 1 from gene VIT_211s0016g04780 matched GTP-binding nuclear protein Ran-3-like (XP_002285018.2 with E-value = 0.0) on chromosome 11.

Preceding the N-terminal acetylated peptide on chromosome 9 and 11 at positions 5,754,839 to 5,754,880 and 4,088,884 to 4,088,925, respectively, the inferred translation initiation codon in each case was TAG, located within the intron, but this is a stop codon not an initiation codon, and does not appear to be a known alternative

translation initiation codon in plants [581-584]. However, it is possible that other non-AUG TIS sites not commonly known of could explain its presence. The presence of the peptide could also be explained through the identification of an AUG translation initiation codon found through splicing. For chromosome 9 there was a GT-AG (U2 spliceosome) donor-acceptor site within an already identified intron, with the GT donor site at position 5,754,701 and the AG acceptor site at position 5,754,837. Similarly, on chromosome 11 there was a potential GT donor site at position 4,088,830 and at position 4,088,684. The evidence suggested that this could lead to an alternative initiation codon, with 2 PSMs, significant spectral E-values, and likely donor-acceptor sites for 5' Methionine capping. However, the peptide could also be derived from the other two proteins where the N-terminal acetylated peptide agreed with the annotation. Any one protein cannot be identified as being expressed over the others due to the protein inference problem, with no proteins identified from any unique peptides to pinpoint the most likely expressed protein candidate. Further evidence is therefore needed before claiming this as an alternative TIS site/protein isoform.

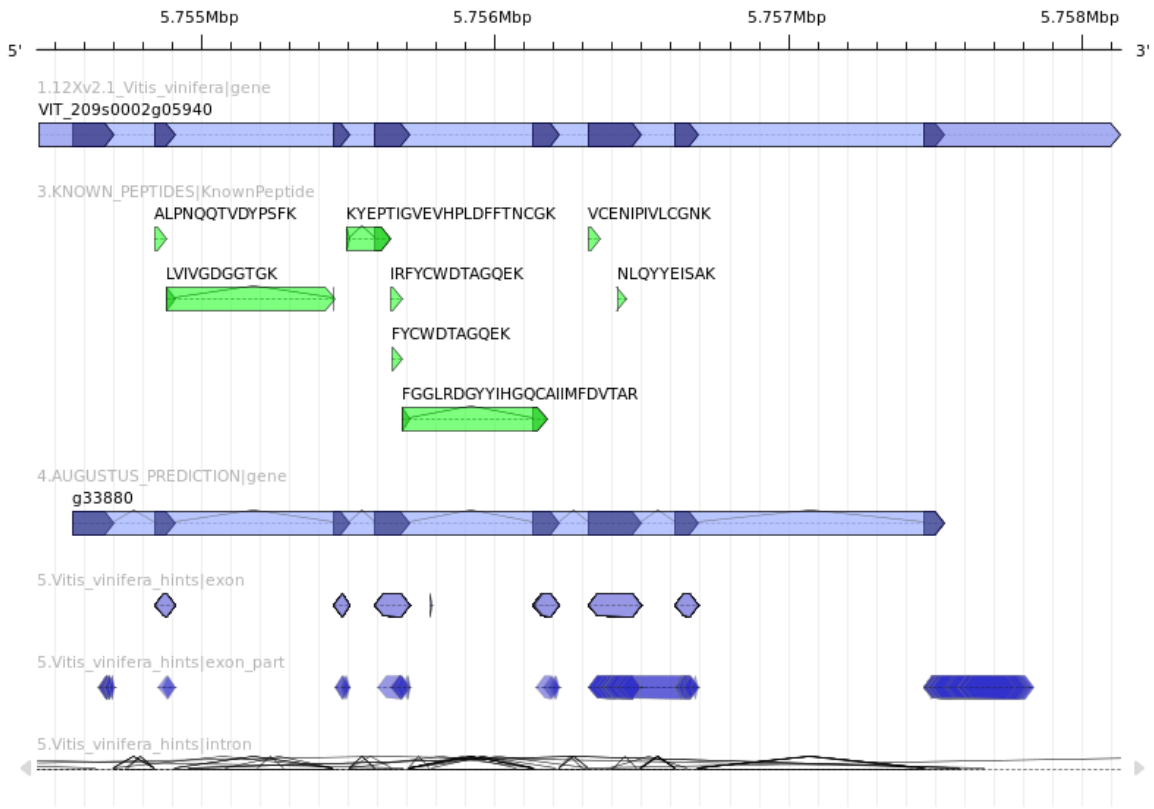


Figure 5.15 Known N-terminal acetylated peptide on chromosome 9

The known N-terminal acetylated peptide "ALPNQQTVDYPSFK" was located at the start of the second CDS. There appeared to be no known initiation codons prior to the peptide, however there was an AG acceptor site at position 5,754,837 and a GT donor site at position 5,754,701. In addition, these peptides were shared with other similar proteins in the genome that have annotations in agreement with the N-terminal acetylated peptide. Augustus did not predict any other protein isoforms for this gene, given the extrinsic evidence (exon, exon_part and intron hints). Taken together, this indicated that the inferred alternative TIS site/protein isoform for this gene was ambiguous and so required further evidence. The same was true for the corresponding peptide on chromosome 11, with similarly mapped peptides and prediction.

From the 5 shared N-terminal acetylated peptides that appeared to conflict with the 12Xv2.1 annotation, there was no conclusive evidence, except for peptides "ALPNQQTVDYPSFK" and "ANAASGMAVHDDCK", which could possibly be accounted for due to nearby splicing donor and acceptor sites, and peptide "MDTFLFTSESVNEGHPDK" which had a known alternative translation initiation codon at its N-terminal end. However, due to the protein inference problem and with no unique peptides identified to unambiguously identify the most likely protein the N-terminal acetylated peptides are derived from, further evidence would be needed. Interestingly, all 5 shared peptides lacked their 5' Methionine cap at their N-terminal most-end, which 1) indicates that N-terminal Methionine Excision (NME) [248] has

likely occurred and 2) that there may be other peptides with retained 5' Methionine caps which are unable to be identified through the current proteogenomics methodology.

Revisiting the analysis in the future, by generating further MS/MS spectral datasets selected only for peptides from the N-terminal end of proteins using N-terminomics [452], could improve the coverage at the N-terminal end. Particularly in combination with multiple replicates and proteases to reduce the ambiguity of which protein the N-terminal peptide is represented from, and which could identify non-AUG translation initiation codons as well as a variety of non-acetylated N-terminal peptides. To account for peptides with a retained 5' Methionine cap, an additional protein sample could be digested, followed by N-terminal peptide enrichment with an addition of Methionine aminopeptidase to cleave all N-terminal most Methionine residues *in vitro*. The peptides could then be mapped to their genomic locations and compared to the untreated sample. In addition, methods from a proteomics-only context could be applied to resolve the protein inference problem, and help with validating the N-terminal end of the known proteins, such as those outlined in Section 2.3.7, listed in Table 2.6 and further discussed in Section 2.3.9.

The use of N-terminomics and how best to resolve the protein inference problem in proteogenomics remains an open problem. At this time only complete coverage through strategies such as top-down proteomics can possibly resolve the ambiguities from the protein inference problem to confidently identifying the N-terminal end of putative proteins in a proteogenomics analysis.

5.3.14 Impact of search space

As highlighted in Section 2.4.2 the search space can have an impact on the sensitivity of a proteogenomics search, which was more pronounced in this study due to the larger search space of the grape genome. To improve the sensitivity of identifications two

different searches using 2.0 Da and 3.0 Da precursor mass tolerances were used, however the impact of the inflated search space was still evident using the combined FDR strategy. A total of 2,773 out of 55,373 proteins from the 12Xv2.1 annotation were identified during MS/MS database searches using MS-GF+, while the total number of mapped proteins was 7,536. Of these, 1,117 high confidence proteins had ≥2 peptides with 1 unique peptide (Table 5.1). When the same search was conducted against only the 12Xv2.1 protein predictions, combining the raw results from the 2.0 Da and 3.0 Da database searches, a total of 5,795 proteins were identified. Comparisons between proteomics- and proteogenomics-only searches revealed a loss of 3,022 proteins out of 5,795 or a loss of 52%, which would also infer a significant loss to novel identifications. This was significantly higher than the 30% loss found in Section 4.3.13 or outlined in previous studies [81, 324, 439], due to the larger sized genome of *V. vinifera* of 487.1 Mbp, compared to *Arabidopsis thaliana* with a genome size of 135 Mbp or *Bradyrhizobium diazoefficiens* with a genome size of 9.1 Mbp. This trend in the loss of sensitivity would most likely continue to increase as the genome size increased. This subject is addressed later in Chapters 6 and 7, by taking a different approach to MS/MS database searching and applying different FDR filtering approaches to improve known and novel peptide identification rates and to better discriminate between true and false positives.

5.4 SUMMARY

This study highlighted a number of advantages, as well as a few caveats in the course of conducting proteogenomics analysis, and which has provided a good example of how to bring different legacy –omics datasets (e.g. genomics, proteomics and transcriptomics) together for the genomic annotation of *Vitis vinifera* (grape). The proteogenomics analysis identified 133 novel peptides contributing to 341 novel annotation events (103 exclusively), consisting of 5 frame-shifts, 37 translated UTRs, 16 exon boundaries, 1

novel splice, 9 novel exons, 160 gene boundaries (24 exclusively), 112 reverse strands (10 exclusively) and 1 novel gene event in a total of 216 genes (67 exclusively) and 326 proteins (101 exclusively).

Among these annotations, 110 novel peptides directly led to 57 predicted proteins via Augustus gene prediction. Through the identification of these annotation events a possible over-assembly of the genome was identified, putatively resulting from the incorporation of non-nuclear reads into the nuclear chromosome assemblies. In addition, a large proportion of the CDS phase throughout the 12Xv2.1 annotation were recognised as incorrect, and subsequently corrected for gene prediction. The methods employed in this study have identified improvements as well as gaps in the understanding of proteogenomics approaches. Specifically, the selection of multiple precursor mass tolerances for low-accuracy MS/MS spectra, with an aggregation of results, improved coverage, and that using the combined FDR strategy to conduct proteogenomics significantly reduced the sensitivity, particularly as the genome size increased. This finding indicates a need for a proteogenomics approach with more refined control on the search space and FDR filtering stage of analysis. This requirement could be achieved by segregating and reducing the search space to only necessary sequences for more sensitive identification of novel and known peptides and to provide better discrimination between true and false positives, as well as potentially reducing the post-processing overhead.

5.5 CONCLUSIONS

The present study was able to identify a significantly larger number of proteogenomics annotation events than previously reported in Chapman et al [8], by using an improved methodology with larger and more diverse datasets. However, the loss in sensitivity due to a proteogenomics search continued to be a problem, resulting in a loss of 52% of

known proteins when using a combined FDR strategy. The inclusion of a splice graph derived from a large RNA-seq dataset contributed to the exclusive identification of 15 peptides identified in multiple annotation events, however only 1 novel splicing event was identified, which was proven to be a false positive and was not incorporated into the Augustus gene predictions. As previously demonstrated in Chapter 4, clustering MS/MS spectra and selecting the most appropriate precursor mass tolerances proved effective for selecting efficient parameters for proteogenomics searches. Applying multiple (2.0 Da and 3.0 Da) precursor mass tolerances, and merging the results, proved effective at improving the identification rate of PSMs when the mass accuracy of the MS/MS spectra was low.

Annotation events in this study were screened by searching the identified novel peptides against NR, manually looking for identifications in known proteins located in the same genomic coordinates. Although this approach improved specificity and sensitivity, combined with the event probability, it also had a negative impact on throughput. Different strategies are needed when identifying outliers from the applied event probabilities to determine effective event probability thresholds, in the absence of a method to determine the annotation event FDR. These could take the form of improving throughput by using automated processes to interrogate unique peptide matches found from entries in NR. Or, for example, by increasing the stringencies of the match, such as only considering matches of unique peptides with 100% identity and coverage to highly curated entries in protein repositories such as RefSeq protein, and possibly removing the manual component to analysis entirely by automating checks in GenBank entries for genomic coordinates and orthogonal evidence. In addition, 23 accepted novel peptides could not be included into the predictions. This posed some questions: were there any novel peptides included which should not have been included? And how could peptides such as these be excluded from peptide clusters and

annotation events to prevent similar situations in the future? One possible solution could be to apply further filtering steps when selecting novel peptides for inclusion into peptide clusters and annotation events using additional sources of evidence. Scenarios of this nature should be considered in future versions of the Enosi tool.

Another caveat of the analysis was the use of a fixed peptide linkage distance that resulted in some annotation events being incorrectly categorized, as was previously illustrated as a problem in Section 4.3.4. This highlights a need to determine the peptide linkage distance for each peptide cluster dynamically. For example, by determining the likely distribution and size of genes in the local region and using machine-learning approaches, as well as assigning annotation events based on other additional evidence; such as known peptide evidence within the same gene being annotated or genes in close proximity, particularly those with the same frame located within close proximity. Any conflicts between the novel and known peptides from overlapping ORFs, could be resolved using parsimony of known versus novel peptides. This approach could remove much of the manual interpretation and validation of the annotation events required throughout analysis and would also improve the accuracy when assigning annotation event types, as well as when applying appropriate event probability thresholds, which could, as a result reduce the occurrence of false positive annotation events and predictions.

In addition, the Enosi tool did not automatically cater for some annotation events, such as over-predicted genes. These types of annotation events could be identified by N-terminal acetylated peptides located within the gene region not in agreement with the TIS site. This is difficult to interpret, primarily due to the protein inference problem, however the identification of proteins in the genome containing a unique N-terminal acetylated peptide could be used to unambiguously confirm the presence of such peptides to correct the annotation, given that the unique status of the

N-terminal peptide was accurate. Conversely, the identification of these N-terminal acetylated peptides could be used to validate the start sites of already known proteins, as was demonstrated in this study. However, it was also found that no TIS sites in the known proteins could be confidently identified and so a time-consuming manual approach was required. An approach of this type could feasibly be automated to improve throughput and could be integrated into the Enosi tool, potentially by identifying only unique N-terminal acetylated peptides to avoid ambiguity. A further means to resolve the identification of N-terminal acetylated peptides could be to perform a proteomics-only analysis and resolve the protein inference problem using approaches and tools like those listed in Table 2.6. Another alternative method could be to enrich for N-terminal peptides employed in methods such as N-terminomics [452].

A problem which became apparent during the analysis was the post-processing of large sets of results through the combined FDR strategy, requiring the merging of all results into 4 separate TSV files prior to FDR filtering to allow processing within a practical time-frame. However this would have negatively impacted on the accuracies of the applied FDR threshold and calculated local FDRs for each PSM, and consequently the event probabilities. In future studies, ways to limit this impact could be applied, such as improvements to the efficiency of the algorithm for FDR filtering and reducing the overall MS/MS spectral dataset and database size prior to FDR filtering. The benefits of such an approach would lead to improved sensitivity of the database searches and, at the same time reduce the overhead for processing the MS/MS spectral datasets and databases, thus enabling the results to be processed together and reducing or removing the negative impact of FDR and event probability accuracy.

This study identified a possible over assembly with the 12X genome assembly. Some clues indicating that this was the case was with the large number of the unassigned chromosome sequences (ChrUn), fragmented chromosomes and the

identification of a number of reference proteins, Augustus gene predictions and novel peptides finding significant matches to chloroplast and mitochondrial proteins in NR. One convincing example was with a reference protein on chromosome 7, with both the original reference protein and the Augustus gene prediction matching significantly to ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (RuBisCo large subunit), a highly abundant plant protein predominantly confined to the chloroplast genome. This prediction contained a high number of PSMs to both unique/shared novel and known peptides and highlighted two problems: 1) the genome was most likely over-assembled due to the presence of contaminant non-nuclear genomic reads, which would have hampered proper assembly, and indicating that contaminant reads should have been filtered out prior to assembly (or optimally sequenced after chromosome sorting); and 2) the proteomics data, which unavoidably was legacy in nature should have been filtered by cell component and/or tissue fractionation prior to MS/MS analysis in order to improve the depth and breadth of coverage and also reduce contamination from non-nuclear derived proteins which potentially contributed to false positive identifications.

Apart from the issues posed by the possible over-assembly there were many revised annotations identified, including corrections to incorrect CDS phases that conflicted with the appropriate phase inferred from the reference protein prediction. This indicated a possible flaw in the method employed for the original reference annotation [7].

The proteogenomics annotation of grape should be re-visited in the future with a much larger MS/MS spectral dataset, preferably obtained from a higher accuracy mass spectrometer, such as an LTQ Orbitrap or QTOF, and with sufficient sampling depth and breadth of coverage, that tighter stringencies on MS/MS spectral quality thresholds could be afforded with little or no negative impact on overall peptide coverage. Such an approach was not feasible in the present study due to the relatively limited MS/MS

spectral dataset, which necessitated more careful filtering to prevent significant losses. The addition of proteomics data derived from Pinot Noir instead of Cabernet Sauvignon would keep the analysis consistent between the proteome and genome. Further RNA-seq data could also be utilised, spanning other cultivars, for sequence variant identification, as previously undertaken [474]. Studies of this type would prove even more worthwhile once the question of genomic coverage and suspected over-assembly have been addressed, to reduce any spurious identifications. In the meantime, the assembly and annotation would benefit greatly from a review and assessment using tools such as BUSCO [588] to identify and resolve any caveats moving forward with improving the assembly and annotation.

5.6 ACKNOWLEDGEMENTS

Chapter 5 is in preparation for publication along with new grape gene and protein predictions, which will be submitted to NCBI. The dissertation author is the primary author of this paper and designed the proteogenomics workflow, ran the analysis and wrote the paper. The dissertation author would like to thank Steven Van Sluyter for providing the MS/MS spectral dataset and Ryan Ghan for providing the large RNA-seq dataset across multiple cultivars. The dissertation author would also like to thank the Centre for Comparative Genomics for their compute resources and guidance and the Pawsey Supercomputing Centre for the use of their compute resources, which were supported by funding from the Australian Government and the Government of Western Australia.

6 HUMAN PROTEOGENOMICS

6.1 INTRODUCTION

The human genome has been the focus of intense study since it was sequenced in 2001 [35], followed by the final accepted complete draft published in 2004 [49]. In more recent years several large scale studies have revealed more about the human genome, with one of these studies being the ENCyclopedia Of DNA Elements (ENCODE) project, which redefined genes and their transcription [589], leading to additional studies to define the coding, non-coding transcripts, transcriptional regulatory regions and more [30, 590]. A sub-project of ENCODE, called GENCODE was tasked with manually annotating the genome to identify the protein-coding transcripts. However, these transcripts were identified from indirect sources such as ESTs, and protein sequences available in repositories such as UniProtKB/SwissProt, which, as outlined in Section 2.4, only consists of 5% direct protein sequences. In addition, more than 50% of the transcripts that were identified have no protein-coding potential [590], and many could not be identified as non-coding RNAs either, leaving their interpretation ambiguous.

A recent proteogenomics study which sheds some light on the hidden protein-coding regions of the human genome produced proteomics data from a number of different human cell lines, including the ENCODE Tier 1 cell lines K562 and GM12878 [11], and later followed-up with a focus on GM12878, looking at the maternal and paternal genomic sequences from the diploid genome NA12878 [449], with both studies using the proteogenomics tool Peppy [591].

6.1.1 Outline of this study

The present study improves upon that of previous work [11, 449], increasing sensitivity and identifying previously unidentified refinements to the current annotations, and highlights that there is still much room for improvement to highly curated data sets like

the human genome. The present study utilizes the MS/MS spectra generated from the study in [11], from cell line GM12878, the human genome sequence Hg19 (GRCh37 patch 13), GENCODE v19 reference annotation and the ENCODE spliced alignment results for GM12878, to generate a splice graph database. The latest proteogenomics tool Enosi [82] was used to perform the proteogenomics analysis.

As outlined in Section 2.4.2 the search space during a proteogenomics search can be a problem, particularly with larger genomes, and as demonstrated in Chapter 5, even with a 487.1 Mbp sized genome such as with *Vitis vinifera* there can be a loss of up to 52% when looking at the identified known proteins. In the present study with a 3.3 Gbp human genome the loss would be significantly more pronounced. This issue is addressed using a two-pass search approach, combined with an improved two-stage false discovery rate (FDR) strategy over the more conservative strategy used in [474]. This enhanced methodology increases the sensitivity of the search, by optimising the search space and applying FDR filtering to discrete known and novel search spaces separately to improve the separation between true and false positives.

6.2 MATERIALS AND METHODS

6.2.1 Proteomics and genomics datasets

The human genome assembly (Hg19 (GRCh37 patch 13)) and GENCODE v19 human genome annotation, which included the protein predictions and GFF file, were obtained from GENCODE, downloaded from the web site (http://www.GENCODEgenes.org/releases/19.html) (Appendix File 6.1).

The MS/MS spectra used in the previous studies [11, 449] were generated from ENCODE tier 1 cell line GM12878, which is a lymphoblastoid cell line immortalized using Epstein Barr Virus (EBV). Proteins from the GM12878 cell line went through a series of processes before MS/MS analysis. The proteins first went through subcellular

fractionation, GELFREE fractionation [592], filter-aided sample preparation (FASP) [593] and finally microwave-assisted tryptic digestion [594] prior to analysis on an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific).

A total of 1,054,278 MS/MS spectra from cell line GM12878 were downloaded from The Giddings Lab at Boise State University web site (http://giddingslab.org/data/encode/proteome-commons), which in turn were generated from different cell component fractions, as outlined in [11]. The cell component fractions included cytosol, membrane, nuclear, mitochondria, and a whole cell lysate. The cytosol fraction was of poor quality as determined in [11], and was removed from that study. However, in the present study all cell component fractions were pooled, including cytosol, clustered and quality filtered to improve the overall quality of all MS/MS spectra. The benefit of using MS/MS spectra pooled from multiple cell component fractions is that it allows for the detection of both low and high abundant proteins, improves on the proteome coverage and could potentially increase the novel annotation event identification rate.

As outlined in Section 3.1.1, a source of contaminants was appended to the protein sequence predictions from GENCODE v19 before being used in the MS/MS database search to identify any contamination.

6.2.2 RNA-seq datasets

A source of RNA-seq data to generate a splice graph for proteogenomics analysis was used. Instead of obtaining reads from the Sequence Read Archive (SRA) and running alignments, alignment results in the form of BAM files from the ENCODE project itself were employed. BAM files in the ENCODE project were generated using the spliced alignment tool, STAR [127]. The BAM files were obtained from the ENCODE project, through NCBI GEO accession GSE30567 and downloaded directly from the University

of California Santa Cruz (UCSC) Genome Bioinformatics web site (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wGENCODECshlLongRnaSeq).

6.2.3 MS/MS database searching

The MS/MS database search was performed by MS-GF+, as outlined in Section 3.3. In this case study trypsin was used as the protease, the instrument was set to low-res LTQ (Ion Trap), and the precursor mass tolerance used was 9.0 ppm, as determined from a preliminary MS/MS spectral dataset assessment, detailed below in Section 6.2.4.

6.2.4 Dataset processing

The GENCODE v19 protein sequence FASTA file and GFF file required formatting into a compatible format for proteogenomics analysis, as outlined in Section 3.4.1.

The total of 1,054,278 MS/MS spectra obtained for this study were first assessed by searching against the known proteome, examining the effects of using MS-Cluster to cluster the MS/MS spectra, PepNovo to quality filter the MS/MS spectra, and with an assessment of optimal precursor mass tolerances, as outlined in Section 3.4.6. Since all the MS/MS spectra were of high-accuracy, derived from an LTQ Orbitrap Velos mass spectrometer, this factor needed to be reflected in the search parameters. Therefore, an assessment of the optimal precursor mass tolerance was needed using a range of tolerances, 0.5, 1.0, up to 10.0 in 1.0 ppm increments, and then up to 100.0 ppm in 5.0 ppm increments (Appendix File 6.2).

All BAM files from the ENCODE project for cell line GM12878 were merged and used to generate a splice graph FASTA database. A six-frame translation of the genome was also generated. The steps involved in the generation of both splice graph and six-frame translation are outlined in Sections 3.4.4 and 3.4.5, respectively.

6.2.5 Proteogenomics pipeline

The proteogenomics pipeline was used as outlined in Section 3.5, with an included two-pass search approach, combined with an improved two-stage FDR strategy, which reduced the database size and applied discrete 1% peptide-spectrum match (PSM) FDRs to the separate known and novel search spaces. This greatly improved the discrimination between true and false positives and thus improved on the number of known and novel PSMs identified. After a preliminary assessment of the MS/MS spectra, as outlined previously in Section 6.2.4, the choice of clustering the MS/MS spectra and quality filtering to a PepNovo score threshold of 0.01 was decided, resulting in 613,432 clustered MS/MS spectra, and it was determined that the optimal precursor mass tolerance was 9.0 ppm.

The total 613,432 MS/MS spectra were then split into separate MGF files of 20,000 MS/MS spectra each, using an in-house MGF splitting tool, before running each MS/MS spectral file through MS-GF+ on a cluster against the known proteome, six-frame translation and splice graph, using the two-pass search approach and then the improved two-stage FDR strategy as outlined in Section 3.5.1 and 3.5.2, respectively.

The improved two-stage FDR strategy was applied in this study, as outlined in Section 3.5.2 and illustrated in Figure 3.1, to see how it could improve on the identification rate in comparison to the more conservative strategy in [474] and the more traditional combined FDR strategy. All three methods were compared in this study. For the conservative two-stage FDR strategy all MS/MS spectra identified as matching the known proteome and protein contaminants were removed. The remaining MS/MS spectra, now considered ‘novel’, were then searched against the proteogenomics search space, after which the now novel PSMs had a 1% PSM FDR applied. With the combined FDR strategy all MS/MS spectra identifying PSMs in both the known proteome and proteogenomics search space were used together and all

results were merged across 2 TSV files due to the processing limits on large result files (See Section 3.5), before a 1% PSM FDR was applied to both resulting TSV files (Appendix File 6.3).

The choice of parameters for the proteogenomics pipeline, as outlined in Section 3.5, included a minimum event probability for novel genes, distal events and proximal events of 90%, a peptide linkage distance of 150,000 bp representing ≥95% of gene sizes in GENCODE v19, a minimum cluster size of 1, and a minimum of 1 unique peptide per cluster.

The annotation events were further screened, as mentioned in Section 3.5, however a number of differences were incorporated. In Chapter 5, many of the accepted annotation events were heavily biased on sequence homology results, and manual visual inspection possibly led to higher false positive rates due to human error. In the present study novel genes and distal events were first screened with ≥2 unique peptides and/or ≥99.9% event probability and proximal events were filtered with an event probability of ≥99.8%. Any outliers, such as those derived from single and unique peptide annotation events were identified through BLASTP searches against the human curated protein repository NCBI RefSeq protein with support from known mapped peptides to the genes annotated, particularly for proximal events where the known peptides mapped to the same ORF as the novel peptide and spectral counts were also often an indication of a likely real annotation event that showed correlation with the event probability. In this study, to improve throughput of screening the outlier single unique peptide annotation events, the results from sequence homology searches to human proteins in RefSeq protein entailed acceptance of matches with 100% query coverage and a maximum of 2 mismatches with an E-value of at least 1E-03.

The investigation of a number of annotation events in Section 6.3 uses BLASTP searches against NCBI NR to broaden the range of evidence supporting the annotation during the discussion. Venny (http://bioinfogp.cnb.csic.es/tools/venny/) was used within this chapter to illustrate Venn diagrams of identified peptides across different methods.

6.2.6 Improving gene predictions

Once the novel annotations were filtered and reviewed the gene prediction tool Augustus [102] was used to improve the overall gene models of GENCODE v19. No training of Augustus was carried out in the present study, with only the available human gene model used from Augustus version 3.02. The novel annotations and GENCODE v19 annotations were used as hints for Augustus gene prediction, along with repeat region hints generated from the genome using RepeatMasker [576]. The Augustus gene prediction tool was then run using parameters as outlined in Section 3.5.3, except that only the novel peptides, GENCODE v19 gene models and repeats were used as hints.

6.3 RESULTS AND DISCUSSION

The present study outlined improvements to the GENCODE v19 annotation of *Homo sapiens*, demonstrating the benefits that proteogenomics presents by integrating the different –omics platforms used from the ENCODE project to contribute to human genome annotation and identifying 617 (126 exclusively) novel annotation events. The study also improved on the proteogenomics findings from the ENCODE project, by using a two-pass search approach with improved two-stage FDR strategy, and which was directly compared to previous methods: the combined FDR and ‘conservative’ two-stage FDR strategy. The improved methodology is particularly useful for larger genomes that are accompanied by a reduced sensitivity and a higher false positive rate due to their larger proteogenomics search space.

6.3.1 Evaluation of pre-processing MS/MS spectra

Prior to running the proteogenomics pipeline, the MS/MS spectra were evaluated for the optimal pre-processing strategy and precursor mass tolerance (Appendix File 6.2). All 1,054,278 MS/MS spectra were clustered by a factor of 1.54. It was found that clustering reduced the peptide FDR after an initial 1% PSM FDR filtering from 4.5% peptide FDR to 2.4% peptide FDR and reduced the protein FDR from 25.7% to 13.9% (Appendix Figure 6.1A-C). As can be seen in Appendix Figure 6.1A, the number of total MS/MS spectra lost after quality filtering can range from 3.2% at the lowest end to 67.8% at the most stringent cut-off. Applying scores between 0.05 – 0.1, recommended by PepNovo (detailed in the Help File bundled with the tool), resulted in around 32.5% to 49.6% of the MS/MS spectra being lost, however at a score of 0.01 only 10.3% of MS/MS spectra were lost while maintaining a peptide FDR of ~2%, as well as retaining many PSMs and unique peptides that were lost when using a higher score than 0.01 (Appendix Figure 6.1D-E). With higher losses in MS/MS spectra there was a reduced number of PSMs and unique peptides reported, and there was only negligible improvement in peptide FDR. This observation indicated that the dataset was not improving but was losing valuable MS/MS spectra, which could be used to identify novel proteogenomics annotations.

The results indicated that a clustered MS/MS spectral dataset with a PepNovo quality filtering score cut-off of 0.01 was most suitable for further proteogenomics analysis, as it resulted in the best balance between keeping the maximum number of PSMs and filtering out poor quality MS/MS spectra.

6.3.2 MS/MS database search parameter optimization

As outlined in Section 3.4.6 high-accuracy MS/MS spectra are well suited to precursor mass tolerance optimization since tighter tolerances often improve the sensitivity of PSM identification. This observation also held true for the present study as high-

accuracy MS/MS spectra were obtained, generated from an LTQ Orbitrap mass spectrometer.

A clustered set of MS/MS spectra, quality filtered to a PepNovo score of 0.01, was used to assess the precursor mass tolerances over a range, as outlined in Section 6.2.4 (Appendix File 6.2). It was determined from this analysis that 9.0 ppm was the optimal precursor mass tolerance to use (Appendix Figure 6.2). After ≤1% PSM FDR filtering the maximum number of PSMs obtainable was 97,930 at 9.0 ppm, while the peptide FDR was 2.13%.

6.3.3 Proteogenomics pipeline

A proteogenomics pipeline was customised using Enosi with MS-GF+, as outlined in Section 3.5 and illustrated in Figure 3.1. The pipeline included a two-pass search approach with improved two-stage FDR strategy by applying FDR filtering to the known and novel search spaces separately. This step reduced the search space and improved sensitivity and discrimination between true and false positives.

The two-pass search approach with improved two-stage FDR strategy reduced the overall database size, and as a result the TSV file sizes were significantly reduced in size. In addition, the TSV files were split between the known identifications and the novel identifications, with a 1% PSM FDR applied to each. This change improved the accuracy of the FDR, since the results from the known search space and novel search space were aggregated in a separate TSV file for known and novel results as they have widely different false positive rates.

The choice of key variables for the proteogenomics pipeline were chosen as they have been in preceding chapters. The peptide linkage distance was chosen based on the size of ≥95% of genes in the current genome annotation, which was found to be 150,000 bp. As previously discussed in Section 5.3.6, the use of a fixed peptide linkage distance

across the entire genome brings with it the caveat that it inflates the number of gene boundary and reverse strand events, thus requiring manual validation and negatively impacting throughput, and which could lead to ambiguity with overlapping annotation event types. This ambiguity could also in turn lead to an increased false positive rate when applying event probability thresholds across the different annotation event types. This possibility becomes more of a problem in human, as the peptide linkage distance used is the largest used so far within this thesis. Therefore, how the peptide linkage distance is defined in Enosi needs to be addressed in future versions of the tool. Until a solution is reached manual screening and interpretation of the annotation events and also by exclusively assigning peptide clusters to annotation events to negate the impact of large peptide linkage distances is necessary to come to a closer value for the true number of different annotation event types, as well as the number of genes and proteins with inferred annotations.

Contamination was not an issue in this study, compared to Chapter 5 and the grape genome, where there may have been over-assembly of the genome, with chloroplast and mitochondrial reads inadvertently incorporated into the assembly. In this study there were no such findings, as would be expected from the human genome, which has benefitted from over a decade of sequencing and rigorous manual curation on an international level. Additionally, the mitochondrial genome was available in GENCODE v19, allowing for a known proteome target for any mitochondrial peptides that might have inadvertently been identified as novel within nuclear chromosomes by chance in its absence. Also, with the peptide mass spectrometry data being of higher accuracy, more specific matches with less error was possible. Additionally, the data were generated with proteogenomics in mind and cellular component fractionation carried out, thus improving coverage and allowing for the identification of some

proteins, which could have otherwise been masked by more abundant proteins from other cellular compartments.

Further screening of novel events was performed by accepting annotation events with high parsimony of unique peptides and/or more stringent event probabilities, with any outliers of single unique peptides identified by searching the unique peptides against well-curated human RefSeq protein database. This step provided a streamlined, faster and more confident approach, compared to screening each annotation event by manually checking against a combination of NR, RefSeq protein and SwissProt databases that may have resulted in human error. Although the matches were not rigorously validated against orthogonal evidence and genomic locations in GenBank, the majority of identifications via this method were also found with higher event probabilities and parsimony, including those annotation events already accepted at higher stringencies. To highlight that the applied method was a valid approach, several random annotation events were confirmed manually by checking the GenBank record to confirm that the genomic coordinates were often the same, in agreement with the proteogenomics analysis (Appendix File 6.3). This proved a viable approach in the present study to improve throughput and apply some control on FDR, although there was no way to determine how effective it was at the annotation event level as the tools to determine the annotation event FDR were not available. However, a more rigorous approach could be undertaken by automatically downloading and then screening the GenBank records after each BLASTP search to check for orthogonal evidence and genomic coordinates to back up the claims from each of the single unique peptide annotation events. The final minimum event probabilities after annotation event filtering were 99.986% for novel genes, 95.616% for distal events, and 98.237% for proximal events.

Similar but more powerful methods that could improve the rate of annotation event identification and validation could be achieved by spectral library searching to validate the identified novel MS/MS spectra, by searching against an already curated spectral library using tools such as Tremolo [397]. Which would have much more sensitive and specific matches than performing BLASTP searches, or as was previously suggested using spectral archives [398]. The spectral archives approach proposes a public repository of clustered MS/MS spectra that gradually improves as the public add further MS/MS spectra. This resource could then be used to validate annotation events, as well as identify truly novel annotation events when an annotation event finds no matching or weakly matching MS/MS spectra. These approaches use the MS/MS spectra to identify spectrum-spectrum matches (SSMs), which have much higher specificity compared to PSMs used in conventional database searches against sequences, which often lead to ambiguous identifications depending on the contents of the sequence database being searched (e.g. misidentifying variant sequences as containing post-translational modifications (PTMs)).

As was identified in previous chapters, a source of false positives can be the actual MS/MS spectra used in the analysis. For this analysis MS/MS spectra were derived from tier 1 cell line GM12878 that was immortalized using EBV. Genetic differences between GM12878 and GENCODE v19 may lead to numerous variant peptides interpreted from the MS/MS spectra which could further be misinterpreted as containing PTMs during the proteogenomics analysis, or lead to the identification of novel peptides in locations where none exist. A viable way to account for these variant peptide sequences, which may inadvertently lead to false positive identifications, is to perform variant call analysis using the same ENCODE alignment results, as previously demonstrated [474], to generate variant peptides that could contribute to the splice graph. This step could also produce some interesting insights into the genetic changes

occurring in GM12878 in becoming an immortalized cell line. However, the approach only became known late in the study and was not pursued, but could be re-visited in future studies.

6.3.4 Proteogenomics analysis

Using the two-pass search approach with improved two-stage FDR strategy, a total of 593 novel peptides were identified with an event probability of ≥90%, at least 1 unique peptide and a minimum cluster size of 1. This is in contrast to the results obtained using the combined FDR and conservative two-stage FDR strategies, which resulted in 240 and 8 novel peptides being identified, respectively. After filtering the novel annotation events and applying the same thresholds across all methods, a total of 77, 46 and 5 novel peptides remained for the two-pass search approach with improved two-stage FDR, combined FDR and conservative two-stage FDR strategies, respectively (Figure 6.1, Table 6.1 and Appendix Files 6.3 and 6.4). Compared to the combined FDR and the conservative two-stage FDR strategies, the two-pass search approach with improved two-stage FDR strategy was able to identify 35 more novel peptides, while the combined FDR strategy identified 4 novel peptides that neither of the other strategies were able to identify. Due to the lower FDR accuracy of identifications with the combined FDR strategy, these 4 novel peptides may have been false positives, incorrectly identified from 'known' MS/MS spectra or fallen just below the applied thresholds within the other two strategies. By comparison, the conservative two-stage FDR strategy only identified 5 novel peptides all of which were identified by both of the other strategies. These results indicate a large improvement in the rate of novel peptide identifications by restricting to separate known and novel search spaces using the two-pass search approach with improved two-stage FDR strategy.

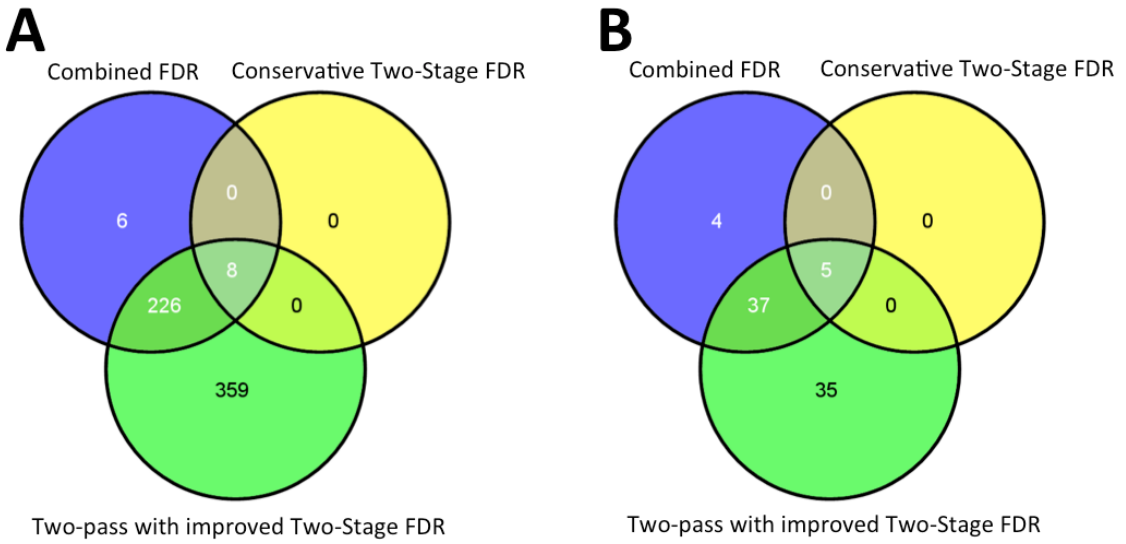


Figure 6.1 Comparison of identified novel peptides between three methods

A comparison between the numbers of identified novel peptides ≥90% event probability, 1 unique peptide per cluster and a minimum cluster size of 1, with (A) an unfiltered set with no screening of annotation events, and (B) after screening of annotation events using the same thresholds for all methods.

A side-benefit of reducing the search space and applying FDR filtering on the known and novel search spaces separately was a reduced final merged TSV file size for each of the search spaces. Previously, in the combined FDR and conservative two-stage FDR strategy the final merged TSV files were much larger when using a large MS/MS spectral dataset and genome, requiring that they be split into two to four files to be able to process the files in a practical time-frame, which resulted in inaccurate PSM and local FDRs due to PSMs being distributed across the files. The two-pass search approach with improved two-stage FDR strategy reduced the size of the TSV files resulting in more efficient and faster processing time with more accurate FDR calculations. As pointed out in Section 5.4, a way to resolve this problem was needed. This method did resolve the problem with the added benefit of increasing the PSM identification rate for both known and novel PSMs. However, the problem may still arise if the number of MS/MS spectra and the genome is many times larger than used in this study.

As was applied in the previous Chapter 5, Section 5.3.4, the exclusive number of annotation events for each peptide cluster, and their associated genes and proteins, were

also determined to account for the inflated number of gene boundary and reverse strand events, as shown in parenthesis in Table 6.1. This was more problematic in this study due to the much larger peptide linkage distance of 150,000 bp that further inflated the number of annotation events identified.

The 77 novel peptides identified from the two-pass search approach with improved two-stage FDR strategy led to 617 annotation events (126 exclusively) amongst 147 genes (29 exclusively) and 609 proteins (116 exclusively) from the GENCODE v19 annotation. The novel annotations along with GENCODE v19 reference annotations were then used as hints for Augustus gene prediction. A total of 29,266 genes and 32,781 proteins (≥66 aa in length) were predicted (Appendix File 6.5), and of these, 52 predicted proteins had 66 novel peptides incorporated (Table 6.1), of which 48 novel peptides were unique and identified in 37 of the 52 predicted proteins (Appendix File 6.6). The number of protein-coding genes predicted by Augustus was much higher than the original reference GENCODE v19 predictions (Table 6.1). In addition to the new predictions previously not identified, the high number of predictions could also be attributed to two factors, as was outlined in Section 3.5.3.

As previously discussed in Section 5.3.4 the reason why many novel peptides, in this case only 11, were not included in the predictions and by extension why peptides possibly included should not have been, needs to be addressed in future versions of the Enosi tool, to better screen novel peptides before inclusion into peptide clusters and annotation events.

By comparison with the combined FDR and conservative two-stage FDR strategies, the two-pass search approach with improved two-stage FDR strategy directly improved the quality of peptide clusters by including more novel peptides, and improved the accuracy of the event probabilities, allowing some annotation events to

now pass the applied event probability thresholds and identify more annotation events as a result. For example, compared to the combined FDR strategy, the number of novel gene annotations with ≥2 unique peptides increased from 4 to 5, and compared to the conservative two-stage FDR strategy, improved from 0 to 5 novel genes. Similarly, the number of proximal events, such as exon boundaries, went from 3 (1 peptide cluster) with the combined FDR strategy, 0 with the conservative two-stage FDR strategy to 27 (7 peptide clusters) using the two-pass search approach with improved two-stage FDR strategy, when applying the same stringent thresholds across all methods.

However, it is worth keeping in mind, the conservative two-stage FDR strategy, although limited in sensitivity, is highly specific. It was warranted in the study in which it was initially demonstrated, i.e. examining very large datasets of numerous variant peptides from cancer, in which the 'known' MS/MS spectra likely containing PTMs could have misidentified novel variants in the splice graph.

To further enhance the improved two-stage FDR strategy to provide discrimination between novel and known identifications, the spectral E-value could be applied. This step could be achieved by comparing the spectral E-values between matches to the known sequences and novel peptide sequences, and accepting only the MS/MS spectra with better spectral interpretation while discarding matches with ambiguous equal spectral E-values, for example, between known protein sequences with PTMs versus the variant peptides in the splice graph. Such an approach would provide another level of discrimination without compromising on sensitivity, and it would not be influenced by the database as the spectral E-value is determined independent of the database.

The number of novel peptides incorporated into the predictions was reasonably high at 66 novel peptides (86%), which was relatively higher, by 3%, than the

proportion included in the previous study reported in Section 5.3.4. The higher number may be attributed to the differences in applied stringencies and the method of filtering single novel unique peptide annotation events between the two studies, as well as differences between the number and accuracy of the MS/MS spectra and genomes used within each study. Of the 66 novel peptides incorporated into the predictions, 8 were exclusively derived from the splice graph, while 2 were identified in both the six-frame translation and splice graph, with the remaining 56 identified exclusively in the six-frame translation.

As demonstrated previously in Section 5.3.4, BLASTP analysis was performed to show how the predictions changed from reference predictions. The analysis was performed by searching the 32,781 Augustus predicted proteins (Table 6.1) against the GENCODE v19 proteins, taking the top match with E-values ≤1E-10. Any sequences that did not match were considered novel predictions, sequences that had a query coverage ≥95% with at least 1 mismatch were considered to be the same prediction as the reference protein, and the remaining matches were considered to be modified predictions, either due to Augustus predicting slightly different models or modified as a direct result of the supporting evidence. From this analysis there were 4,620 non-paralogous novel protein predictions, 12,116 modified predictions and 16,045 predictions considered to be essentially the same as the reference.

Searching all 52 protein predictions that had the novel peptide evidence incorporated, against the GENCODE v19 proteins, taking the top match with E-value ≤1E-10, identified 41 protein predictions likely to be modified predictions, leaving 11 protein predictions, that found no match and were considered as non-paralogous novel protein predictions (Table 6.1).

Based on the annotation events incorporated into the Augustus gene predictions,

the minimum event probabilities which led to a new Augustus gene prediction were: novel gene event 99.986%, gene boundary, translated UTR, and exon boundary 98.167%, reverse strand event 95.615%, novel exon 98.236% and frame-shift 99.799%.

Table 6.1 Summary of human proteogenomics annotations

The results of the proteogenomics analysis of GENCODE v19 showing comparisons between the (1) combined FDR, (2) conservative two-stage FDR and (3) the two-pass search approach with improved two-stage FDR strategy.

	Two-Pass with Improved Two-Stage FDR	Conservative Two-Stage FDR	Combined FDR
Total GENCODE v19 genes	57,820	57,820	57,820
Total 'known' protein-coding genes	20,738	20,738	20,738
Total 'known' proteins	95,309	95,309	95,309
Raw MS/MS search 'known' protein matches ≤1% PSM FDR	6,618	6,607	3,715
Proteogenomics mapping: Total 'known' proteins ≤1% PSM FDR	27,839	27,796	20,793
Proteogenomics mapping: Total 'known' proteins ≤1% PSM FDR (≥2 peptides with 1 unique)	1,047	1,049	769
Total identified 'novel' peptides ≤1% PSM FDR	667	28	270
Raw MS/MS search 'known' peptides ≤1% PSM FDR	38,191	38,151	21,263
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR	38,028	37,987	21,170
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR (≥2 peptides with 1 unique)	9,507	9,512	5,191
Frame-shifts	7	0	0
Translated UTRs	4	0	3
Exon boundaries	27	0	3
Novel splices	0	0	0
Novel exons	23	0	24
Gene boundaries	289 (10)	19 (19)	130 (0)
Reverse strands	262 (50)	9 (9)	77 (8)
Novel genes	5	0	4
Total annotation events	617 (126)	28 (13)	241(42)
Total genes affected	147 (29)	8 (3)	62 (11)
Total proteins affected	609 (116)	28 (13)	240 (41)
Total novel peptides in affected genes/proteins	77	5	46
Total Augustus protein-coding gene predictions	29,266	NA	NA
Total Augustus protein predictions	32,781	NA	NA
Total Augustus gene predictions with incorporated novel peptides	49	NA	NA
Total Augustus protein predictions with incorporated novel peptides	52	NA	NA
Total novel peptides incorporated into Augustus protein predictions	66	NA	NA
Improved protein predictions with incorporated novel peptides	41	NA	NA
Novel non-paralogous protein predictions with incorporated novel peptides	11	NA	NA

NA: Not available

Note: Numbers in parenthesis represent the exclusive numbers. The inflationary effect of a large peptide linkage distance on gene boundaries and reverse strands was removed by assigning a peptide cluster as either a proximal or distal event, not both, with preference placed on proximal events.

In the original proteogenomics study from the ENCODE project [11] a total of 23,085 peptides were identified in both the known and novel search space. In the present study 38,028 known peptides and 77 novel peptides were identified, a total of 38,105 peptides, which was 15,020 more peptides than the ENCODE study was able to identify. Of the 77 novel peptides identified, only 18 were also identified in the ENCODE study. However, the ENCODE study looked at Hg19 GRCh37 and GENCODE v7, while the present study examined Hg19 GRCh37 patch 13 and GENCODE v19. While the GENCODE version would unlikely affect the total known and novel peptides identified, slight differences between the patch versions of genome assembly may have resulted in slightly different numbers of identified peptides, but they would unlikely be as striking as 15,020 peptides. These further identifications were more likely to be due to the differences between the combined FDR strategy, which the ENCODE study used, and two-pass search approach with improved two-stage FDR strategy implemented in the present study. This conclusion was drawn, based on the observed similarities between the 21,533 identified peptides (Table 6.1) using the combined FDR strategy in this study, and the 23,085 peptides identified in the ENCODE project.

6.3.5 Novel gene annotation

A total of 5 novel genes (Table 6.1) were identified, all on unassigned chromosome fragments. An example of one of these novel genes was on unassigned chromosome fragment GL000251.1, spanning positions 2,836,635 to 2,837,405, with an event probability of 99.999%, consisting of 3 unique peptides with 10 PSMs assigned. In addition, Augustus gene predictions were carried out incorporating the novel peptides, predicting two new gene models, gene g28537 spanning positions 2,761,309 to 2,836,886, and gene g28539 spanning positions 2,837,145 to 2,837,411 (Figure 6.2, and with 10 supporting annotated MS/MS spectra in Appendix Figure 6.3).

Performing a BLASTP search against human in NR revealed novel peptide "YFDTAMSR" matched a MHC class I antigen (AAB48498.1 with E-value = 0.045), novel peptide "APWIEQEGPEYWDRNTQIFK" matched MHC class I antigen (AHA90574.1 with E-value = 7E-17), and novel peptide "YLENGKDTLER" matched another MHC class antigen (AGZ87642.1 with E-value = 1E-04), all with 100% query coverage and identity. The Augustus gene prediction g28537 matched human leucocyte antigen B (CAA10522.1 with E-value = 6E-58), with 15% query coverage and 100% identity, also identified as a MHC class I antigen in GenBank, and gene g28539 matched unnamed protein product (BAG64567.1 with E-value = 9E-60), with 100% query coverage and identity, also identified as a MHC class I antigen in GenBank.

Novel peptide "YLENGKDTLER" was incorporated into prediction g28537, while novel peptides "APWIEQEGPEYWDRNTQIFK" and "YFDTAMSR" were incorporated into prediction g28539, with both predictions matching two distinctively different MHC class I proteins in NR. However, prediction g28537 appeared to be over-predicted, with only the first 85 aa (15% query coverage) of the prediction matching human leucocyte antigen B (CAA10522.1). The two 5'-most exons and CDS in the g28537 prediction, which were not supported by the single novel peptide, found no alignment with the human leucocyte antigen B protein, indicating that these two exons may be incorrectly predicted. In contrast, gene g28539, which had two novel peptides incorporated, found a good alignment to unnamed protein product (BAG64567.1), which was also identified as a MHC class I antigen. Therefore, both predictions were probably two distinct proteins. Further evidence, beyond that of the 3 novel peptides is needed to further refine this prediction.

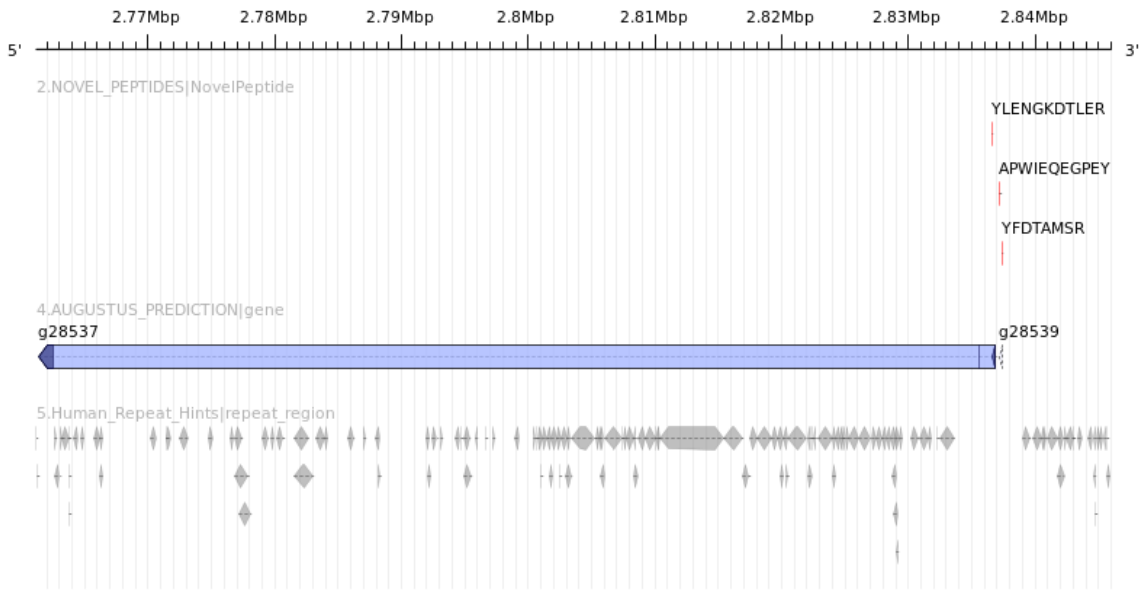


Figure 6.2 Novel gene annotation

A novel gene annotation event is located on chromosome fragment GL000251.1, leading to two new gene predictions on the reverse strand. One novel peptide was incorporated into the large spliced Augustus predicted gene g28537 within the genes first exon/CDS, while the two other novel peptides were incorporated into the smaller neighbouring Augustus predicted gene g28539. Repeat regions span the length of the region, except where the novel peptides were located.

6.3.6 Gene boundary and novel exon annotations

There were 289 (10 exclusive) gene boundary and 23 novel exon annotation events identified (Table 6.1). An example of a gene boundary and two novel exon events were on chromosome 2, spanning positions 69,693,124 to 69,693,567 for the gene boundary event and spanning positions 69,693,124 to 69,693,171 and spanning positions 69,693,190 to 69,693,222 for the two novel exon events, respectively. The gene boundary event was identified for genes ENSG00000115977.14, ENSG00000169599.8 and ENSG00000198380.8, covering protein-coding transcripts ENST00000357308.4, ENST00000361060.5, ENST00000410022.2, ENST00000474230.1, ENST00000303698.3, ENST00000394305.1, ENST00000462320.1, ENST00000450796.2, ENST00000484177.1, ENST00000419370.1, ENST00000438184.2, ENST00000409085.4 and ENST00000406297.3, with an event probability of 99.999% and 3 PSMs identifying 3 novel and unique peptides. The two novel exon events were identified for gene ENSG00000115977.14, with transcript ENST00000409068.1, with an event probability of 99.80%, and 1 PSM identifying 1

novel and unique peptide for each of the two annotation events. Each of the three novel peptides from the gene boundary event was also identified as a novel exon event. Novel peptide “TQNNLESDYLAR” spanning positions 69,693,532 to 69,693,567, fell below the event probability threshold of 99.80% for proximal events, with an event probability of 98.98%, and was therefore removed from further analysis. However, when grouped together in a larger peptide cluster as with the gene boundary event it was included, and contributed to the higher event probability of the gene boundary event. Interestingly, the novel peptides also overlapped the long non-coding RNA (lncRNA) gene ENSG00000188971.4.

Since gene ENSG00000115977.14 was in much closer proximity to the peptide cluster, which also annotated protein-coding transcript ENST00000409068.1 with a novel exon, it was the best candidate for an annotation event amongst the other genes, which were only included due to the large peptide linkage distance. In addition, there were known mapped peptides supporting this gene for these annotations: four shared known mapped peptides to protein-coding transcripts ENST00000406297.3 and ENST00000409085.4, and three shared known mapped peptides to protein-coding transcript ENST00000409068.1. Therefore, the three novel peptides were more likely to belong to this gene, and may extend the boundaries of transcripts ENST00000406297.3 and/or ENST00000409085.4, with transcript ENST00000409068.1 being exclusive for the identified novel exons. Augustus gene predictions were carried out incorporating the novel peptides and reference annotations, predicting a gene model which was an extended version of the original reference gene ENSG00000115977.14, protein-coding transcript ENST00000409085.4, incorporating the novel peptides (Figure 6.3, and with 3 supporting annotated MS/MS spectra in Appendix Figure 6.4).

Performing a BLASTP search against human in NR, revealed that novel peptide “SSMAPSQPEESDVFLR” matched unnamed protein product (BAC86877.1 with E-

value = 2E-09) with 100% query coverage and identity, novel peptide "VPNDDMDIFSK" matched alternative protein AAK1 (CCO13666.1 with E-value = 5E-05) with 100% query coverage and identity, and novel peptide "TQNNLESDYLAR" matched an uncharacterised protein (Q6ZSR9.2 with E-value = 2E-05) with 100% query coverage and identity. All were described as containing a BMP-2-inducible protein kinase C-terminus domain. The reference protein ENST00000406297.3 matched AP2-associated kinase 1 (AAI04843.1 with E-value = 0.0) with 100% query coverage and 99% identity, and reference protein ENST00000409085.4 and ENST00000409068.1, both matched AP2-associated kinase 1 (Q2M2I8.3 with E-value = 0.0) with 100% query coverage and identity, and 98% query coverage and 99% identity, respectively. The new Augustus gene prediction also matched AP2-associated protein kinase 1 (Q2M2I8.3 with E-value = 0.0) with 71% query coverage and 100% identity.

The new prediction showed a 29% reduction in the query coverage compared to the original prediction due to the extension of the new protein prediction towards the proteins C-terminal end, which may be attributed to a single protein-coding transcript, or all transcripts, as evidenced by the known peptides, indicating that each protein isoform appeared to be expressed. Since the known peptides were shared, there was a concern that they were derived from different proteins, however this is not the case, as they were only identified within the same gene across the different protein isoforms within that gene. Overall, the prediction improved, extending the C-terminus of the protein. But further proteomics evidence is needed to improve the coverage of novel peptides across the region, helping to improve the confidence of this annotation event, and also to identify any unique known peptides to unambiguously identify at least one protein isoform as being expressed.

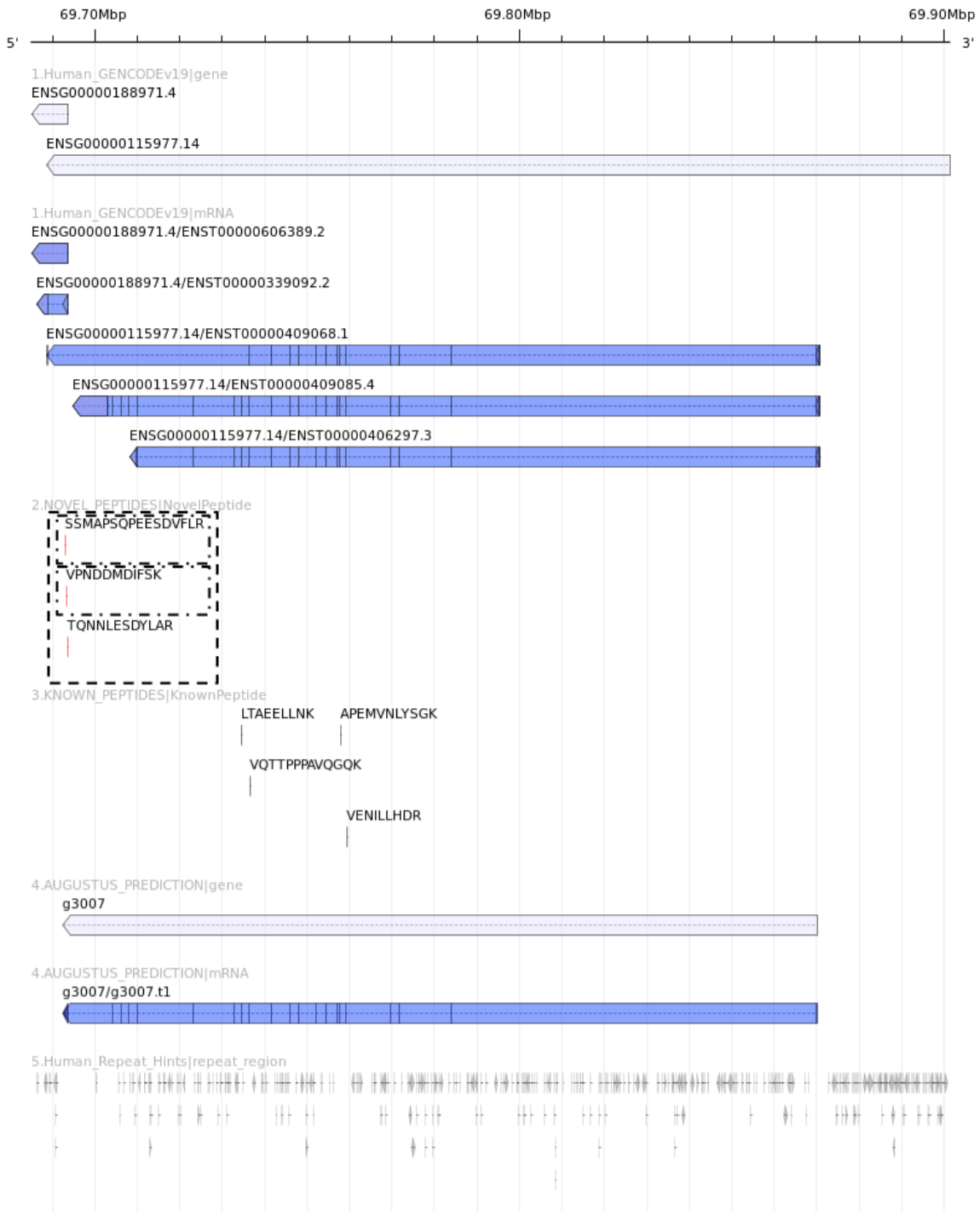


Figure 6.3 Gene boundary and novel exon annotations

A gene boundary annotation (dashed lines) and two novel exon annotation events (dash dotted lines) were located on chromosome 2. These novel peptides led to a new prediction, extending the region of the original transcripts and including a new exon and CDS region previously unknown. Shared known peptides were also found mapping to each of the protein-coding transcripts involved in the annotation event. Repeats were also dotted across the gene region.

6.3.7 Reverse strand and frame-shift annotation leads to new gene predictions

There were 262 (50 exclusive) reverse strand annotation events identified (Table 6.1), and of these, 27 annotation events directly overlapped a gene/protein-coding transcript

(5 exclusively). An example of one of these 27 reverse strand events was on chromosome 10, spanning positions 115,609,828 to 115,654,858 for gene ENSG00000165806.15 with protein-coding transcripts ENST00000429617.1, ENST00000369331.4, ENST00000369321.2, ENST00000345633.4, ENST00000369318.3, ENST00000369315.1 and ENST00000452490.2; gene ENSG00000148735.10 with protein-coding transcripts ENST00000361048.1, ENST00000369312.4, ENST00000369310.3, ENST00000369309.1, ENST00000354462.3 and ENST00000448805.1; gene ENSG00000196865.4 with protein-coding transcript ENST00000369301.3; gene ENSG00000234631.1 with protein-coding transcript ENST00000451472.1; and gene ENSG00000043591.4 with protein-coding transcript ENST00000369295.2, with an event probability of 99.996% and 3 PSMs identifying 2 unique and 1 shared novel peptide. The majority of these genes were included due to the large 150,000 bp peptide linkage distance. The most likely gene for this annotation event was gene ENSG00000196865.4 with protein-coding transcript ENST00000369301.3, which directly overlapped the peptide cluster. In addition, there was 1 unique known peptide supporting the original reference protein-coding transcript ENST00000369301.3, with the unique peptide unambiguously indicating its expression. The peptide cluster also overlapped gene ENSG00000198924.3 with protein-coding transcripts ENST00000361384.2 and ENST00000369305.1, with one of the three novel peptides spanning positions 115,609,828 to 115,609,857 inferring a frame-shift event, falling outside of the 99.80% event probability threshold for proximal events with a 99.799% event probability. This novel peptide was saved due to its presence in the reverse strand event, part of the larger peptide cluster with a 99.996% event probability.

Augustus gene predictions were carried out incorporating the novel peptides and reference annotations and predicting two genes: one in line with the reference gene

ENSG00000196865.4 and protein-coding transcript ENST00000369301.3, and the other prediction on the reverse strand, incorporating two of the novel peptides. There was also a new prediction, incorporating the novel frame-shift peptide and separate from the other two novel peptides in the cluster which split the original reference gene ENSG00000198924.3 with protein-coding transcripts ENST00000361384.2 and ENST00000369305.1 into two predictions, by changing the frame of exon 2 halfway through (Figure 6.4, and with 3 supported annotated MS/MS spectra in Appendix Figure 6.5).

Performing a BLASTP search against human in NR revealed the novel peptide "MGVAAHPK" matched protein S100-A8 (NP_002955.2 with E-value = 20) with 100% query coverage and 88% identity, described as a S100 calcium-binding protein A8 (calgranulin A). Novel peptide "FLCTRHCSK" matched hCG1994383 (EAW52580.1 with E-value = 4.8) with 66% query coverage and 100% identity, described as a provisional hypothetical protein, both of which led to the reverse strand prediction. The novel peptide which led to the frame-shift event and resulting revised prediction was novel peptide "KMMTAVVFLK" which matched SLC35E1 protein (AAH14557.1 with E-value = 9.8) with 80% query coverage and 70% identity, described as containing a Triose-phosphate Transporter family domain. Many of these novel peptides matched proteins in NR with poor significance because all were quite short (≤10aa). For the reference proteins and new Augustus gene predictions in which the novel peptides were incorporated, the majority of the matches were significant. The reference protein ENST00000369301.3 from gene ENSG00000196865.4, for which the reverse strand event was identified, matched NHL repeat-containing protein 2 (NP_940916.2 with E-value = 0.0) with 100% query coverage and identity. Reference proteins ENST00000361384.2 and ENST00000369305.1 from gene ENSG00000198924.3, for which the frame-shift event was identified, both matched

DNA cross-link repair 1A protein (NP_055696.3 with E-value = 0.0) with 100% query coverage and identity. The new Augustus gene prediction g13908 on the reverse strand matched growth-inhibiting protein 6 (ACA14135.1 with E-value = 0.28) with 20% query coverage and 58% identity, indicating what is possibly a truly novel identification. Augustus gene prediction g13909 matched NHL repeat-containing protein 2 isoform X1 (XP_011538071.1 with E-value = 0.0) with 88% query coverage and 100% identity (mostly in line with the original reference prediction on the forward strand). The predictions inferred from the novel frame-shift peptide was gene g13906, which matched DNA cross-link repair 1A protein (NP_055696.3 with E-value = 0.0) with 100% query coverage and identity; and gene g13907 in which the novel frame-shift peptide was incorporated matched DNA cross-link repair 1A protein (NP_055696.3 with E-value = 1E-150) with 76% query coverage and 100% identity.

The new gene prediction g13908 on the reverse strand was further supported by one of the novel peptides at the N-terminal-most end, which was N-terminal acetylated, indicating that the start of translation predicted by Augustus also agreed with the PTM presented by the proteomics evidence, thus further supporting the identification of a novel gene on the reverse strand (Figure 6.4). Although the peptide cluster identified a reverse strand annotation with two of the three novel peptides contributing to a likely novel gene prediction on the reverse strand, the single outlying novel peptide that instead identified a frame-shift indicated that not all peptides should belong within the same peptide cluster. Particularly as this peptide was previously removed as a frame-shift event due to its lower event probability, possibly as a result of only being a single peptide annotation event with a relatively lower intensity MS/MS spectrum when compared to the other MS/MS spectra (Appendix Figure 6.5C). Situations such as these indicate that more stringent MS/MS spectra quality filtering should be considered, given enough MS/MS spectra to tolerate the MS/MS spectral losses with relatively limited

negative impact on the identification rate. The other alternative would be to simply accept the cost of losing a large proportion of MS/MS spectra from a relatively small MS/MS spectral dataset to improve specificity and reduce errors such as these. Another alternative could be for a re-consideration to how annotation events are filtered, perhaps requiring that all annotation events are further screened to identify unique and novel peptides from smaller annotation events within other larger annotation events, which have previously been filtered out to reduce ambiguity and keep annotation event filtering consistent.

The prediction resulting from the frame-shift peptide had its second exon split into two different frames resulting in two different predictions, with both matching the same protein, DNA cross-link repair 1A protein (NP_055696.3). When looking at the proteins corresponding gene DCLRE1A in GenBank, there was sufficient RNA-seq coverage across the gene, particularly across the exon that was split in the new prediction. Given this evidence, the new prediction as a result of the frame-shift peptide, was most likely a false positive. As a result, this particular annotation event should be either discarded or at the very least require further evidence before it could be accepted. Careful screening of any new predictions which are a direct result of proteogenomics annotation should therefore be conducted by comparison with the previous predictions, particularly when the previous predictions were manually curated and considered to be of high confidence, and supported by numerous types of other orthogonal evidence such as RNA-seq and EST. Overall, there were another 7 frame-shift events, derived from only 2 peptide clusters, which were above the event probability threshold. However, these could not be incorporated into gene predictions indicating the possibility that all frame-shift events throughout this particular study were false positives, particularly when considering the annotation events directly challenged the highly curated predictions that contained multiple sources of orthogonal evidence.

Based on this observation, many of the new predictions that were a result of dramatic changes (i.e. such as changes in frame) to the current well-curated GENCODE v19 annotations, were possibly false positives. The likely exceptions being annotation events such as novel genes, translated UTRs, reverse strand and exon boundary events that do not drastically conflict with but only add to or extend the annotation. In less adequately curated reference annotations, this judgement would likely not hold true.

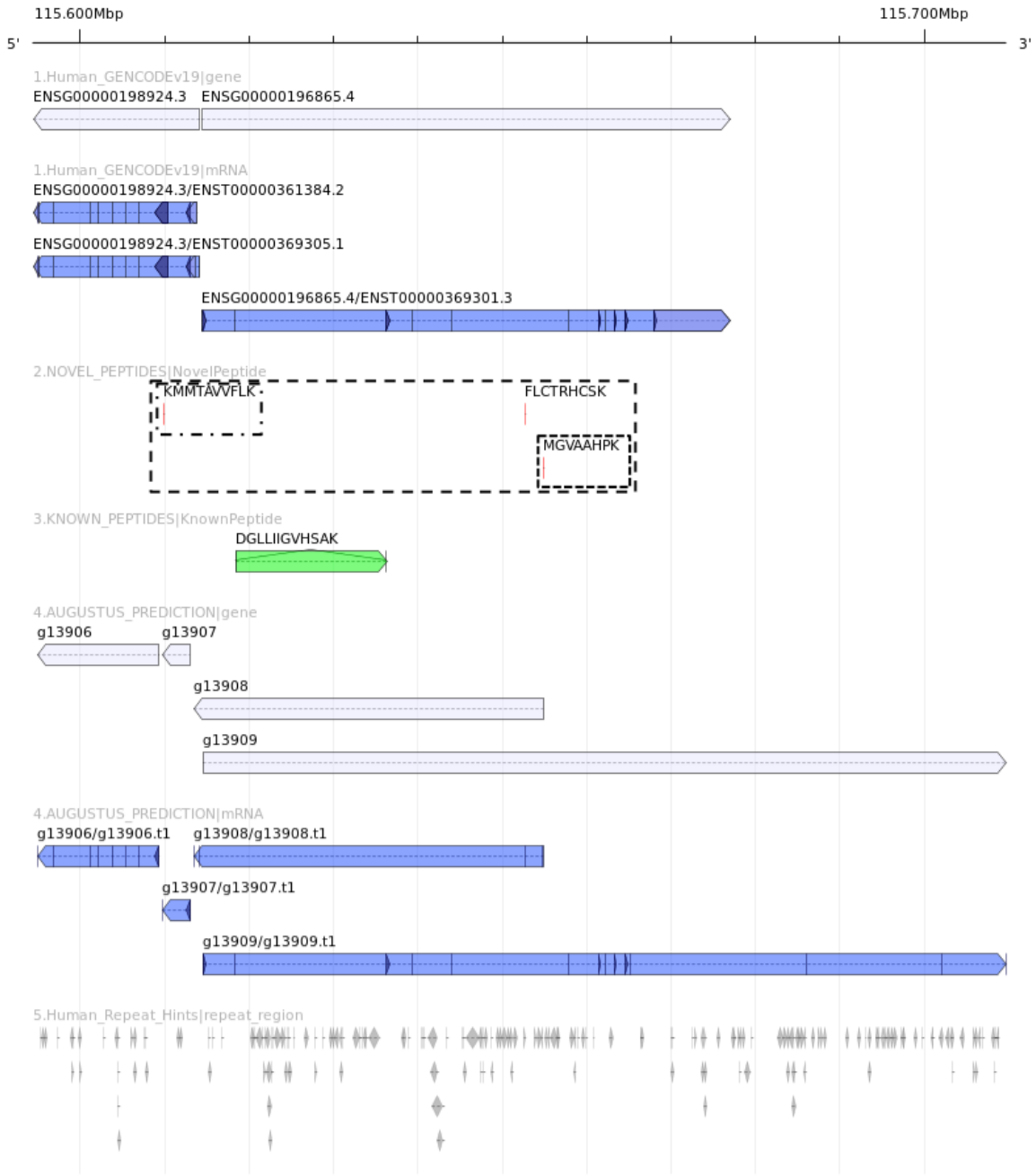


Figure 6.4 Reverse strand and frame-shift annotation

A reverse strand annotation event (dashed lines) and a frame-shift annotation event (dash dotted lines) were located on chromosome 10. An N-terminal acetylated peptide was also identified within the reverse strand event (square dotted lines). These novel peptides led to two new predictions; a novel gene prediction on the reverse strand with the incorporated N-terminal acetylated peptide agreeing with the genes translation initiation start (TIS) site, and a modified prediction from the frame-shift event, possibly a false positive. Repeats were also dotted across the gene regions.

6.3.8 Exon boundary and translated UTR annotation

Four translated UTR annotations and 27 exon boundary annotations were identified (Table 6.1). An example of an exon boundary and a translated UTR annotation was on chromosome 1, spanning positions 45,213,052 to 45,213,121, where a translated UTR

annotation was identified for gene ENSG00000142945.8 with protein-coding transcript ENST00000372222.3, and an exon boundary was identified for the same gene, with protein-coding transcripts ENST00000452259.1, ENST00000372224.4, ENST00000372218.4, ENST00000455186.1 and ENST00000372217.1. Both annotation events had an event probability of 98.167% with 1 PSM identifying 1 novel and unique peptide. In addition, supporting these two annotations, there were known mapped peptides, with 9 mapping to protein-coding transcript ENST00000372222.3, 2 mapping to ENST00000452259.1, 11 mapping to ENST00000372224.4, 9 mapping to ENST00000372218.4, 4 mapping to ENST00000455186.1 and 10 mapping to ENST00000372217.1. However, none of the known peptides were unique, making it impossible to determine which protein-coding transcript was more likely to be expressed over the others, and also made it difficult to determine the protein-coding transcript to which the annotation event most probably belonged. Additionally, the annotation event may be a translated UTR or an exon boundary for one or all of these protein-coding transcripts. Augustus gene predictions were carried out incorporating the novel peptide and reference annotations, predicting a gene model in line with the original reference gene model ENSG00000142945.8, and also with an additional protein-coding transcript with a longer exon extended in the 5' direction incorporating the novel peptide (Figure 6.5, and with one supporting annotated MS/MS spectrum in Appendix Figure 6.6).

Performing a BLASTP search against human in NR revealed the novel peptide matched kinesin-like protein KIF2C isoform 2 (NP_001284584.1 with E-value = 8E-16) with 100% query coverage and 96% identity. The reference protein ENST00000372222.3, with which the exon boundary event was identified for, matched kinesin-like protein KIF2C isoform 1 (NP_006836.2 with E-value = 0.0), with 100% query coverage and identity. The reference proteins which had translated UTR events

identified were ENST00000452259.1 which matched kinesin-like protein KIF2C isoform 2 (NP_001284584.1 with E-value = 0.0) with 100% query coverage and identity; ENST00000372224.4 which matched kinesin-like protein KIF2C isoform 1 (NP_006836.2 with E-value = 0.0) with 100% query coverage and identity; ENST00000372218.4 which matched kinesin-like protein KIF2C isoform 2 (NP_001284584.1 with E-value = 0.0) with 100% query coverage and identity; ENST00000455186.1 which matched mitotic centromere-associated kinesin (AAC27660.1 with E-value = 0.0) with 95% query coverage and 99% identity; and ENST00000372217.1 which matched kinesin-like protein KIF2C isoform 2 (NP_001284584.1 with E-value = 0.0) with 100% query coverage and identity. The new Augustus gene prediction contained two protein-coding transcripts, the first transcript being the altered prediction containing an extended exon, g772 transcript 1, which matched kinesin-like protein KIF2C isoform 1 (NP_006836.2 with E-value = 0.0) with 100% query coverage and 97% identity; and g772 transcript 2 was unaltered, in line with the original prediction which matched kinesin-like protein KIF2C isoform 1 (NP_006836.2 with E-value = 0.0) with 100% query coverage and identity.

The novel peptide overlapped the exon in the 5’ direction by 1 aa and led to a new transcript prediction g772 transcript 1, which extended the exon by 20 aa in the 5’ direction and accounted for the 97% identity with its match to the target sequence in NR: kinesin-like protein KIF2C isoform 1. However, the g772 transcript 2 prediction reproduced the original reference protein, and showed full alignment with the protein with 100% identity. The presence of known peptides around the region of the novel peptide improved the confidence of the annotation event by implying that it was being expressed. However, the known peptides were shared in multiple other identifications in different protein isoforms within the same gene and also found matches in a few other paralogous genes making the identification ambiguous. Although the presence of this

novel peptide, identified as both a translated UTR and exon boundary, looked promising, further evidence will be needed to support and confirm this annotation event and prediction.

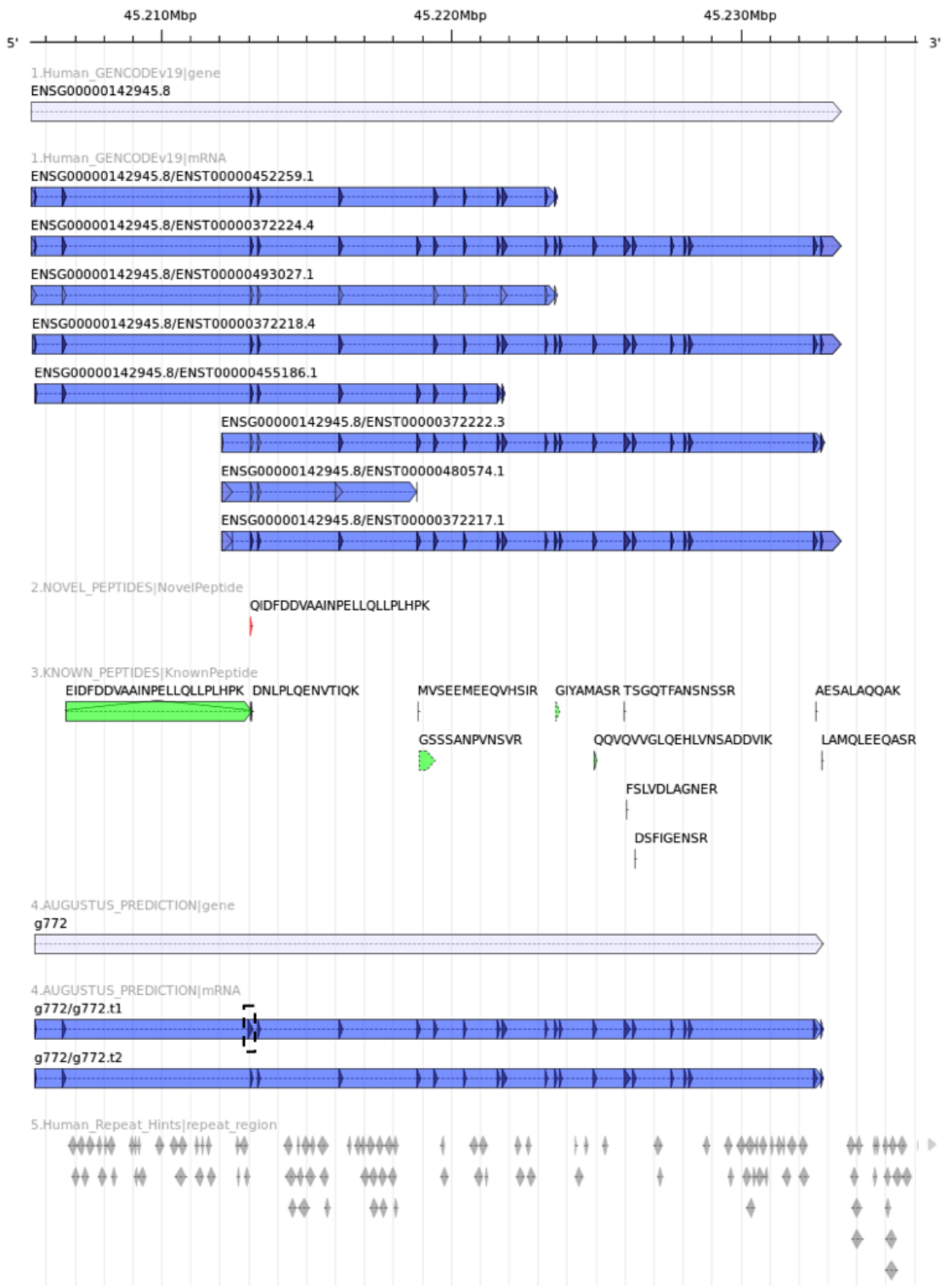


Figure 6.5 Exon boundary and translated UTR annotation

An exon boundary annotation event and a translated UTR annotation event for the same novel peptide cluster were located on chromosome 1. This novel peptide was incorporated into a new prediction, predicting a new transcript in addition to the original transcript from the reference gene, by extending the exon region in the 5' direction (dashed lines). Repeats were also dotted across the gene region.

There were another 3 translated UTR annotation events identified from a single peptide cluster on chromosome 18, spanning positions 3,978,073 to 3,978,145, for gene ENSG00000170579.10, with transcripts ENST00000315677.3, ENST00000515196.2

and ENST00000581527.1 with an event probability of 99.997%, and with 5 PSMs identifying 2 novel and unique peptides. This peptide cluster also happened to overlap an already identified pseudogene ENSG00000250933.1, with transcript ENST00000509132.1. The pseudogene in question was glyceraldehyde 3-phosphate dehydrogenase 66 (GAPDHP66), which is known to be expressed in ENCODE tier 1 cell line GM12878 probably as a consequence of genetic changes during immortalization of the cell line using EBV. Since the reference proteome used in the present study was GENCODE v19, this expressed pseudogene was not accounted for among the known proteins and was misinterpreted as a translated UTR in this instance. This finding outlines a need for caution when using datasets to perform proteogenomics annotations, due to the differences in various cell lines. At the very least, if available, the known proteome for that cell line should complement the target known proteome to remove any inconsistencies between the proteomics data source and target genome annotation. The finding also outlined the lack of any support from the Enosi tool, with the identification of expressed pseudogenes or non-coding RNA genes, which could have identified the annotation event instead of it being misinterpreted as a translated UTR and requiring further manual parsing of the results for correct interpretation.

Continuing the theme of expressed pseudogenes: one identified pseudogene, ENSG00000257907.2, with transcript ENST00000547505.2 and classified as translation elongation factor 1 alpha 1 pseudogene 17, contained unique and novel peptide “SGDAAIVDMVPGKPMCVESFSVYPPLSR” on chromosome 12, spanning positions 44,054,701 to 44,054,784. The novel peptide was primarily identified by proteogenomics analysis as a gene boundary and reverse strand event with an event probability of 98.167% for a number of genes in the region. However, no other novel peptides were in close proximity to indicate a gene spanning this region, possibly indicating a misidentified annotation event, solely based on the peptide linkage distance.

A BLASTP search of the novel peptide revealed a match to a protein-coding gene translation elongation factor 1 alpha 1 (NP_001393.1 with E-value = 4E-19) with 100% query coverage and 93% identity on chromosome 6, however the alignment contained 2 mismatches. It is quite possible that the peptide was derived from an actual protein-coding translation elongation factor 1 alpha 1 from cell line GM12878 elsewhere in the genome and, due to sequence variation, found a match to identified pseudogene ENSG00000257907.2. Further novel peptides mapping to the pseudogene would need to be identified before it would be feasible to re-annotate this pseudogene as a functional protein-coding gene.

6.3.9 N-terminal acetylated peptides

Protein N-terminal acetylation is known to contribute to many functional changes in proteins and the identification of such changes in conflict with the known proteome can potentially indicate over-predicted genes requiring re-annotation, but could also indicate an alternative protein isoform with a different translation initiation start (TIS) site. A total of 729 N-terminal acetylated peptides were identified amongst 2,949 proteins in the GENCODE v19 annotation and 136 were identified from 131 high confidence proteins (≥2 peptides with at least 1 unique), using the two-pass search approach with improved two-stage FDR strategy.

Many of the N-terminal acetylated peptides appeared to agree with the TIS sites of the reference proteins, whereas others did not. These conflicts could not be accounted for because the N-terminal acetylated peptides were shared with other proteins, which supported their TIS sites. Due to the large number of these peptides a thorough manual validation was not practical, requiring a way to automate the process. However, even if this were done, numerous N-terminal acetylated peptides would still be shared between many other proteins due to the protein inference problem, with the problem amplified in this case by the presence of numerous isoforms in GENCODE v19. Thus making any

findings that have conflicts with the annotation ambiguous, as found in Section 5.3.13. Therefore, to improve the specificity of any findings only unique known N-terminal acetylated peptides were identified, as these could more confidently be assigned as alternate TIS sites. Alternatively, as discussed in the previous chapter (Section 5.3.13), N-terminomics [452] and other methods, such as heuristic approaches and top-down proteomics, could be used in the future to resolve such ambiguous analysis.

From the 729 N-terminal acetylated peptides, 132 were found to be unique amongst 123 proteins, all of which did not conflict with the GENCODE v19 annotation.

Of the 77 novel peptides identified (Table 6.1), 4 were found to be N-terminal acetylated (Appendix File 6.7). Of these, 3 were found to be unique, 2 of which were incorporated into different gene predictions at the N-terminal most-end, and one could not be incorporated into the prediction. The fourth peptide was a shared peptide, found in a peptide cluster consisting of two other novel peptides, but it was also mapped to 76 other genome locations. In one of these other locations within the same chromosome, the novel peptide was incorporated into a new gene prediction.

The first of the novel N-terminal acetylated peptides was unique peptide "LTNQDCPGRER", spanning positions 19,524,186 to 19,524,218, on chromosome 8, identified within a gene boundary event and reverse strand event with an event probability of 99.927%, and with 1 annotated MS/MS spectrum in Appendix Figure 6.7. This particular peptide could not be incorporated into any gene prediction.

The second novel N-terminal acetylated unique peptide was "MGVAAHPK", spanning positions 115,654,835 to 115,654,858 on chromosome 10, identified within a gene boundary event and reverse strand event with an event probability of 99.996%, with 1 annotated MS/MS spectrum (Appendix Figure 6.5A). This peptide was incorporated into Augustus gene prediction g13908 at the N-terminal-most end,

highlighted previously as an example of a reverse strand annotation, also shown in Figure 6.4, further supported by N-terminal proteomics evidence.

The third novel N-terminal acetylated unique peptide was “MTLRGCNQK”, spanning positions 16,027,267 to 16,027,293, on chromosome 6, identified within a gene boundary event and reverse strand event with an event probability of 99.995% and with 1 annotated MS/MS spectrum in Appendix Figure 6.8. This peptide was incorporated into Augustus gene prediction g8234 at the N-terminal-most end, spanning positions 16,027,267 to 16,100,195, which found no known matches to proteins in NR, indicating the identification of a truly novel gene, further supported by N-terminal proteomics evidence. Interestingly, the protein sequence of g8234 contains a long chain of arginine, lysine and glutamic acid residues, all of which are highly charged, potentially indicating an important role in ligand binding and protein-folding.

The fourth and final novel N-terminal acetylated shared peptide was “PGDSIRSHR”, spanning positions 16,165,749 to 16,165,775 on chromosome 6, identified within the same gene boundary event and reverse strand event as peptide “MTLRGCNQK”, with 2 annotated MS/MS spectra in Appendix Figure 6.9. The peptide was not incorporated into a gene prediction at this location; however, the peptide was found with 76 other genomic locations. At one of these locations, specifically spanning positions 83,808,134 to 83,808,160 on the same chromosome 6, residing within an intron region of gene ENSG00000083097.10 with protein-coding transcripts ENST00000349129.2, ENST00000237163.5, ENST00000536812.1, and ENST00000369739.3, there was a possible novel exon event. This particular annotation event was not formally identified because it was filtered out as it only contained shared peptides. At this location the novel peptide was incorporated into Augustus gene prediction g8936 within exon 2 and CDS 2, shared by two transcripts, which spanned positions 83,806,697 to 83,877,886.

Performing a BLASTP search of the two protein-coding transcripts from gene g8936 against human in NR revealed that protein-coding transcripts 1 and 2 matched protein dopey-1 isoform b (NP_001186871.1 with E-value = 0.0) with 100% query coverage and 99% identity and 100% query coverage and 98%, respectively, within the same chromosome 6 and within the same genomic region. However, the region where the novel peptide was incorporated did not show complete homology with the protein, indicating a potential new gene structure or isoform for this protein. Because the peptide was also located in another 75 locations, it was likely that the N-terminal acetylated status of the peptide agrees more with one of the other locations, and due to the protein inference problem the true location of this particular peptide will remain ambiguous until further evidence becomes available.

6.3.10 Impact of search space

The impact that an inflated search space has on proteogenomics analysis has often been a challenge, as outlined previously in Chapters 4 and 5. The combined FDR strategy has been shown to be limited in its sensitivity and reduces the identification rate in both known and novel search spaces. While the conservative two-stage FDR strategy outlined in [474] showed improved identification rate of known proteins, it was overly conservative for the novel identifications and did not take advantage of improved sensitivity using a two-pass search approach such as that mentioned in [427]. The enhanced method of controlling FDR by implementing the use of a two-pass search approach with improved two-stage FDR strategy demonstrated better balance of FDR control, improved the discrimination between true and false positives in the known and novel search spaces, and allowed for more peptides and proteins to be identified when compared to the combined FDR and conservative two-stage FDR strategies (Table 6.1). This was accomplished by reducing the impact that the proteogenomics search space has on analysis, through a reduction in the size of the search space to only sequences

containing likely matches to either the known or novel search spaces, prior to 1% PSM FDR filtering.

At the known protein level protein identification rates during MS/MS database search changed from 3.9% of all proteins with combined FDR, to 6.93% with the conservative two-stage FDR, and 6.94% with the two-pass search approach with improved two-stage FDR strategy. Overall, there were 44% more known proteins identified with the two-pass search approach with improved two-stage FDR strategy than compared with the combined FDR strategy. At the novel peptide level, identification rates during MS/MS database search changed from 270 with combined FDR, down to 28 with the conservative FDR (a 10x drop in sensitivity), with an increase to 667 with the two-pass search approach with improved two-stage FDR strategy (a 2.5x improvement over the combined FDR strategy) (Table 6.1). This new method resulted in improvements in sensitivity, where previous losses were towards 30% of the known proteins demonstrated in Chapter 4 and 52% of the known proteins demonstrated in Chapter 5. Using the new method, no losses of known proteins were observed. Instead, slight gains in the identification of known proteins could be seen using the two-pass search approach when compared to the conservative two-stage FDR strategy, which did not implement a two-pass search approach. However, the opposite was true for the high confidence protein identifications, although this was a negligibly small difference. Future algorithmic improvements will likely result in better identification rates and provide a means of reducing the impact that the proteogenomics search space has on analysis.

6.4 SUMMARY

This study integrated different –omics platforms: genomics, proteomics and transcriptomics, available from a previous proteogenomics study which was part of the

ENCODE project, and identified an improved methodology for conducting proteogenomics. This new methodology, which improves on the sensitivity and specificity of a proteogenomics search, was directly compared to two previous methods: the combined FDR and 'conservative' two-stage FDR strategy. In addition, this study identified 15,020 more peptides when compared to the previous ENCODE proteogenomics study.

The present study made a significant contribution to the annotation of the human genome, identifying 77 novel peptides contributing to 617 novel annotation events (126 exclusively), consisting of 7 frame-shifts, 4 translated UTRs, 27 exon boundaries, 23 novel exons, 289 gene boundaries (10 exclusively), 262 reverse strands (50 exclusively), and 5 novel gene events, amongst a total of 147 genes (29 exclusively) and 609 proteins (116 exclusively).

Of these annotations 66 novel peptides directly led to 52 predicted proteins via Augustus gene prediction. The two-pass search approach with improved two-stage FDR strategy proved that after filtering it was able to identify 35 more novel peptides than either the combined FDR or conservative two-stage FDR strategies, and as a result it identified more annotation events, thus reducing the negative impact of the inflated proteogenomics search space and reducing the overhead for the post-processing of raw results.

The number of annotation events identified in this study was far higher than the relatively few peptides identified, and more pronounced than found in the previous study in Chapter 5, due to the much larger peptide linkage distance used in this study, as was explained in Section 6.3.4. Although this caveat was accounted for by also providing the number of exclusive annotation events and associated genes and proteins, the problem still remains and should be addressed appropriately in future studies by

applying a novel way to determine the most appropriate peptide linkage distance for each peptide cluster. Other identified caveats included misidentified annotation events due to the lack of proper annotation event coverage; such as the absence of a way to identify expressed pseudogenes and non-coding RNAs, discrepancies between the proteome of GENCODE v19 and cell line GM12878 and overlapping annotation events causing ambiguities when annotation events were filtered, leading to false positive predictions.

6.5 CONCLUSIONS

The implementation of a two-pass search approach with improved two-stage FDR strategy proved highly effective at improving the PSM identification rate, and which reduced losses of known proteins as the searches were conducted separately and did not impact on the sensitivity of known and novel peptide sequence identification within their respective searches. This can be seen when comparing the identified known proteins between the two-stage FDR strategy and the combined FDR strategy (Table 6.1), as well as from the proteomics-only searches during the precursor mass tolerance optimization step from Section 6.4.1. As a result, this led to improvements in the discrimination of true and false positives and also reduced the overhead needed to process the results for FDR filtering.

As in the previous studies presented in Chapters 4 and 5, clustering MS/MS spectra and selecting the most appropriate precursor mass tolerances was also effective, particularly with the high-accuracy MS/MS spectra utilized in the present study. Although higher stringency MS/MS spectral quality filtering may have improved specificity and reduced the occurrence of some identified false positives found during analysis, this would have been at the cost of impacting true positive identifications. Much larger MS/MS spectral datasets may afford such higher MS/MS spectral quality

stringencies with a relatively limited negative impact on the identification rates.

The present study was compared to the ENCODE study from [11]. In this study, a total of 38,105 peptides were identified, of which 77 were novel peptides. Compared to the original ENCODE study, an additional 14,961 known peptides and 59 novel peptides were identified.

In the absence of a method to determine the annotation event FDR to identify appropriate event probability thresholds, the screening of annotation events by applying tighter stringencies when searching against RefSeq protein proved an effective way of identifying more annotation events outside of the applied stringent event probability thresholds.

An issue arose in Section 6.3.7 where a frame-shift peptide with an event probability below the applied threshold was incorporated into a larger peptide cluster with a higher event probability, identified as a reverse strand event. Because the single frame-shift peptide was still retained, this led to a probable false positive prediction. Suggestions to resolve this issue could be that, when applying filtering, an additional level of filtering should be applied to further remove any single unique and novel peptide annotation events that have been previously filtered out. This step could be achieved by also filtering from larger annotation events, thus reducing ambiguity, improving specificity and ensuring consistency across all annotation event filtering, as well as avoiding the occurrence of false positive annotation events being included in the revised annotation.

Another issue arose in Section 6.3.8, with the identification of a false positive translated UTR event, which was really a protein-coding pseudogene, namely glyceraldehyde 3-phosphate dehydrogenase 66 (GAPDHP66), identified as only a protein-coding gene in GM12878. This finding highlighted two problems: 1) the

proteome from GENCODE v19 did not reflect the proteome of cell line GM12878, which should be appended, if possible; and 2) the Enosi tool did not accommodate the identification of possible coding pseudogenes and/or non-coding RNA genes, and instead misidentified an annotation event, which required further manual curation of all identified annotation events. The ability to identify such annotation events automatically is important, as it removes the possibility of misidentifications and ambiguity, and time spent manually interpreting the results. Enabling such a feature could be relatively simple as interpreting the reference annotations and creating a new annotation event type, or running the analysis in parallel with other proteogenomics tools, such as PGTools [496], which can handle such annotation events.

Another caveat of the analysis, which was highlighted in the previous two chapters, was how annotation events could be better defined using other evidence, such as known peptides in close proximity, and dynamically defined peptide linkage distances across the genome for each peptide cluster.

Another consideration for improving how annotation events could be defined came to light during the analysis in Section 6.3.7 with the application of N-terminal acetylated peptides, by examining the location of unique N-terminal acetylated peptides in the context of the peptide cluster, to define the cluster boundaries. However, this tactic would be conditional on the fact that the provided genome was near complete with little fragmentation, in order to reduce as much as possible the unambiguity when assigning an N-terminal acetylated peptide as unique.

Four N-terminal acetylated novel peptides were identified in this study, with three being unique, of which, two were included in Augustus gene predictions at the N-terminal-most end, while there were numerous known N-terminal acetylated peptides, with 132 found to be unique among 123 proteins, none of which conflicted with the

GENCODE v19 annotation.

Besides the use of unique peptide parsimony, other methods to identify these N-terminal acetylated peptides in the context of known proteins and novel annotations are needed, e.g. in a proteomics-only context using tools such as those listed in Table 2.6, top-down proteomics and N-terminomics [452].

Future studies on the human genome could further expand on this work by including other cell lines, besides GM12878, e.g. K562: an immortalized cell line produced from a female patient with chronic myelogenous leukaemia (CML); A549: an adenocarcinomic human alveolar basal epithelial cell line; H1 human embryonic stem cell line: H1-neuron cell line from neurons derived from H1 embryonic stem cells; and A431: a human vulvar cancer cell line. All of these cell lines are publically available from ENCODE and other sources. In addition, much larger comprehensive data sources could be obtained from previous proteogenomics studies, such as the recent human proteome mapping study [450, 451], which mapped the proteome from multiple cell lines and tissues. Sources of RNA-seq data could be obtained from large comprehensive studies, such as those conducted in the 1,000 Genome Project [65], a large international project to sequence 1,000 genomes and identify genomic variants. Such studies would greatly benefit from using comparable proteomics data from the same biological samples to be truly meaningful, thus providing a means of confidently identifying variant peptides between different individuals.

6.6 ACKNOWLEDGEMENTS

Chapter 6 is in preparation for publication along with new human gene and protein predictions, which will be submitted to NCBI. The dissertation author is the primary author of this paper. The dissertation author designed the proteogenomics workflow, ran the analysis and wrote the paper. The dissertation author would like to thank the Centre

for Comparative Genomics for their compute resources and guidance and the Pawsey Supercomputing Centre for the use of their compute resources, which were supported by funding from the Australian Government and the Government of Western Australia.

7 WHEAT PROTEOGENOMICS

7.1 INTRODUCTION

Bread wheat (*Triticum aestivum*) is a major food crop for human consumption, contributing to a large proportion of our staple diet, comprising approximately 20% of calories consumed, being a good source of protein, vitamins and minerals, and also as a good livestock feed. Originally derived from a cross-hybridization of cultivated tetraploid emmer wheat (AABB, *Triticum dicoccoides*) and diploid goat grass (DD, *Aegilops tauschii*) approximately 8,000 years ago, the net result was a allohexaploid genome (6x), consisting of an A, B and D genome with a combined genome size of 17 Gbps, making it one of the largest known plant genomes around [595, 596].

Recently, the International Wheat Genome Sequencing Consortium (IWGSC) sequenced and assembled the genome of *Triticum aestivum* cultivar Chinese Spring done on a per chromosome arm basis, by isolating and sorting each chromosome arm prior to sequencing [12]. A total of 124,201 gene loci were identified, evenly distributed across the chromosomes and subgenomes. In the same study, the dissertation author contributed towards proteogenomics analysis, which validated 50 high confidence genes, and identified 16 novel peptides which contributed to 33 novel annotations in 13 genes: 4 frame-shifts, 3 translated UTRs, 2 exon boundaries, 13 novel exon events, 4 gene boundary events, 2 reverse strand events and 5 novel gene events. However, the proteogenomics study was limited in scope with only 11,334 MS/MS spectra and implementing an earlier version of the Enosi tool, with no splice graph to define the splice junction search space, and with no method employed to address the huge loss in sensitivity resulting from searching a six-frame translation of the 17 Gbp genome [12].

7.1.1 Outline of this study

The present study re-visited the proteogenomics annotation and addressed the limitations from the earlier study by the dissertation author within the study by Mayer

and colleagues [12] by adding an additional sources of MS/MS spectra: 1) Wheat flour from cultivar Butte 86 under three different digest protocols (Trypsin (as per the same as the previous proteogenomics study in [12]), chymotrypsin, and thermolysin [13]); and 2) meiotically developing anthers from a cross between rye *(Secale cereale* cultivar Petkus) and wheat (*Triticum aestivum* cultivar Chinese Spring), digested with trypsin and AspN [14]. The analysis utilised an improved Enosi tool, with incorporated MS-GF+ MS/MS database search tool and a splice graph derived from a large source of RNA-seq data. The splice graph included RNA-seq data used in the study from [12], RNA-seq data derived from maturing grain in a related study [15], as well as a selection of four publically available datasets from the Sequence Read Archive (SRA). The loss in sensitivity was also addressed by applying a two-pass search approach with improved two-stage false discovery rate (FDR) strategy, which aimed to improve the sensitivity and discrimination between true and false positive identifications. In addition, due to the complexity of the allohexaploid wheat genome combined with the highly fragmented draft assembly, the difficulties of performing proteogenomics annotation and accurately predicting genes with such a dataset were highlighted.

7.2 MATERIALS AND METHODS

7.2.1 Proteomics and genomics datasets

The latest version of the assembled *Triticum aestivum* genome sequence and Wheat MIPS version 2.2 annotations as GTF file and protein FASTA file from the study by Mayer and colleagues [12] were downloaded from the URGI web site (https://urgi.versailles.inra.fr/download/iwgsc/). The GTF file was subsequently converted to GFF, using the gtf2gff perl script supplied with the Augustus gene prediction tool (Appendix File 7.1).

The MS/MS spectra for this study were derived from a number of different sources. The first source was from the study in [13], totalling 42,909 MS/MS spectra from trypsin (11,334), thermolysin (16,776) and chymotrypsin (14,799) digests. The proteins were extracted from finely ground *Triticum aestivum* wheat flour cultivar Butte 86 and run on a 2D electrophoresis gel, with 233 spots excised and separately digested with trypsin, chymotrypsin and thermolysin. The resulting digests were then run on a QSTAR Pulsar i quadrupole time-of-flight mass spectrometer (QTOF) (Applied Biosystems/MDS Sciex), with attached nano-electrospray source and nano-flow HPLC. The MS/MS spectra were kindly provided and downloaded from Susan Altenbach and William Vensel at USDA (California), co-authors of a prior proteomics study [13].

The second source of MS/MS spectra was derived from meiotic tissue from the anthers of a wheat-rye hybrid with Ph1 deletion (Ph-) from the study in [14], totalling 42,528 MS/MS spectra from a trypsin digest, with a sub-fraction of 5,392 MS/MS spectra digested with trypsin and AspN. The wheat-rye hybrid came from a cross between rye (S*ecale cereale* cultivar Petkus) and wheat (*Triticum aestivum* cultivar Chinese Spring). Protein extracts from the anthers, were run on a 1D gel, 8 bands were excised and digested using trypsin with 1 band digested with trypsin and AspN. The samples were then run on an LTQ Orbitrap (Thermo Scientific) mass spectrometer with attached nano-flow HPLC. The MS/MS spectra was kindly provided and downloaded from Ali Pendle and Graham Moore, co-authors of a prior proteomics study [14].

The protein sequence predictions from the Wheat MIPS version 2.2 annotations were appended to a source of contaminants before being used in the MS/MS database search as outlined in Section 3.1.1, to identify any contamination. In this study, for the thermolysin and AspN specific searches, thermolysin and AspN contaminant sequences downloaded from UniProtKB/Swiss-Prot were also added to account for the protease used in both cases.

7.2.2 RNA-seq datasets

Illumina RNA-seq datasets were obtained from six different sources. One source was obtained from the *Triticum aestivum* cultivar Chinese Spring study [12], derived from five tissues (root, leaf, stem, spike and grain) of 5 different pooled conditions (SE library). In addition, RNA-seq data that was not used in the study from [12] were used, which included the same five tissues from 5 different conditions and 3 different developmental stages (PE library). These datasets were downloaded from URGI (http://wheat-urgi.versailles.inra.fr/Seq-Repository/RNA-Seq).

The second source of data was obtained from a different study [15], which ran in parallel with [12], derived from the aleurone layer of developing endosperm tissue at different time points from the *Triticum aestivum* cultivar Chinese Spring. These datasets were kindly provided by Odd-Arne Olsen one of the co-authors from the studies [12, 15].

The third to sixth sources of data were obtained from the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) via the DNA Nexus (http://sra.dnanexus.com/). The identifiers used were: SRA053323 (a study which looked at the alueurone layer and starchy endosperm layer at different developmental time points); SRA048049 (a study which examined different tissues from cultivar Kukri); SRA059240 (a study which surveyed various cultivars for variation, including AC_Barrie, Alsen, Baxter, Chara, Excalibur, Kukri, Pastor, RAC875, Westonia, Xianyon54, Volcani, Yitpi, Steele, Stephens, Truman, Caledonia, Grandin, ID0444, Jaypee, and Jupateco); and SRA071558 (a study which looked at heat stress of cultivar HD2985 (thermotolerant) and HD2329 (thermosusceptible)).

7.2.3 MS/MS database searching

The MS/MS database search was performed by MS-GF+, as outlined in Section 3.3. In this case study, trypsin, chymotrypsin, thermolysin and AspN were used as the

proteases for the different sources of MS/MS spectra. For the wheat flour dataset digested with protease trypsin, chymotrypsin and thermolysin, the instrument was set to TOF (QTOF) and for the meiotic tissue dataset, digested with trypsin, with one sub-set dataset digested with trypsin and AspN, the instrument was set to low-res LTQ (Ion Trap). MS-GF+ does not have thermolysin as an option for protease and therefore 'unspecific cleavage' as the protease was chosen, based on evidence where MS-GF+ has previously been shown to identify peptides from unspecified proteases, such as αLP, outperforming Mascot [292]. The MS-GF+ results from unspecified cleavage (thermolysin), were then screened for peptides which contained >1 missed cleavage sites, based on the enzyme chemistry for thermolysin according to the Expasy peptide cutter tool (http://web.expasy.org/peptide_cutter/peptidecutter_enzymes.html). The R package 'cleaver' [597] from Bioconductor was used to identify which peptides contained >1 missed cleavage, and these peptides were then removed from further downstream analysis.

7.2.4 Dataset processing

The Wheat MIPS version 2.2 protein sequence FASTA file and GFF file required formatting into a compatible format for proteogenomics analysis, as outlined in Section 3.4.1. In particular, the exon coordinates in the GFF file were in the wrong orientation on the reverse strand, in a 5' to 3' direction, and were subsequently corrected to a 3' to 5' orientation using an in-house script.

The total 85,437 MS/MS spectra from all datasets obtained for this study, were first assessed by searching against the known proteome, examining the effects of using MS-Cluster to cluster the MS/MS spectra, PepNovo to quality filter the MS/MS spectra, and an assessment of optimal precursor mass tolerances, as outlined in Section 3.4.6. Since all the MS/MS spectra were of high-accuracy, derived from a QTOF and also an LTQ Orbitrap mass spectrometer, this factor needed to be reflected in the search

parameters. Therefore, assessment of the optimal precursor mass tolerance, involved using a range of tolerances, 0.5, 1.0, up to 10.0 in 1.0 ppm increments, and then up to 150.0 ppm in 5.0 ppm increments (Appendix File 7.2).

All RNA-seq data were pre-processed for quality and aligned to the wheat genome as detailed in Section 3.4.3. The resulting alignment BAM files were then merged and used to generate a splice graph FASTA database. A six-frame translation of the genome was also created for proteogenomics analysis. The methods used for both splice graph and six-frame translation generation are outlined in Sections 3.4.4 and 3.4.5, respectively. When running a MS/MS database search using the larger precursor mass tolerances for trypsin and chymotrypsin digests, as outlined below in Section 7.2.5, smaller file sizes for the six-frame translation, of around 50 MB, were needed to keep the MS/MS database search within the walltime.

7.2.5 Proteogenomics pipeline

The proteogenomics pipeline used a two-pass search approach, combined with an improved two-stage FDR strategy as outlined in Section 3.5. The choice of not clustering the MS/MS spectra and not quality filtering with PepNovo was selected due to the small dataset size and the significant loss of MS/MS spectra when quality filtering. Based on the preliminary assessment outlined in Section 7.2.4, 75.0 ppm, 110.0 ppm, 5.0 ppm, 10.0 ppm and 5.0 ppm precursor mass tolerances were chosen for the wheat flour trypsin digest, chymotrypsin digest, thermolysin digest, meiotic tissue trypsin digest, and meiotic tissue AspN digests, respectively.

Using MS-GF+ on a cluster, the MS/MS spectra from each dataset were searched against the known proteome, six-frame translated genome and splice graph, and further processed using the two-pass search approach, and then the improved two-stage FDR strategy as outlined in Section 3.5.1 and 3.5.2, respectively.

The choice of parameters for the proteogenomics pipeline, as outlined in Section 3.5, included a minimum event probability for novel genes, distal events and proximal events of 90%, a peptide linkage distance of 8,500 bp representing ≥95% of gene sizes in the Wheat MIPS version 2.2 annotation, a minimum cluster size of 1, and a minimum of 1 unique peptide per cluster.

The annotation events were further screened, as outlined in Section 3.5 and taking the same approach as in Section 6.2.5, accepting novel gene and distal events with ≥2 unique peptides and/or ≥99.9% event probability and proximal events filtered with an event probability of ≥99.8%. Single unique peptide annotation events identified outside these thresholds were screened against a protein repository before acceptance. However, the protein repository used in this study was NCBI NR, instead of NCBI RefSeq protein as was used in the previous chapter, because no wheat proteins (apart from mitochondrial and chloroplast) were represented in the highly curated RefSeq protein repository. Therefore, the NR protein repository was used to screen against with a higher level of specificity to account for the lack of a curated protein database. The outliers were identified through BLASTP searches against bread wheat in NR, with 100% query coverage, no mismatches, and an E-value of at least 1E-03, with any peptides matching required to have a length of >10aa. The findings outlined later in Section 7.3 also used BLASTP searches against bread wheat in NR to identify supporting evidence for discussion.

7.2.6 Improving gene predictions

Once the novel annotations were filtered and reviewed, the gene prediction tool Augustus [102], was used to improve the overall gene models of the Wheat MIPS version 2.2 annotation. Augustus was first trained with 6,137 filtered FL-cDNA from Komugi-TriFLDB (http://trifldb.psc.riken.jp/download.pl) and 1,286,040 EST sequences from NCBI to generate a base gene model. The training was kindly

conducted by Stefanie Koenig, from Mario Stanke's group (http://bioinf.uni-greifswald.de/bioinf/group/), the developer of Augustus. The novel annotation events, reference Wheat MIPS version 2.2 annotations and other additional extrinsic hints were then used as hints during gene prediction. Stefanie generated the additional extrinsic hints by: 1) masking the wheat genome using RepeatMasker [576] and generating repeat hints; 2) aligning 6,137 filtered FL-cDNA from Komugi-TriFLDB and all available EST sequences from NCBI, to the masked wheat genome using BLAT [122]; and 3) aligning all RNA-seq reads available from the URGI web site (http://wheat-urgi.versailles.inra.fr/Seq-Repository/RNA-Seq) using Bowtie2/Tophat2 [126]. The Augustus gene prediction tool was then run on a compute cluster, as mentioned in Section 3.4.3, across all chromosome fragments, split into 10 MB sizes to improve throughput in a highly parallel manner, and using parameters as outlined in Section 3.5.3, except that Augustus version 3.03 was used, and parameters "UTR" and "singlestrand" were set to on and off, respectively. The need to change some of the applied parameters arose due to a different version of Augustus (v3.03), which allowed for UTR predictions in this instance but not on both strands independently, which may have also been due to specific features of the Augustus wheat gene model prepared by Stefanie.

7.3 RESULTS AND DISCUSSION

The present study outlined improvements to the Wheat MIPS version 2.2 annotation of *Triticum aestivum* cultivar Chinese Spring, demonstrating the benefits of proteogenomics by integrating –omics datasets from genomics, proteomics and transcriptomics. Primarily, the study identified 189 (187 exclusively) novel annotation events, value-adding to the previous proteogenomics annotation by the dissertation author [12], outlining the benefits of performing a two-pass search approach with two-stage FDR strategy, suitable for very large plant genomes, and also highlighting the

challenges for proteogenomics with fragmented genomes.

7.3.1 Evaluation of pre-processing MS/MS spectra

Prior to running the proteogenomics pipeline the MS/MS spectra were evaluated for the optimal pre-processing strategy and precursor mass tolerance (Appendix File 7.2). It was decided that clustering should not be used due to the small dataset size, with >100,000 MS/MS spectra being ideal, while quality filtering negatively impacted on each of the MS/MS spectral datasets, resulting in the majority of the MS/MS spectra being filtered out, which would have resulted in very few or no annotation events being identified in the downstream analysis.

For the trypsin-digested wheat flour dataset, quality filtering reduced the peptide FDR after an initial 1% peptide-spectrum match (PSM) FDR filtering from 2.99% to 1.88%, and the protein FDR from 11.59% to 3.8%, at the most stringent PepNovo quality score of 0.2 (Appendix Figure 7.1B-C). However, as can be seen in Appendix Figure 7.1A, the total number of MS/MS spectra lost after quality filtering ranged from 15.52% at the lowest end to 62.33% at the most stringent cut-off. Applying scores between 0.05 – 0.1, as recommended by PepNovo (detailed in the Help File bundled with the tool), resulted in around 43.57% - 52.73% of the MS/MS spectra being lost. Approximately half of the MS/MS spectra were lost when using a mid-range score threshold, while the peptide FDR was maintained at around 2% - 3% with no improvements noticed. In addition, the number of MS/MS spectra, PSMs, and unique peptides dropped dramatically from 11,334 to 4,270, from 1,667 to 213, and from 535 to 106, respectively, at the highest stringency (Appendix Figure 7.1A, D-E).

For the chymotrypsin-digested wheat flour dataset, quality filtering reduced the peptide FDR after an initial 1% PSM FDR filtering from 2.44% to 0.00%, probably due to inaccuracies because of the small peptide numbers, and the protein FDR from 5.55%

to 0.00%, also likely due to inaccuracies with small protein numbers, at the most stringent PepNovo quality score of 0.2 (Appendix Figure 7.2B-C). However, as can be seen in Appendix Figure 7.2A, the total number of MS/MS spectra lost after quality filtering ranged from 10.61% at the lowest end to 74.49% at the most stringent cut-off. Applying scores between 0.05 – 0.1, as recommended by PepNovo, resulted in around 53.56% - 64.36% of the MS/MS spectra being lost. Approximately half of the MS/MS spectra were lost when using a mid-range score threshold, while the peptide FDR was seen to vary widely. In addition, the number of MS/MS spectra, PSMs, and unique peptides dropped dramatically from 14,799 to 3,775, from 213 to 53, and from 41 to 6, respectively, at the highest stringency (Appendix Figure 7.2A, D-E).

For the thermolysin-digested wheat flour dataset, which required MS/MS searching with 'unspecific cleavage', quality filtering reduced the peptide FDR after an initial 1% PSM FDR filtering from 6.98% to 3.30%, and the protein FDR from 19.51% to 7.14%, at the most stringent PepNovo quality score of 0.2 (Appendix Figure 7.3B-C). However, as can be seen in Appendix Figure 7.3A, the total number of MS/MS spectra lost after quality filtering ranged from 11.77% at the lowest end to 71.88% at the most stringent cut-off. Applying scores between 0.05 – 0.1, as recommended by PepNovo, resulted in around 51.35% - 61.56% of the MS/MS spectra being lost. Approximately half of the MS/MS spectra were lost when using a mid-range score threshold, while the peptide FDR was seen to increase and then drop significantly at the higher quality score cut-offs. In addition, the number of MS/MS spectra, PSMs, and unique peptides dropped dramatically from 16,776 to 4,718, from 1,036 to 183, and from 129 to 30, respectively, at the highest stringency (Appendix Figure 7.3A, D-E).

For the trypsin-digested wheat meiotic tissue dataset, quality filtering reduced the peptide FDR after an initial 1% PSM FDR filtering from 1.25% to 0.8%, and the protein FDR from 2.6% to 0.97%, at the most stringent PepNovo quality score of 0.2

(Appendix Figure 7.4B-C). However, as can be seen in Appendix Figure 7.4A, the total number of MS/MS spectra lost after quality filtering ranged from 63.43% at the lowest end to 94.42% at the most stringent cut-off. Applying scores between 0.05 – 0.1, as recommended by PepNovo, resulted in around 88.74% - 91.78% of the MS/MS spectra being lost. More than three quarters of the MS/MS spectra were lost when using a mid-range score threshold, while the peptide FDR was maintained at around 1%, with no improvements noticed. In addition, the number of MS/MS spectra, PSMs, and unique peptides dropped dramatically from 42,528 to 2,371, from 5,324 to 258, and from 4,250 to 243, respectively, at the highest stringency (Appendix Figure 7.4A, D-E).

For the AspN-digested wheat meiotic tissue dataset, a small sub-set of the trypsin digest dataset, quality filtering reduced the peptide FDR after an initial 1% PSM FDR filtering from 0.91% to 0.00%, probably due to inaccuracies because of the small peptide numbers, and the protein FDR from 1.25% to 0.00%, also likely due to inaccuracies associated with small protein numbers, at the most stringent PepNovo quality score of 0.2 (Appendix Figure 7.5B-C). However, as can be seen in Appendix Figure 7.5A, the total number of MS/MS spectra lost after quality filtering ranged from 66.64% at the lowest end to 95.73% at the most stringent cut-off. Applying scores between 0.05 – 0.1, as recommended by PepNovo, resulted in around 90.28% - 92.97% of the MS/MS spectra being lost. Almost all of the MS/MS spectra were lost when using a mid-range score threshold, while the peptide FDR was essentially 0% due to the small number of peptides. In addition, the number of MS/MS spectra, PSMs, and unique peptides dropped dramatically from 5,392 to 230, from 116 to 8, and from 110 to 8, respectively, at the highest stringency (Appendix Figure 7.5A, D-E).

These results indicated that the datasets were not improving as further MS/MS spectra were removed. Also, any apparent improvements, such as reduced peptide FDR, were inaccurate because the small numbers made any statistical analysis difficult, and

any numbers that appeared to improve were likely due to the significant losses of MS/MS spectra on such small datasets. Additionally, if any quality filtering was to be applied to these datasets, very few or no annotations would likely be identified further downstream in analysis due to the FDR inaccuracies from such small datasets.

The results indicated no clustering and no PepNovo quality filtering should be applied as clustering was not suitable for small MS/MS spectral datasets, and quality filtering reduced the dataset sizes significantly, resulting in too many lost MS/MS spectra and as a result poor accuracy of the statistical calculations required to ascertain any improvements to the dataset. Although no quality filtering was applied, as mentioned in Section 4.3.1, MS-GF+ applied a level of quality filtering based on log-likelihood ratios [292], and so sufficiently poor MS/MS spectra would still have been removed from the analysis to reduce potential false positives.

7.3.2 MS/MS database search parameter optimization

As outlined in Section 3.4.6, high-accuracy MS/MS spectra are well suited to precursor mass tolerance optimization. This held true for the present study, as high-accuracy MS/MS spectra were generated from a QTOF and LTQ Orbitrap mass spectrometer.

The original MS/MS spectra, with no clustering and quality filtering, were used to assess the precursor mass tolerances over a range, as outlined in Section 7.2.4 (Appendix File 7.2). From this analysis it was determined that for the trypsin-digested wheat flour, chymotrypsin-digested wheat flour, thermolysin-digested wheat flour, trypsin-digested meiotic tissue and the AspN-digested meiotic tissue, that the optimal precursor mass tolerance should be 75.0 ppm, 110.0 ppm, 5.0 ppm, 10.0 ppm, and 5.0 ppm, respectively (Appendix Figures 7.6, 7.7, 7.8, 7.9 and 7.10). After ≤1% PSM FDR filtering the following number of PSMs and peptide FDRs were obtained for each dataset. For the trypsin-digested wheat flour the maximum number of PSMs obtainable

was 2,755 at 75.0 ppm, while the peptide FDR was 3.5%. For the chymotrypsin-digested wheat flour the maximum number of PSMs obtainable was 713 at 110.0 ppm, while the peptide FDR was 4.49%. For the thermolysin-digested wheat flour the maximum number of PSMs obtainable was 274 at 5.0 ppm, while the peptide FDR was 4.08%. For the trypsin-digested meiotic tissue the maximum number of PSMs obtainable was 5,352 at 10.0 ppm, while the peptide FDR was 1.2%. For the AspN-digested meiotic tissue the maximum number of PSMs obtainable was 126 at 5.0 ppm, while the peptide FDR was 0.8%.

7.3.3 Proteogenomics pipeline

A proteogenomics pipeline was customised using Enosi with MS-GF+, as outlined in Section 3.5, and illustrated in Figure 3.1. As in Chapter 6, for each MS/MS spectral dataset that was processed a two-pass search approach with improved two-stage FDR strategy was used.

As was similarly applied in previous chapters, key variables for the proteogenomics pipeline were chosen. The peptide linkage distance was chosen based on the size of ≥95% of genes in the Wheat MIPS version 2.2 genome annotation, which was found to be 8,500 bp. As previously determined, a fixed peptide linkage distance brings with it problems when defining annotation events. However, in this study, due to the highly fragmented genome the peptide linkage distance now overshot the size of many smaller scaffolds containing genes, which could further contribute to misidentified annotation event types.

Contamination was also not an issue in this study, as compared with Chapter 5, with no noticeable over-abundance of peptides matching chloroplast and/or mitochondrial proteins. In all MS/MS spectral datasets, digests were obtained from 1D and 2D gels, acting as a filtering step and improving the abundance of target proteins

within a given mass range. No fractionation of samples by tissue or cellular components was conducted, except for filtration by 1D and 2D gels. However, this filtration resulted in significant losses, as shown from the limited number of MS/MS spectra obtained, but was partially compensated for by using different tissues (flour and meiotic tissue) and proteases in an attempt to improve the peptide coverage.

As applied previously in Section 7.2.5, the screening of novel and unique peptides in the novel annotation events was performed by accepting annotation events, that contained ≥2 unique peptides, and/or more stringent event probabilities, with single unique peptide outliers identified from homology searches against bread wheat in NR (Appendix Files 7.3 and 7.4). It was found that many accepted annotation events above the event probability/parsimony threshold also matched well to wheat proteins in NR. The majority of the outliers were found to just fall outside the event probability/parsimony thresholds applied, implying that for some annotation event types, further refinement of the applied thresholds could improve annotation event identification rates without relying so much on homology searches for their identification. This method is thus a viable approach to help define appropriate event probability thresholds when lacking a direct method to determine the annotation event FDR for a given event probability. The final minimum event probabilities after the applied filtering were 95.284% for novel genes, 95.615% for distal events and 98.167 for proximal events.

Similarly, as outlined in Chapter 6, due to the multiple sources of MS/MS spectra and RNA-seq data used in this analysis, with some being derived from different cultivars, there was the possibility of a number of variant sequences being misinterpreted as post-translational modifications (PTMs), or the false identification of novel peptides in incorrectly identified locations, both of which would affect the calculated event probabilities. These possibilities could be accounted for by including

and identifying known variants in the splice graph, as was previous demonstrated [474]. However, as mentioned in previous chapters, this approach only became known late in the study and so it was not pursued. The approach could, however, be used in a later wheat proteogenomics studies to compare variants from different cultivars that could account for differences in traits and gene expression profiles. Another source of false positives could be derived from the poor coverage of the assembly, which was more of an issue in this study, and could be addressed via a number of possible approaches, as previously mentioned in other chapters. For example: de novo sequencing, matching MS/MS spectra against closely homologous sequences, or modifications to approaches such as template proteogenomics.

Another caveat was that due to the highly fragmented nature of the wheat genome, a number of genes probably span across multiple scaffolds in the assembly, thus artificially inflating the number of identified gene regions, which could also hamper attempts to identify novel and unique coding regions. Therefore, revisiting a proteogenomics annotation as the draft assembly improves is an important consideration in future studies, as the status of unique/shared peptides and the peptide linkage distance of annotation events and consequently their event probabilities may well change. Additionally, in the future, adding a larger source of MS/MS spectra and RNA-seq data with a focus on a single cultivar, as well as specifically targeting the identification of variants between cultivars, will further improve the annotation and improve the event probabilities of annotation events that fell below the set thresholds in the present study.

During the screening of annotation events containing single novel and unique peptides no obvious correlation could be found with spectral counts as the event probability increased. This was probably due to the relatively small number of MS/MS spectra used. A number of novel and unique peptides had very high spectral counts compared to other annotation events (Appendix File 7.3) and so were considered as

possible contaminants from chloroplast and/or mitochondria, as was found in Chapter 5. However, after screening of a number of these particular unique and novel peptides, no evidence could be found to indicate that these were indeed peptides derived from chloroplasts and/or mitochondria. The novel peptides with high spectral counts were mostly distributed amongst the novel gene and distal events, where false positive rates are often higher and may also be due to the draft genome being highly fragmented, with a higher proportion of MS/MS spectra being misinterpreted to the same sequence, due to the limited coverage and also possibly because no MS/MS spectral quality filtering was applied.

7.3.4 Proteogenomics analysis

A total of 6,834 novel peptides were identified with an event probability of ≥90%, at least 1 unique peptide per cluster and a minimum cluster size of 1, using the two-pass search approach with improved two-stage FDR strategy. After filtering the novel annotation events as detailed in Section 7.2.5, there were a total of 290 novel peptides remaining (Table 7.1 and Appendix Files 7.3 and 7.4).

The final 290 novel peptides led to 189 novel annotation events (187 exclusively) in total amongst 96 genes (96 exclusively) and 189 proteins (187 exclusively) from the Wheat MIPS v2.2 annotation. The novel annotations along with the Wheat MIPS v2.2 reference annotation, and the extrinsic hints as generated by Mario Stanke's group outlined in Section 7.2.6, were then used as hints for Augustus gene prediction. A total of 413,587 genes and 426,007 proteins (≥66 aa in length) were predicted (Appendix File 7.5), and of these, 70 predicted proteins had 180 novel peptides incorporated (Table 7.1), of which 80 novel peptides were unique and identified in 49 of the 70 predicted proteins (Appendix File 7.6). The number of protein-coding genes and proteins predicted by Augustus was far higher than the original reference Wheat MIPS v2.2 predictions (Table 7.1). In addition to the new predictions

previously not identified, the high number of predictions could also be attributed to two factors, as outlined in Section 3.5.3.

The number of novel peptides incorporated into the predictions was reasonably high at 180 novel peptides (62%), but relatively lower compared to the previous two chapters by about 20%, probably due to some genes which were unable to be predicted within short scaffolds, as well as the relatively lower number of MS/MS spectra due to the limited dataset size and absence of MS/MS spectra quality filtering. Of the 180 incorporated novel peptides 20 were exclusively derived from the splice graph, 62 were identified in both the six-frame translation and splice graph, with the remaining 98 identified exclusively in the six-frame translation.

As demonstrated previously in Section 6.3.4, BLASTP analysis was performed to show how the predictions changed from the reference predictions. This was undertaken by searching all 426,007 Augustus predicted proteins (Table 7.1) against the Wheat MIPS v2.2 proteins, taking the top match, with E-value ≤1E-10. Any sequences that did not match were considered novel predictions, sequences that had a query coverage ≥95% with at least 1 mismatch were considered to be the same prediction as the reference protein, and the remaining matches were considered to be modified predictions, either due to Augustus predicting slightly different models or as a direct result of the supporting evidence. From this analysis, there were 257,341 non-paralogous novel protein predictions, 111,004 modified predictions and 57,662 predictions considered to be essentially the same as the reference.

Searching all 70 protein predictions that had the novel peptide evidence incorporated, against the Wheat MIPS v2.2 proteins, taking the top match with E-value ≤1E-10, identified 46 protein predictions likely to be modified predictions, leaving 24 protein predictions, that found no match and were considered as non-paralogous novel

protein predictions (Table 7.1).

Based on the annotation events incorporated into the Augustus gene predictions, the minimum event probabilities which led to a new Augustus gene prediction were: novel gene event 98.785%, gene boundary 99.639%, reverse strand event 99.899%, translated UTR and frame-shift 99.8%. No filtered annotation events from exon boundaries and novel exons could be incorporated into the Augustus gene predictions.

Of the 290 novel peptides identified in this study, 127 were derived from the trypsin-digested wheat flour, of which 86 were included in new predictions, with 41 unique peptides. A total of 97 peptides were derived from chymotrypsin-digested wheat flour, of which 68 were included in new predictions, with 19 unique peptides. A total of 3 peptides were derived from thermolysin-digested wheat flour, of which 2 were included in new predictions, 1 of which was a unique peptide. A total of 55 peptides were derived from trypsin-digested meiotic tissue, of which 20 were included in new predictions, 16 of which were unique. A total of 9 peptides were derived from AspN digested meiotic tissue, of which 4 were included in new predictions with 3 unique. One novel peptide was identified in both trypsin-digested wheat flour and trypsin-digested meiotic tissue.

In the initial study 16 novel peptides were identified in 33 novel annotation events [12]. Of these 5 novel peptides were re-identified from the 290 identified in this study, 127 of which were from the trypsin-digested wheat flour. This demonstrated that the present study was able to significantly improve upon the previous work, even though in this study tighter event probability stringencies were applied. Including the novel peptides derived from the trypsin-digested wheat flour which were not re-identified in this study, overall this new study demonstrated a ~8x improvement in the novel peptide identification rate. Overall, compared to the initial study, a total of 156

additional novel annotation events were identified.

The reason why an additional 122 novel peptides were identified from the trypsin-digested wheat flour in this study was probably due to a number of factors: 1) the use of MS-GF+ over InsPecT; 2) a tighter precursor mass tolerance (75.00 ppm versus 2.0 Da); 3) the two-pass search approach with improved two-stage FDR strategy versus the combined FDR strategy; and 4) the application of a later version of the Wheat MIPS annotation (November 2013 version used in the previous proteogenomics study [12] versus July 2014 (v2.2) for this study).

The previous study [12] validated 50 high confidence proteins, consisting of 152 peptides, whereas in this study 107 high confidence proteins were identified consisting of 557 peptides (Table 7.1). Of the 557 peptides that were mapped to 107 high confidence proteins, 174 peptides were derived from trypsin-digested wheat flour, and of these 54 peptides were re-identified from the 50 validated proteins from the previous study. The differences in identification could be attributed to the same reasons for the novel peptides mentioned above, as well as the difference in method used to identity high confidence proteins. In the previous study, the protein probability was used based on the product of the local FDR of mapped peptides, whereas in this study the use of protein probability was discarded from the proteogenomics pipeline in favour of simply identifying proteins containing at least 2 peptides with at least 1 unique peptide.

Table 7.1 Summary of wheat proteogenomics annotations

The results of the proteogenomics analysis of the Wheat MIPS v2.2 annotation, using the two-pass search approach with improved two-stage FDR strategy.

Total Wheat MIPS v2.2 genes	99,386
Total 'known' protein-coding genes	99,386
Total 'known' proteins	293,053
Raw MS/MS search 'known' protein matches ≤1% PSM FDR	1,970
Proteogenomics mapping: Total 'known' proteins ≤1% PSM FDR	16,635
Proteogenomics mapping: Total 'known' proteins ≤1% PSM FDR (≥2 peptides with 1 unique)	107
Total identified 'novel' peptides ≤1% PSM FDR	14,151
Raw MS/MS search 'known' peptides ≤1% PSM FDR	4,471
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR	4,474
Proteogenomics mapping: Total identified 'known' peptides ≤1% PSM FDR (≥2 peptides with 1 unique)	557
Frame-shifts	46
Translated UTRs	9
Exon boundaries	17
Novel splices	0
Novel exons	39
Gene boundaries	17 (15)
Reverse strands	24 (24)
Novel genes	37
Total annotation events	189 (187)
Total genes affected	96 (96)
Total proteins affected	189 (187)
Total novel peptides in affected genes/proteins	290
Total Augustus protein-coding gene predictions	413,587
Total Augustus protein predictions	426,007
Total Augustus gene predictions with incorporated novel peptides	67
Total Augustus protein predictions with incorporated novel peptides	70
Total novel peptides incorporated into Augustus protein predictions	180
Improved protein predictions with incorporated novel peptides	46
Novel non-paralogous protein predictions with incorporated novel peptides	24

Note: Numbers in parenthesis represent the exclusive numbers. The inflationary effect of a large peptide linkage distance on gene boundaries and reverse strands was removed by assigning a peptide cluster as either a proximal or distal event, not both, with preference placed on proximal events.

7.3.5 Novel gene annotation

A total of 37 novel genes (Table 7.1) were identified. An example of one of these novel genes was on chromosome 5BL, fragment 5BL10923515, spanning positions 6,935 to 8,434, with an event probability of 100%, consisting of 5 unique and 3 shared peptides and with 24 PSMs assigned, derived from trypsin-digested wheat flour protein. In addition, an Augustus gene prediction was carried out incorporating the novel peptides, exon, exon_part and intron hints, predicting 1 new gene model (Figure 7.1, and with a sample of 14 of 24 representative MS/MS spectra supporting annotated MS/MS spectra

in Appendix Figure 7.11).

Performing a BLASTP search against bread wheat in NR revealed that novel peptide “AFLQPSHYDADEIAFVR” matched an unnamed protein product (CDM82143.1 with E-value = 1.6), with 35% query coverage and 100% identity; novel peptide “ALAFPQQAR” matched gamma-gliadin (AGZ20266.1 with E-value = 1.6) with 88% query coverage and 75% identity; novel peptide “AQPESVFVAGPQQQR” matched globulin 3B (ACJ65515.1 with E-value = 0.17) with 53% query coverage and 88% identity; novel peptide “EGDVFVIPAGSIVYSANTHR” matched storage protein (AAA34269.1 with E-value = 6E-04) with 90% query coverage and 72% identity; novel peptide “EGEGVLVLLR” matched unnamed protein product (CDM83362.1 with E-value = 1.3) with 80% query coverage and 88% identity; novel peptide “GGGGGSGSEKEDIQPR” matched homeobox protein (BAH03543.1 with E-value = 0.96) with 50% query coverage and 100% identity; novel peptide “IYIVVEGR” matched CBL-interacting protein kinase 23 (AFR90218.1 with E-value = 5.0) with 75% query coverage and 83% identity; and novel peptide “SLTGEKPR” matched phenylalanine ammonia-lyase (AAA50849.1 with E-value = 6.5) with 87% query coverage and 86% identity. The Augustus gene prediction g265543 matched to storage protein (AAA34269.1 with E-value = 1E-54) with 94% query coverage and 38% identity. The most significant novel peptide “EGDVFVIPAGSIVYSANTHR” and the Augustus gene prediction g265543 both matched a storage protein. Many of these matches were also of low coverage, identity and E-value, indicating that the new gene prediction may be a truly novel gene, possibly a storage protein homolog or similarly functional protein.

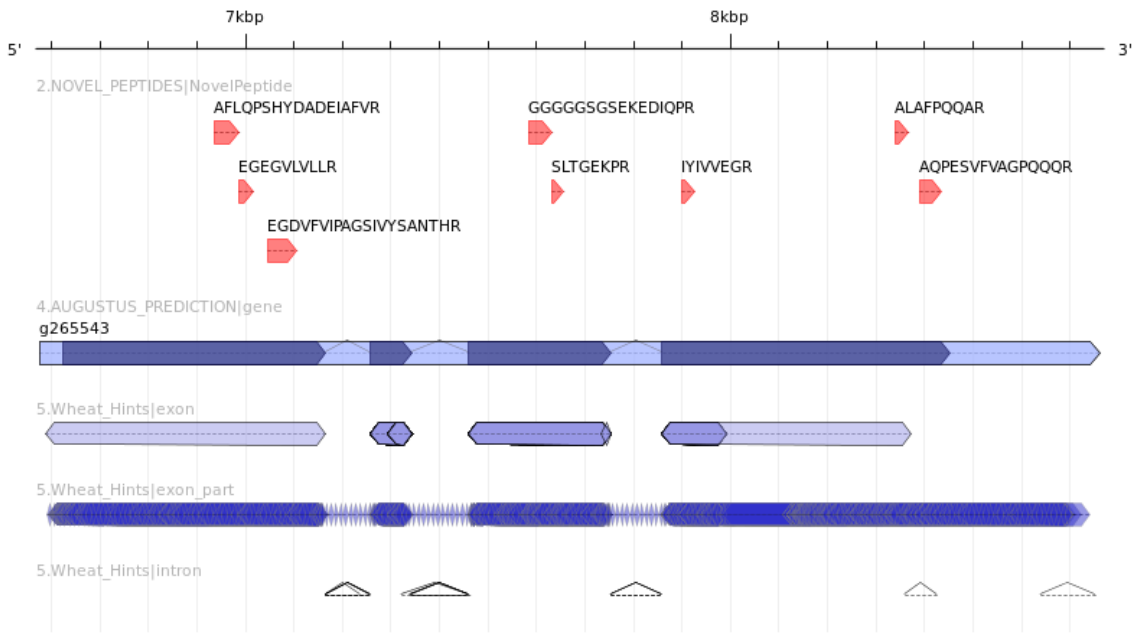


Figure 7.1 Novel gene annotation

A novel gene annotation event was located on chromosome 5BL, fragment 5BL10923515, and led to a new gene prediction. The novel peptides were incorporated into the Augustus predicted gene g265543. Extrinsic hints from EST, cDNA and RNA-seq evidence indicate exon, exon_part and intron hints, in agreement with the novel peptides and new prediction.

7.3.6 Gene boundary annotation event leads to a new gene prediction

There were 17 (15 exclusive) gene boundary annotation events identified (Table 7.1). An example of a gene boundary annotation event was on chromosome 1AL, fragment 1AL3923345, spanning positions 10,484 to 12,796, with an event probability of 100%, consisting of 8 unique and 14 shared peptides, and with 50 PSMs assigned, 38 derived from trypsin-, 11 derived from chymotrypsin- and 1 derived from thermolysin-digested wheat flour protein. The gene boundary event was identified for gene Traes_1AL_52FE5D716, protein-coding transcripts 2 and 3. In addition, an Augustus gene prediction was carried out incorporating the novel peptides, reference gene Traes_1AL_52FE5D716 and exon_part and intron hints, predicting two new gene models. One gene was an identical prediction to the reference gene Traes_1AL_52FE5D716, whereas the other gene was a completely novel prediction in close proximity to the reference gene Traes_1AL_52FE5D716 prediction (Figure 7.2, and with a sample of 25 of 50 supporting annotated MS/MS Appendix Figure 7.12).

Performing a BLASTP search against bread wheat in NR revealed the 22 novel peptides matched to a high molecular weight (HMW) glutenin (CAC84119.1, P02861.1, CAC84121.1, CAC40684.1, CAC83002.1, AAA62315.1, CAE00624.1 and BAN29068.1 with E-value range: 0.086 – 2E-24) with 100% query coverage and identity. The two reference protein-coding transcripts 2 and 3 matched a protein kinase (ABG68041.1 with E-value = 0.0) with 100% query coverage and 95% identity, and an unnamed protein product (CDM85023.1 with E-value = 3.3) with 30% query coverage and identity, respectively. The Augustus gene prediction g4797 matched a HMW glutenin (AHZ62762.1 with E-value = 0.0) with 100% query coverage and 96% identity, and gene g4798 matched a protein kinase (ABG68041.1 with E-value = 0.0) with 100% query coverage and 96% identity.

The novel peptide "QPGYYSTSPQQLGQGQPR" (Figure 7.2), which matched a HMW glutenin (BAN29068.1) was not incorporated into the prediction and was a shared peptide, probably derived from a different genomic region, and also probably from a HMW glutenin gene. The novel peptide and hints evidence (Figure 7.2) indicated that the gene boundary annotation identified was actually a novel gene annotation, and probably a HMW glutenin. No known peptides were identified for the reference gene Traes_1AL_52FE5D716, and both gene predictions g4797 and g4798 matched to two different proteins in NR, indicating that it was unlikely that these two genes were simply a fragmented Augustus gene prediction.

The misidentification of this annotation as a gene boundary event also highlighted again the difficulty of using a fixed value peptide linkage distance to conduct proteogenomics analysis. If the peptide linkage distance were significantly shorter (i.e. the distance between the peptide cluster and the neighbouring reference gene), this annotation would have been identified correctly as a novel gene instead of a gene boundary annotation. To improve the accuracy of the assignment of annotation

event types such as these, the determination of a peptide linkage distance through a dynamic approach using additional sources of evidence could be used.

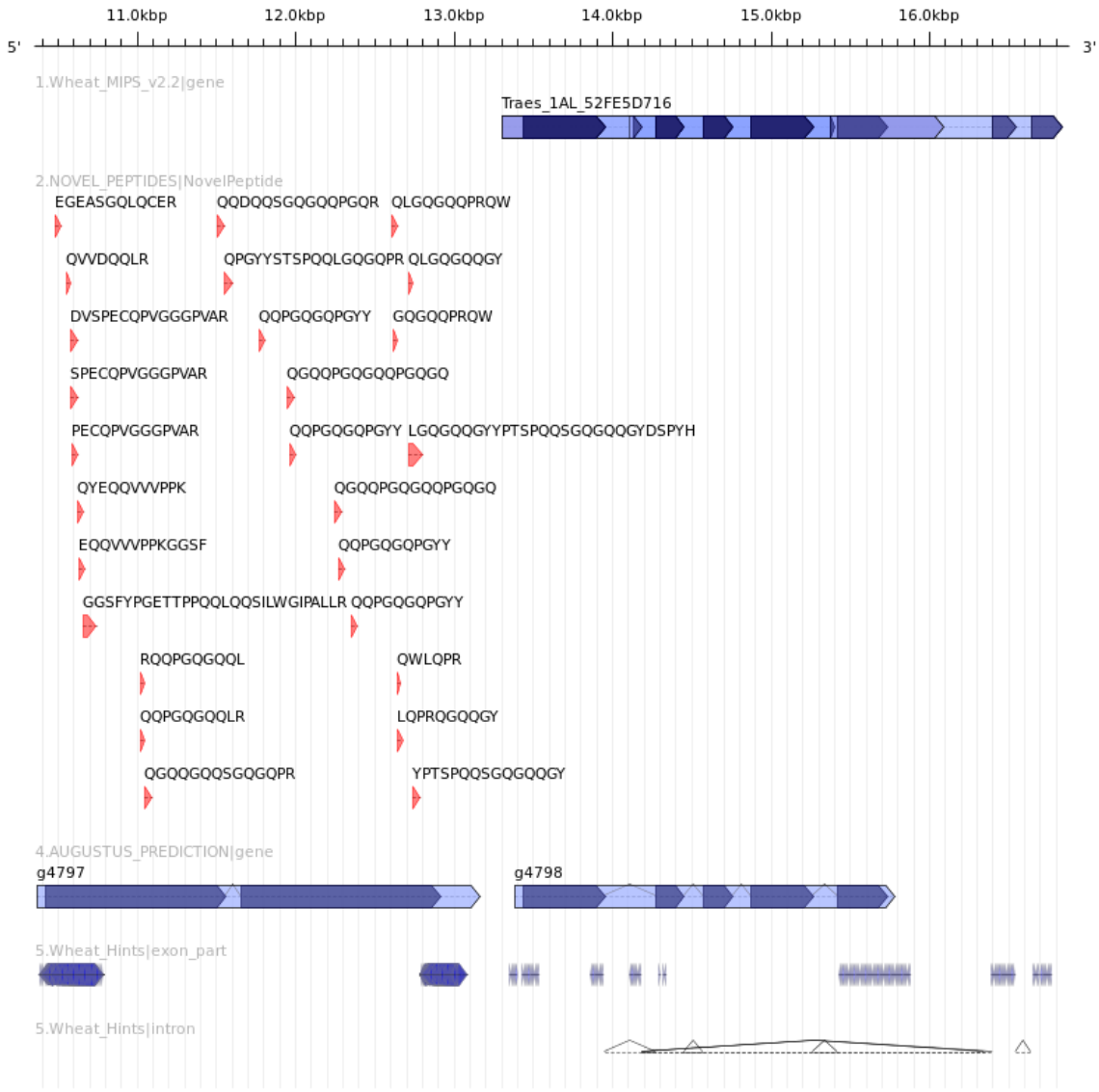


Figure 7.2 Novel gene annotation misidentified as a gene boundary event

A gene boundary annotation event that led to two gene predictions; g4798 in line with the original reference gene Traes_1AL_52FE5D716 and gene g4797 a new gene prediction in close proximity to the reference gene on chromosome 1AL, fragment 1AL3923345. The novel peptides and exon_part hints led to a novel gene prediction, while the reference Traes_1AL_52FE5D716 gene, exon_part, and intron hints led to a similar prediction to the reference. No known peptides were identified for the reference gene indicating that both gene predictions were unlikely to be simply fragmented predictions.

7.3.7 Reverse strand annotation

There were 24 (24 exclusive) reverse strand annotation events identified (Table 7.1). An example of a reverse strand annotation event was on chromosome 3AL, fragment 3AL4447768, spanning positions 2,753 to 2,968, with an event probability of 99.90%,

consisting of 1 unique and 2 shared peptides, and with 3 PSMs assigned, derived from trypsin-digested meiotic tissue. The reverse strand event was identified for gene Traes_3AL_8EF777719, and its single protein-coding transcript. In addition, an Augustus gene prediction was carried out, incorporating the novel peptides, reference gene Traes_3AL_8EF777719 and exon_part hints, predicting a new gene model on the opposite strand and removing the previous reference prediction. This indicated that the original reference gene was predicted on the wrong strand (Figure 7.3, and with 3 supporting annotated MS/MS spectra in Appendix Figure 7.13).

Performing a BLASTP search against bread wheat in NR revealed that the three novel peptides matched to an unnamed protein product (CDM87003.1 with E-value range: 1.1 – 6E-07) with 100% query coverage and percentage identity ranging between 92% and 100%. The single reference protein-coding transcript matched to wall-associated kinase-like 1 (AAY34779.1 with an E-value = 0.47) with 25% query coverage and 34% identity. The Augustus gene prediction g50683 matched unnamed protein product (CDM87003.1 with E-value = 3E-86) with 90% query coverage and 89% identity. This evidence indicated that the novel peptides and the new prediction found a better match to a *Triticum aestivum* protein than the reference protein from the Wheat MIPS v2.2 annotation, confirming the proteogenomics annotation was probably correct.

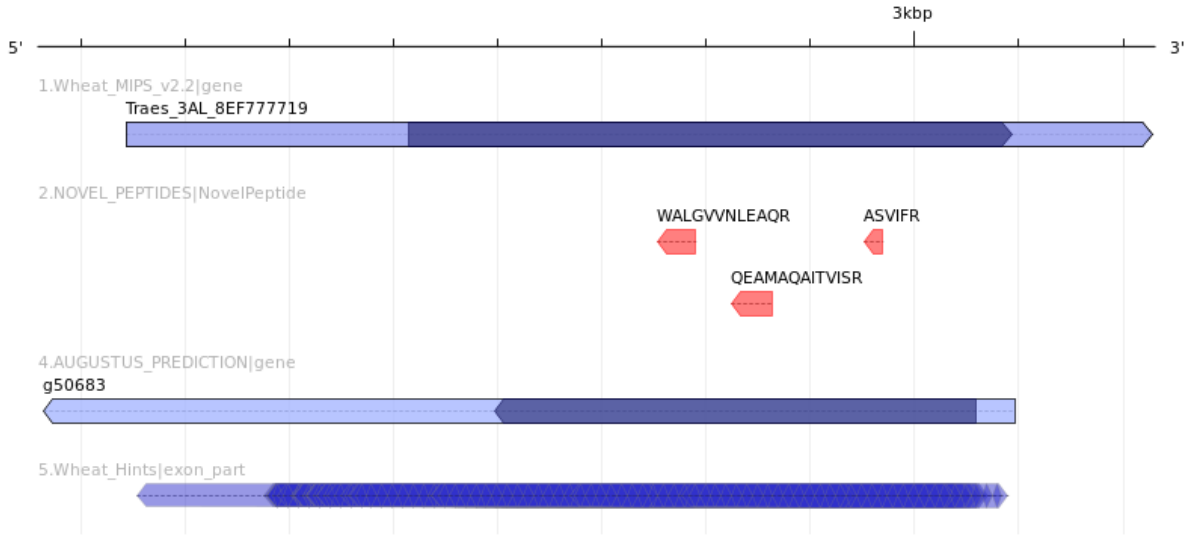


Figure 7.3 Reverse strand annotation

A reverse strand annotation event, located on chromosome 3AL, fragment 3AL4447768, where the original Traes_3AL_8EF777719 prediction was predicted on the forward strand. An Augustus gene prediction on the reverse strand was suggested with incorporated novel peptides and exon_part hints.

7.3.8 Translated UTR annotation hides an exon boundary event

There were 9 translated UTR annotation events identified (Table 7.1). An example of a translated UTR annotation event was on chromosome 1DL, fragment 1DL2289899, spanning positions 56 to 2,161, with an event probability of 100%, consisting of 3 unique and 16 shared peptides, and with 34 PSMs assigned, 5 derived from trypsin- and 14 derived from chymotrypsin-digested wheat flour protein. The translated UTR event was identified for gene Traes_1DL_757719220 and protein coding transcript 4, which also directly overlapped protein coding transcripts 2 and 3, both of which had a 41 bp region of ambiguous X amino acid residues at the 5' end of their protein predictions (Appendix File 7.1). In addition, an Augustus gene prediction was conducted incorporating the novel peptides, reference gene Traes_1DL_757719220 and exon_part and intron hints, predicting a gene model, and revised from the original reference prediction, replacing the intron from the reference gene with an exon region by incorporating the novel peptides (Figure 7.4, and with a sample of 25 of 34 supporting annotated MS/MS spectra in Appendix Figure 7.14).

Performing a BLASTP search against bread wheat in NR revealed that the novel peptides matched to a HMW glutenin (AAS67321.1, ABB05179.1, AEF32781.1, AEL99901.1, AHI62992.1, BAH96595.1, CAC40684.1 and P02861.1 with E-value range: 0.003 – 7E-21) with 100% query coverage and identity. The reference protein-coding transcript 4 matched protein kinase (ABG68032.1 with E-value = 2E-13) with 71% query coverage and 93% identity, while reference protein-coding transcripts 2 and 3 matched HMW glutenin (AAS67319.1 with E-value = 9E-55 and AAS67320.1 with E-value = 1E-46, respectively) with 46% query coverage and 100% identity, and 66% query coverage and 96% identity, respectively. The Augustus gene prediction g307507 matched HMW glutenin (P08489.1 with E-value = 0.0) with 99% query coverage and identity, with the 1% coverage in disagreement located at the 5'-most end of the prediction.

The novel peptides, reference protein-coding transcripts 2 and 3, and the revised prediction all matched HMW glutenin in NR, while the reference protein-coding transcript 4, which was identified specifically for this annotation event, matched protein kinase. In addition, all novel peptides and the 7 identified known peptides were within the same frame as transcripts 2 and 3 (frame +2), while transcript 4 was in a different frame (frame +3) (Figure 7.4). The revised prediction failed to predict the exon/CDS region of transcript 4, which also lacks support from the extrinsic exon_part and intron hints (Figure 7.4). Additionally, the 5' region of the gene prediction appeared truncated, with a different exon in an alternate frame (frame +1) and an intron, which conflicted with all the evidence, including the BLASTP results, which found inconsistencies with the 5' end of the match to a HMW glutenin. This is because there was only one Methionine translation initiation codon (ATG) located at the 5' end of the scaffold, in agreement with the Augustus gene prediction. The absence of any more translation initiation sites further upstream was probably because the actual gene overlapped the

edge of the scaffold. This could also explain why the original reference protein-coding transcripts 2 and 3 were truncated at the 5' end with a 41 bp ambiguous X protein sequence in their protein predictions, because the available evidence used for the Wheat MIPS v2.2 predictions overlapped the end of the scaffold and, instead of producing half a protein sequence, X residues were used instead to indicate incompleteness.

Although this annotation event was identified as a translated UTR event for transcript 4, it was also unreported by Enosi as an exon boundary event for transcripts 2 and 3. The annotation event did not completely agree with the hints evidence, however it is worth noting that many of the extrinsic hints were also originally used for the Wheat MIPS v2.2 annotation, and so should not be considered as complete evidence. An exon boundary event was never identified for transcripts 2 and 3 because the only identified unique peptides within the peptide cluster mapped within their 5' regions which, as described previously, contained a 41 bp region with ambiguous X amino acid residues and so were never identified as known peptides belonging to those transcripts. This is because, as described in Section 3.5, this analysis only accepts peptide clusters with at least 1 unique peptide. However, if the unique peptides were identified mapping to transcripts 2 and 3, they would have been identified as known peptides and therefore the annotation event would not have been identified. If, to allow detection of this annotation event, the analysis was to accept all annotation events with a minimum of zero unique peptides, then the annotation event would have been identified in this case. However, all possible events would have also been identified, including many which are ambiguous due to multiple locations across the genome, which would have inflated the FDR at the annotation event level.

Overall, the evidence was in agreement with transcripts 2, 3 and the revised Augustus gene prediction, but required a correction to exon 1 and further extension to the 5' region of the gene. Further extending the scaffold towards the 5' region would

improve the prediction and also resolve the 41 bp ambiguous X amino acid residue region of the original transcript 2 and 3 protein predictions, which should then reach full coverage and identity with the identified HMW glutenin.

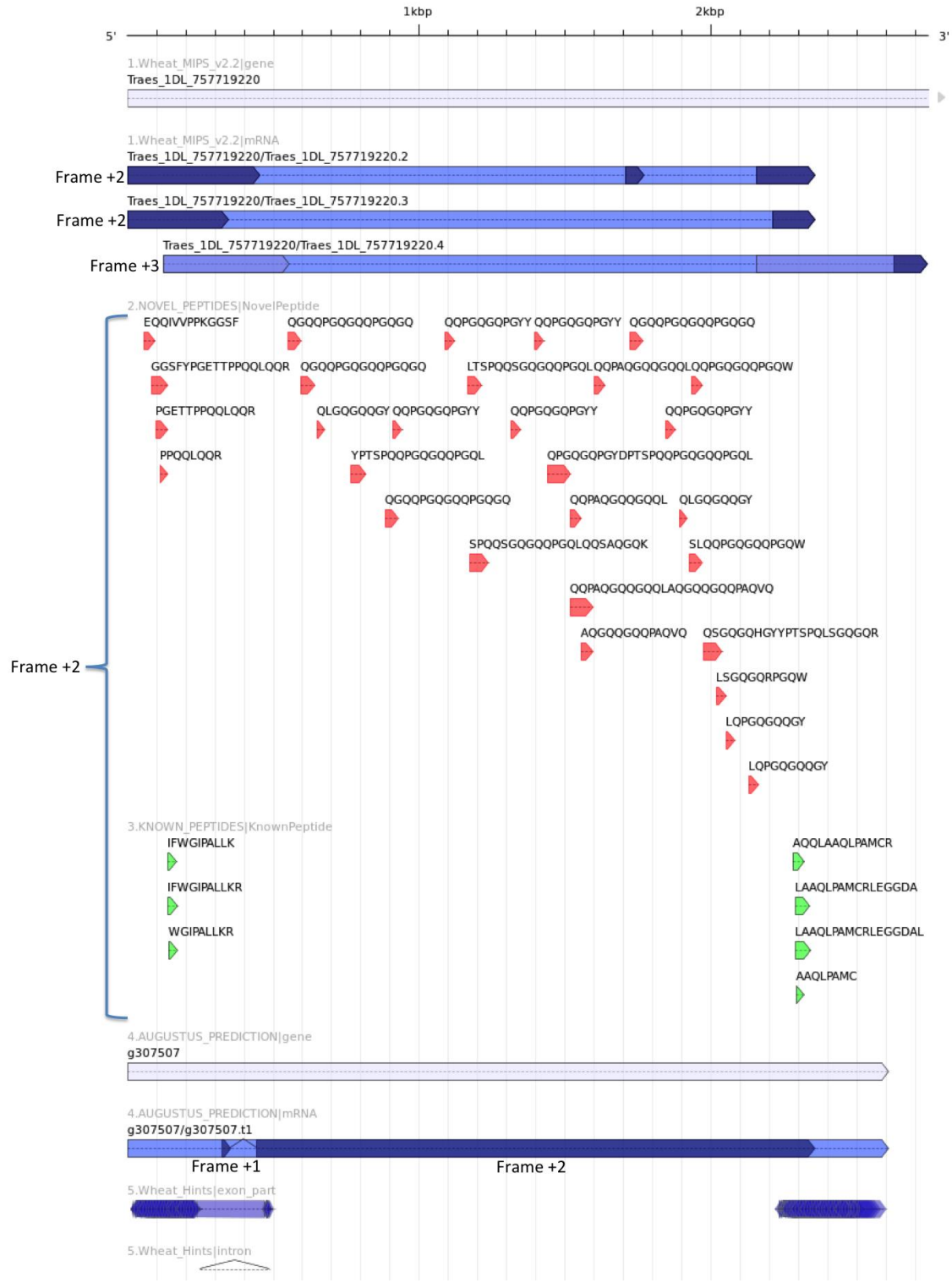


Figure 7.4 Exon boundary annotation via a translated UTR event

An exon boundary annotation identified via a translated UTR event, located on chromosome 1DL, fragment 1DL2289899. The novel peptides appeared within the untranslated UTR region of transcript 4 but within a different frame and they also appeared as an exon boundary for transcripts 2 and 3, within the same frame. Known peptides appeared mapped to both transcripts 2 and 3, within the same frame as the novel peptides. The evidence indicates an annotation event more in line with transcripts 2 and 3. There was also no hints evidence for many of the novel peptides spanning the intron region. The four 5'-most novel peptides map to a 41 bp ambiguous X amino acid region at the 5' end of transcripts 2 and 3 indicating incompleteness of the prediction. The revised Augustus gene prediction agrees with transcripts 2, 3 and the novel peptides.

7.3.9 Exon boundary annotation

There were 17 exon boundary annotation events identified (Table 7.1). An example of an exon boundary annotation event was on chromosome 6AL, fragment 6AL2841856, spanning positions 6 to 71, with an event probability of 99.899%, consisting of 1 unique and 1 shared peptide, and with 2 PSMs assigned, derived from trypsin-digested meiotic tissue. The exon boundary event was identified for gene Traes_6AL_895062375 and its single protein-coding transcript. In addition, an Augustus gene prediction was carried out incorporating the novel peptides, reference gene Traes_6AL_895062375 and exon_part hints. However, no Augustus gene prediction could be generated, regardless of the proteogenomics evidence and hints supplied. This was likely due to the small fragment size of 6AL2841856 which was just 315 bp in size, with only 3 Methionine translation initiation codons across the entire length of the fragment, and no translation initiation codons towards the 5' region of the original Traes_6AL_895062375 gene where the novel peptides were mapped. Also, in the absence of a suitable translation initiation codon, Augustus probably penalises any evidence and resulting prediction. Extending this fragment length further given the supporting evidence would probably see a prediction along the length of this chromosome fragment (Figure 7.5, and 2 supporting annotated MS/MS spectra in Appendix Figure 7.15).

Performing a BLASTP search against bread wheat in NR revealed the novel peptides matched to a putative pyruvate dehydrogenase (ADD73514.1 with E-value range: 0.002 – 2E-08) with 100% query coverage and identity. The single reference protein-coding transcript matched putative pyruvate dehydrogenase (ADD73514.1 with E-value = 1E-55) with 100% query coverage and 99% identity. The 1% discrepancy in identity was attributed to a region missing from the original prediction, of around 77 bp according to a BLASTP alignment to the putative pyruvate dehydrogenase in NR. All other exon boundary events identified in this study could not be incorporated into

Augustus gene predictions, some probably due to their small fragment sizes and/or missing appropriate translation initiation start (TIS) sites, while others were possibly due simply to false positive peptide identifications.

As pointed out with this annotation event, and the translated UTR event previously discussed in Section 7.3.8, the fragmented nature of the genome can be a limiting factor for gene prediction. In addition, many scaffolds making up the genome were smaller than the peptide linkage distance. This size constraint further limited the approach to recruit more genes and peptide clusters into annotation events, reducing the upper bounds for the identification of novel annotation events and also further increasing the chances of misidentifying an annotation event. As a result, it is important to revisit the proteogenomics analysis as future versions of the wheat genome assembly become available.

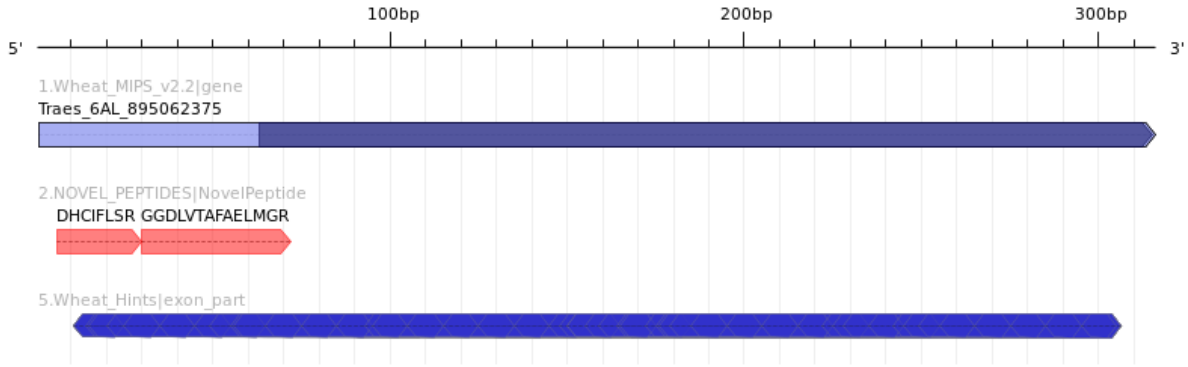


Figure 7.5 Exon boundary annotation

An exon boundary annotation event, located on chromosome 6AL, fragment 6AL2841856. Two novel peptides overlapped the exon boundary of gene Traes_6AL_895062375, which was supported by exon_part hints. This evidence, as well as the reference gene was used as hints for gene prediction. However, no prediction resulted, probably due to the short fragment length of 315 bp, resulting in a missing TIS site from further upstream needed for Augustus gene prediction. Later extensions of his fragment on chromosome 6AL will likely allow for a prediction, given the supporting evidence that is already available.

7.3.10 Frame-shift annotation

There were 46 frame-shift annotation events identified (Table 7.1). An example of a frame-shift annotation event was on chromosome 1AL, fragment 1AL3886502, spanning positions 4,108 to 4,197, with an event probability of 99.80%, consisting of 1

unique peptide, and with a single PSM assigned, derived from trypsin-digested meiotic tissue. The frame-shift event was identified for gene Traes_1AL_C8971FBEE and its single protein-coding transcript. In addition, an Augustus gene prediction was carried out incorporating the novel peptide, reference gene Traes_1AL_C8971FBEE, exon_part and intron hints, which predicted a gene model revised from the original reference prediction, by splitting the original prediction with 1 exon/CDS into a gene prediction with 2 exons/CDS, with the novel peptide incorporated into the first exon/CDS in a different frame to that that of the original (Figure 7.6, and a single supporting annotated MS/MS spectrum in Appendix Figure 7.16).

Performing a BLASTP search against bread wheat in NR revealed the novel peptide matched to an unnamed protein product (CDM84219.1 with E-value = 6.3) with 53% query coverage and 57% identity. The single reference protein-coding transcript matched unnamed protein product (CDM82578.1 with E-value = 1.5) with 16% query coverage and 29% identity. The Augustus gene prediction g2008 matched unnamed protein product (CDM81340.1 with E-value = 0.026) with 36% query coverage and 41% identity.

The novel peptide, reference protein-coding transcript and gene prediction g2008 all matched different unnamed protein products. However, of significant note was that the E-value, query coverage and percentage identity of gene prediction g2008 were better than that found with the reference protein-coding transcript. One shared known peptide was mapped to the reference protein-coding transcript and also appeared in the g2008 gene prediction. Additionally, as can be seen in Figure 7.6, many of the hints appeared located around the two new exons, more so around exon 2 which can be seen when the exon_part hints are not stacked on one another indicating potential splicing patterns, even though there were no intron hints to indicate definitive evidence. Many of the other 45 frame-shift events also contained single unique peptides, all with

the same event probability and most with the same or similar number of PSMs identified.

Of the remaining 45 frame-shift events, 27 were incorporated into predictions. These were visually inspected and appeared to conflict with the available hints. Also, in some instances a number of known peptides mapped to the same region as the frame-shift novel peptide, indicating a direct conflict within the proteogenomics evidence itself. This demonstrated that careful consideration of all the provided evidence, particularly with the frame of mapped peptides, is needed before acceptance of a new annotation and the inferred prediction, and that the acceptance of an annotation event should never solely rely on the confidence of the event probability alone, particularly when the evidence is only in the form of a single unique peptide.

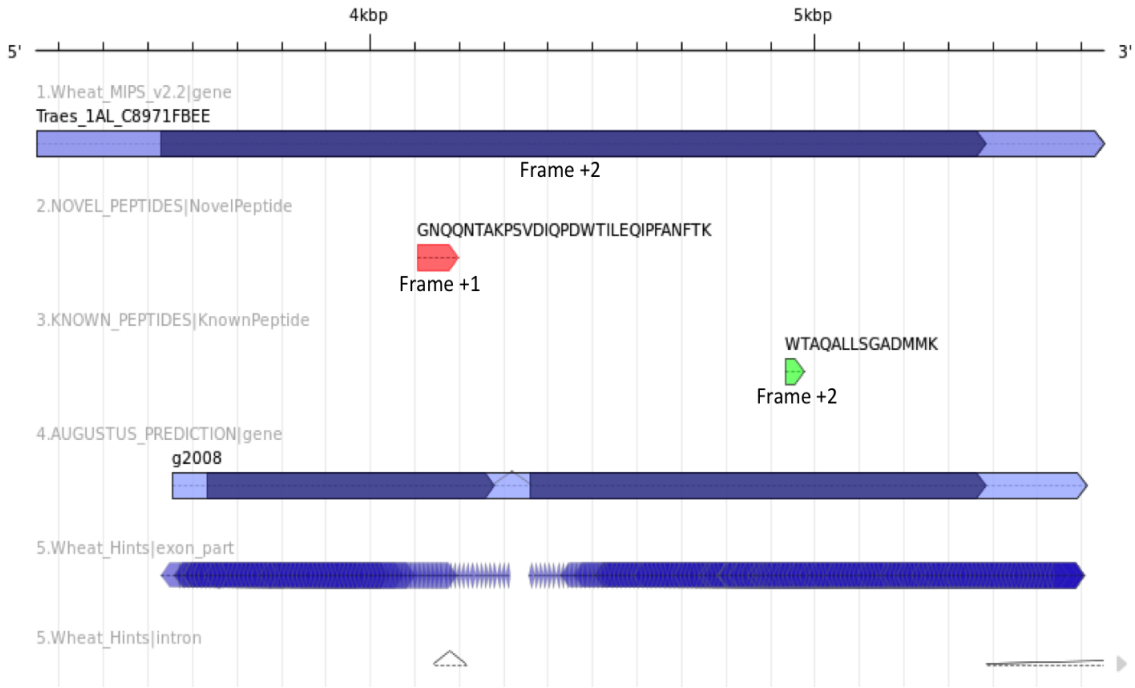


Figure 7.6 Frame-shift annotation

A frame-shift annotation event located on chromosome 1AL, fragment 1AL3886502. A novel and unique peptide was identified in a different frame to reference gene Traes_1AL_C8971FBEE. The novel peptide, reference gene Traes_1AL_C8971FBEE, and exon_part and intron hints were incorporated into new prediction g2008, which consisted of two exons in two different frames, in agreement with the novel peptide, and also the single supporting known mapped peptide.

7.3.11 Novel exon annotation

There were 39 novel exon annotation events identified (Table 7.1). An example of a novel exon annotation event was on chromosome 1DS, fragment 1DS1899380, spanning positions 5,423 to 5,449, with an event probability of 99.80%, consisting of 1 unique peptide, and with a single PSM assigned, derived from chymotrypsin-digested wheat flour. The novel exon event was identified for gene Traes_1DS_947F6918F, protein-coding transcript 1, 2 and 3. In addition, an Augustus gene prediction was carried out using the novel peptides, reference gene Traes_1DS_947F6918F and exon_part and intron hints. However, the Augustus gene prediction did not incorporate the novel peptide, only the reference gene, exon_part and intron hints (Figure 7.7, and a single supporting annotated MS/MS spectrum in Appendix Figure 7.17).

Performing a BLASTP search against bread wheat in NR revealed that the novel peptide matched to an unnamed protein product (CDM83565.1 with E-value = 2.2) with 88% query coverage and 75% identity. The three reference protein-coding transcripts matched chloroplast MDAR6 protein (AKA43771.1 with E-value range: 2E-06 – 4E-07) with ~48% query coverage and 32% identity.

The three reference protein-coding transcripts appeared to be poorly annotated, with no entries in NR, since they found no matches with significant coverage. However, the proteins did have 11 shared peptides mapped, supporting their prediction, ranging in lengths from 7 aa to 24 aa. By comparison, the novel peptide only found poor significant matches to known proteins in NR, with a relatively short peptide of only 9 aa and with little supporting evidence, indicating that the annotation event may be a false positive.

None of the other 38 novel exon annotations identified in the proteogenomics analysis led to new predictions, with the majority of novel exon annotations appearing

to conflict with the extrinsic EST, cDNA and RNA-seq evidence from the exon, exon_part and intron hints. However, in a few novel exon annotation events, including this example novel exon event for gene Traes_1DS_947F6918F, there appeared be to intron and exon_part hints across the same region (Figure 7.7), which may account for the single novel exon peptide identified. Until further evidence comes to light in support of these novel exon annotation events, it is assumed that all novel exon annotation events in the present study are likely false positives.

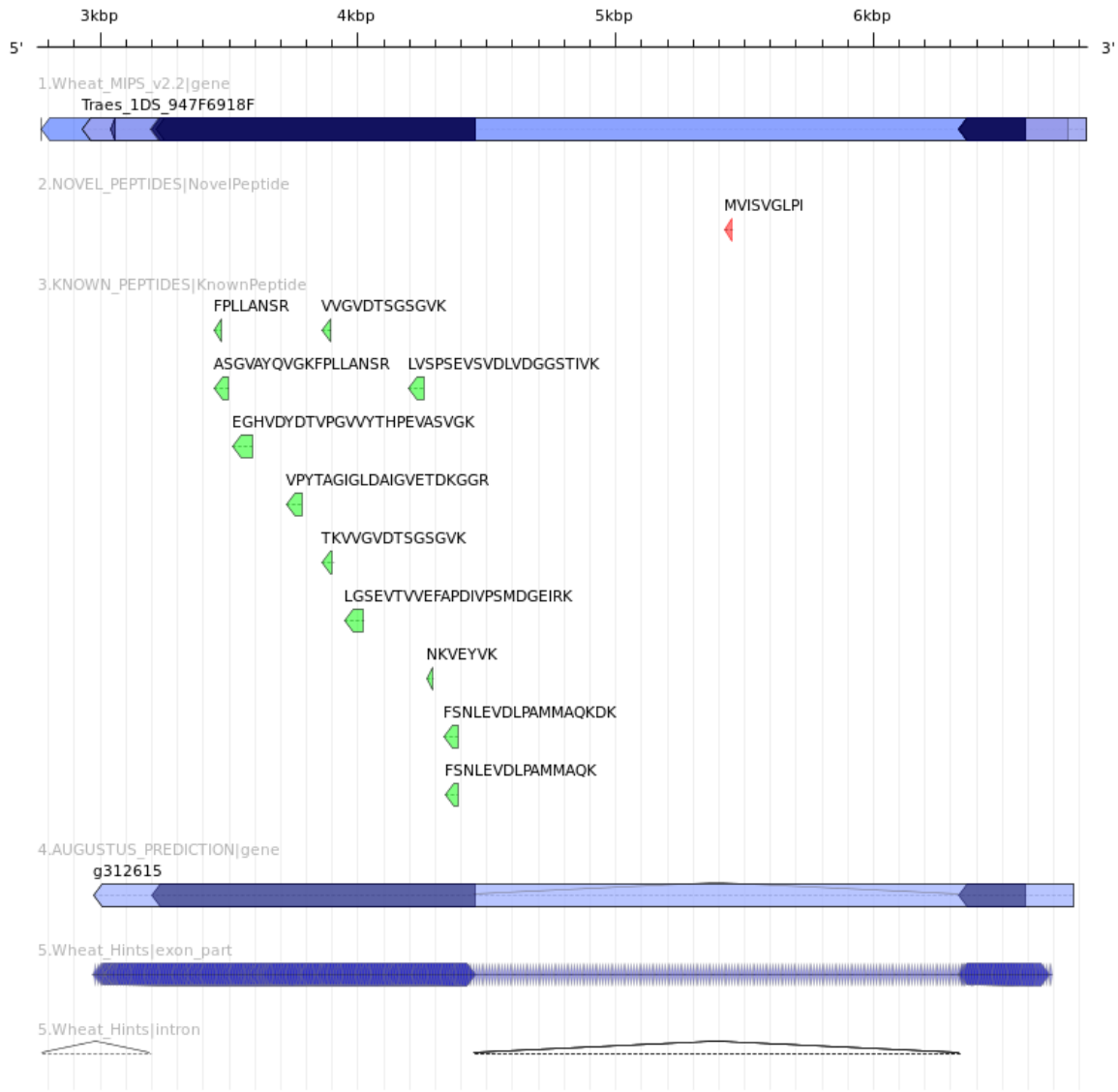


Figure 7.7 Novel exon annotation

A novel exon annotation event located on chromosome 1DS, fragment 1DS1899380. A novel and unique peptide was identified in the middle of an intron from gene Traes_1DS_947F6918F. The reference gene, novel peptide, exon_part and intron hints were used for gene prediction. All but the novel peptide could be incorporated into the prediction, even though some exon_part hints can be seen spanning the intron region.

7.3.12 N-terminal acetylated peptides

As applied in previous chapters and the overall study, N-terminal acetylation was used as a variable modification to infer the location of TIS sites, in order to further validate novel annotation events and identify potentially over-predicted known genes, or identify alternative protein isoforms with different TIS sites. A total of 87 N-terminal acetylated peptides were identified among 348 genes (795 protein isoforms) in the Wheat MIPS v2.2 annotation (Appendix File 7.7), and 10 were identified from 10 high confidence proteins (≥2 peptides with at least 1 unique). Of the total 348 genes there were 208 genes (~60%) that had N-terminal acetylated peptides in agreement with their TIS sites, and the other 140 genes (~40%) had N-terminal acetylated peptides in disagreement with their TIS sites. Due to the protein inference problem only unique N-terminal acetylated peptides were considered and these were found to map to 11 genes (11 protein isoforms), all of which were in agreement with their TIS sites.

Among the 290 novel peptides identified (Table 7.1), 12 were found to be N-terminal acetylated (Appendix File 7.7). Of these, 11 were unique and 1 was shared, with 10 of the unique peptides not incorporated into any Augustus gene predictions. The two remaining novel N-terminal acetylated peptides incorporated into gene predictions were not incorporated at the N-terminal-most end of the predictions, but instead resided towards the C-terminal or the middle of the predictions.

The one unique N-terminal acetylated peptide “RRGGQGIWRRH” was incorporated into Augustus gene prediction g151373, located at positions 18,926 to 24,839, inferred by a novel gene annotation event, on chromosome 2BL, fragment 2BL8085194, spanning positions 18,334 to 21,094. The N-terminal acetylated peptide was located well within the prediction at positions 21,062 to 21,094, conflicting with its suggested TIS site. The ambiguity that the locations of these N-terminal acetylated peptides presents could probably be attributed to the level of fragmentation of the

genome, with the large number of missing sequence regions causing misrepresentation of the majority of truly unique and shared peptides within the genome.

The single shared N-terminal acetylated novel peptide “RWERRPR” was identified within a gene boundary event on chromosome 1AS, fragment 1AS3271894, spanning positions 3,166 to 3,186, for gene Traes_1AS_16ED0B5C3 and its two protein-coding transcripts, as well as gene Traes_1AS_70E8717D0 and its single protein-coding transcript. The N-terminal acetylated peptide was, however, incorporated into two Augustus gene predictions: g12834 and g320082, located on chromosome 1AS, fragment 1AS3271894, and spanning positions 3,094 to 3,715, in line with the proteogenomics evidence, and on chromosome 2DL, fragment 2DL9871754, spanning positions 1 to 785, respectively. The location of the peptide from gene g320082 on chromosome 2DL was not reported within any novel annotation events, due to the lack of at least a single unique and novel peptide within the peptide cluster at that genomic location.

There were only 11 unique and known N-terminal acetylated peptides identified amongst the known proteins and only 1 unique and 1 shared novel N-terminal acetylated peptide identified amongst the novel annotation events, probably due to the high level of fragmentation of the genome. As a result, the ability to identify all of these N-terminal acetylated peptides as unique and thus provide a level of unambiguity to reliably identify alternative TIS sites is sorely lacking. Moving forward, this would require the genome to be further assembled before any enhanced interpretation of the locations of these N-terminal acetylated peptides.

Methods to further assist in identifying the N-terminal ends of proteins to help validate the known proteins and improve the identification of novel annotation events could be implemented through the use of N-terminomics [452], and other methods such

as heuristic and database techniques, as well as top-down proteomics, as was outlined in the preceding two chapters (Sections 5.3.13 and 6.3.9).

7.3.13 Impact of search space

Applying the two-pass search approach with two-stage FDR strategy, as previously demonstrated in Chapter 6, significantly reduced the impact that the search space had on the proteogenomics analysis. Using the same strategy in the present study 1,970 of 293,053 proteins from Wheat MIPS v2.2 annotation were identified during the MS/MS database searches using MS-GF+, while the total number of proteins mapped by proteogenomics was 16,635. Of these, 107 high confidence proteins had ≥2 peptides with 1 unique peptide.

7.4 SUMMARY

This present study has highlighted the advantages of proteogenomics and how different legacy –omics datasets from genomics, proteomics and transcriptomics can be repurposed for genomic annotation. Primarily, the study made a significant contribution to the genome annotation of *Triticum aestivum* (Bread wheat), specifically the Wheat MIPS v2.2 annotation, and which further contributed to the dissertation author's previous proteogenomics study [12]. The present study identified 290 novel peptides contributing to 189 novel annotation events (187 exclusively), consisting of 46 frame-shifts, 9 translated UTRs, 17 exon boundaries, 39 novel exons, 17 gene boundaries (15 exclusively), 24 reverse strands (24 exclusively), and 37 novel gene events, among a total of 96 genes (96 exclusively) and 189 proteins (187 exclusively). Among these annotations, 180 novel peptides directly led to 70 predicted proteins via Augustus gene prediction. This study, which implemented a new methodology, benefitted from utilising MS/MS spectra over a range of proteases and tissues and contributed a large RNA-seq dataset as a splice graph, which will become a valuable resource for further

wheat proteogenomics studies in the future. This study also identified some key highlights for consideration in future proteogenomics analysis: 1) the high level of fragmentation of the genome caused problems with annotation event identification and gene prediction; 2) the large fixed peptide linkage distance caused misidentified annotation events, as has been found to be the case in previous chapters, which also impacted the identification of annotation events where the scaffolds are much shorter than the peptide linkage distance used; and 3) the spectral datasets used were small and derived from gels which prevented clustering and quality filtering and probably resulted in retained spurious spectra.

7.5 CONCLUSIONS

In this study, as in Chapter 6, a two-pass search approach with improved two-stage FDR strategy was used. This approach greatly benefitted the study; given the further inflated search space of the 17 Gbp wheat genome and it's much larger proteogenomics search space.

Due to the small MS/MS spectral datasets, and being derived from 1D and 2D gel, clustering and quality filtering were not feasible. This decision may have inadvertently impacted the false positive rate, possibly resulting in a number of misidentifications due to spurious partially fragmented MS/MS spectra as a result of the absence of clustering, and retained poor quality MS/MS spectra because of the absence of any quality filtering. However, the use of multiple proteases probably offset these drawbacks to a degree, by reinforcing the identification of numerous annotation events derived from different sources. In addition, selection of the most appropriate precursor mass tolerances for each of these datasets proved effective, particularly with the high-accuracy MS/MS spectra. This approach, in combination with the various proteases

used to improve coverage, provided an effective method to achieve sufficient coverage and deliver an informative proteogenomics analysis.

The study was able to significantly improve the novel peptide identification rate from the trypsin-digested wheat flour by comparison with the initial study by the dissertation author [12], by an additional 122 novel peptides, demonstrating a ~8x improvement in sensitivity and with similar improvements observed in the identification of known proteins.

The merging of multiple search results from the various protein extracts and tissues digested with different proteases resulted in a number of annotation events containing multiple different sources of MS/MS spectra from different digests, while others consisted of only one source of MS/MS spectra from a single digest. This demonstrated that, by using multiple proteases, the coverage within an annotation event could be significantly increased and as a result could improve the event probability.

The MS/MS spectra and RNA-seq data sources in the present study were limited to reported findings obtainable as legacy data at the time. Due to the apparent differences between proteomics data and RNA-seq data obtained from the different tissues, cultivars and wheat-rye hybrids, there was a real potential for increasing the FDR, which needed to be taken into consideration when identify any annotation events and new predictions. Future studies, with a focus solely on obtaining MS/MS spectra and RNA-seq data for proteogenomics analysis, could concentrate on the Chinese Spring cultivar and generate sufficient depth and breadth of sampling to reduce the occurrence of any false positives. Additionally, when different sources of MS/MS spectra, genomic and RNA-seq data are included from different cultivars, the addition of variant calls in the splice graph could identify and account for the different sequence

variations, allowing for informative side studies with a focus on identifying these variants.

The total number of predicted genes and proteins in this study were probably overestimated due to the fragmented nature of the genome, with some longer genes spanning scaffolds being identified two or more times. The number of genes could also be overestimated due to Augustus splitting the prediction based on the evidence. Revisiting the genomic annotation in the future as knowledge of the genome improves would likely see these numbers drop. Another problem related to the highly fragmented nature of the genome would be the application of the peptide linkage distance, which probably would be larger than many chromosome fragments, and which would have resulted in a number of annotation events being misidentified, particular on the boundaries of a scaffold where a locus may span two or more scaffolds.

As mentioned in previous chapters, a dynamic way to determine the peptide linkage distance was needed. The problem with using a fixed peptide linkage distance was clearly illustrated in Section 7.3.6, where a novel gene event was misidentified as a gene boundary event due to its close proximity to the neighbouring gene. One possibility would be to determine the peptide linkage distance dynamically through machine-learning approaches, considering the distribution of genes in the local region as well as considering the distribution of all mapped peptides, protein, EST and RNA-seq alignment evidence.

Another problem that arose in this study as a result of the fragmented genome was the occurrence of partially predicted genes in the reference annotation. In one example, detailed in Section 7.3.8, a partially predicted gene caused problems with an assigned annotation event that contained a 41 bp ambiguous stretch of amino acid X residues, because the reference prediction overlapped the end of the chromosome

fragment. This resulted in confusion with the annotation event, as the only unique and novel peptide identified in this peptide cluster resided in the ambiguous, supposedly known coding part of the prediction, and therefore would not have been identified if sufficient coverage of the reference prediction had been available.

The Enosi tool could be improved to recognise annotations containing ambiguous protein sequences and flag these as problems prior to annotation event inference, or discard them from the analysis to prevent misinterpretation of annotation events. At the very least, the reference predictions used for analysis should be complete predictions, either full predictions or truncated, with the GFF and protein prediction in complete agreement and with no ambiguous regions confusing the assignment of the annotation event.

Overall, the fragmented wheat genome was a problem when identifying annotation events accurately and the final gene predictions were often erroneous or incomplete due to a gene overlapping the end of the chromosome fragment. As mentioned in previous chapters, possible methods which could be used to overcome such limitations in the future could be *de novo* sequencing of unassigned MS/MS spectra, matching the MS/MS spectra against closely homologous sequences, or approaches such as template proteogenomics, which could be used to build up complete coverage of proteins given closely homologous sequences to use as the genomic template and enough MS/MS spectra to provide complete coverage. However, such a method has only ever been used for small niche studies, looking at single genes, such as antibody genes, but it feasibly could be modified to work on the scale of an entire genome. Moving forward, tools such as BUSCO [588], as suggested in Chapter 5, Section 5.4, could be applied to conduct an assessment of the genome assembly and annotation, to identify any problem areas and to resolve difficult genomic regions.

7.6 ACKNOWLEDGEMENTS

Chapter 7 is in preparation for publication along with new wheat gene and protein predictions, which will be submitted to NCBI. The dissertation author is the primary author of this paper. The dissertation author designed the proteogenomics workflow, ran the analysis and wrote the paper. The dissertation author would like to thank the IWGSC for providing the wheat genome and allowing the author to contribute towards its annotation, Odd-Arne Olsen for providing an RNA-seq dataset, Susan Altenbach and William Vensel for providing the wheat flour MS/MS spectra, Ali Pendle and Graham Moore for providing wheat meiotic tissue MS/MS spectra, and Rudi Appels for providing these individual contacts. The dissertation author would also like to thank the Centre for Comparative Genomics for their compute resources and guidance and the Pawsey Supercomputing Centre for the use of their compute resources, which were supported by funding from the Australian Government and the Government of Western Australia.

8 GENERAL CONCLUSIONS

Proteogenomics has evolved greatly over the last decade, holding promise for improving genome annotation to match the accelerating rate of genome sequencing technologies. However, a number of computational challenges remain, many of which have been addressed throughout this thesis, including: how to address the reduction in sensitivity due to an inflated search space; how to accurately identify various annotation events; how to define the search space; how best to control the false discovery rate (FDR); and how best to define the parameters for a MS/MS database search.

This thesis identified a bioinformatics framework for conducting proteogenomics analysis, defined in a methodology that was iteratively improved and demonstrated across four case studies using the Enosi proteogenomics tool. During the time Enosi was being further developed, since its publication in [81], the dissertation author provided the developers with extensive input through debugging and suggested features, while conducting early preliminary studies for grape [8] and wheat [12], which were later expanded on within the thesis.

The dissertation author made further contributions to methodology development by adding an evaluation of search parameters, FDR filtering approaches, the screening of annotation events to improve the number of annotation events identified, and modifications to the search space to improve sensitivity and discrimination between true and false positives.

Finally, the thesis draws conclusions and provides further insight into proteogenomics, and highlights a number of considerations for the future of this new and rapidly growing field.

8.1 CASE STUDIES

Throughout the thesis a proteogenomics methodology was applied to a number of different case studies and it was understood that each case study would highlight the benefits and caveats of the applied methodology in an iterative improvement. The identified known proteins and numerous novel annotation events identified in each of these case studies are illustrated below in Table 8.1.

Table 8.1 Summary of proteogenomics annotations across four case studies

	Bacteria	Grape	Human	Wheat
Known proteins	3,550	11,779	27,849	16,635
High confidence known proteins	2,194	5,048	1,047	107
Frame-shifts	9	5	7	46
Translated UTRs	NA	37	4	9
Exon boundaries	22	16	27	17
Novel splices	NA	1	0	0
Novel exons	NA	9	23	39
Gene boundaries	19	160 (24)	289 (10)	17 (15)
Reverse strands	45	112 (10)	262 (50)	24 (24)
Novel genes	60	1	5	37
Total annotation events	155	341 (103)	617 (126)	189 (187)
Annotated genes	145	216 (67)	147 (29)	96 (96)
Augustus gene predictions	NA	84,948	29,266	413,587
Augustus gene predictions as a direct result of proteogenomics	NA	55	49	67

Note: Numbers in parenthesis represent the exclusive numbers. The inflationary effect of a large peptide linkage distance on gene boundaries and reverse strands was removed by assigning a peptide cluster as either a proximal or distal event, not both, with preference placed on proximal events.

In light of these findings, there were a number of problems encountered during proteogenomics analysis, which included: the underlying quality of the reference annotation; the quality of the genome assembly; the limitation of the available MS/MS spectra; and/or the different shortcomings of the applied proteogenomics approach.

A common challenge across all studies was the use of a fixed peptide linkage distance, which was an over-simplification of gene sizes and intergenic distances. The use of a fixed peptide linkage distance likely contributed to misidentified annotation events due to the many varying degrees of gene overlap, gene size and intergenic

distances across genomes the size of bacteria to larger, eukaryotic genomes. The use of a fixed peptide linkage distance was a significant problem and was resolved in part by identifying the exclusive annotation events, as highlighted in Figure 8.1, however, a more robust solution will be needed in future methodologies and/or later versions of Enosi, to more accurately identify the annotation events and by extension perform more accurate annotation event level filtering for more confident gene predictions.

The majority of case studies were able to utilise MS-Cluster to improve the overall quality of the MS/MS spectra, by clustering and merging and with PepNovo to remove any lower quality MS/MS spectra. However, there were two case studies, namely the bacterial and wheat studies, where this was not entirely possible. In the bacterial case study, the MS/MS spectra were suitable for clustering. However, the spectra were derived from 1D gels, and had relatively few peaks and lower peak intensities, resulting in relatively high spectral losses when using PepNovo. In the wheat case study, there were too few spectra per dataset, which did not allow for clustering, and the spectra were derived from 1D and 2D gels, which led to relatively fewer peaks and peak intensities and again resulted in higher spectral losses when using PepNovo. Therefore, there may have been a higher level of false positives derived from these datasets, particularly with the wheat case study, as both clustering and quality filtering were not possible. Nevertheless, the multiple sources of protease digests for the wheat study may have partially compensated, by providing higher coverage and thus reinforcing some identifications with multiple different sources of MS/MS spectra.

Revisiting these studies in the future with larger MS/MS spectral datasets, optimally derived from whole cell lysates, would alleviate these problems. While the outcomes from the pre-processing steps were not always consistent across all case studies, the use of a precursor mass tolerance optimization step greatly improved on the

sensitivity of peptide-spectrum match (PSM) identification, while keeping the peptide FDR low, which in particular was useful for high-accuracy MS/MS spectra.

8.1.1 Bacterial

In Chapter 4, proteogenomics analysis was applied to the nitrogen-fixing bacteria *Bradyrhizobium diazoefficiens*, identifying 155 novel annotation events (Figure 8.1). One particular major finding and a number of challenges were identified, as detailed below.

1) ***Major finding***

One of the novel annotation events identified was found to be an ambiguous annotation event, which hinted at a possible sequencing error. The possible sequencing error was further assessed and successfully validated, through the identification of a guanine insertion indicated from multiple sequence alignment of closely related species. Through this identification another avenue for proteogenomics has been highlighted: a means to assess the quality of the genome sequence itself within protein-coding regions.

2) ***Challenges encountered***

During analysis a number of caveats were also identified, such as the inability to utilise the identified annotation events, at least for bacteria, in heuristic gene prediction tools, apart from simply validating predictions and performing manual curation. The dissertation author is currently unaware of any prokaryotic gene prediction tools that consider external peptide evidence as hints during gene prediction. This challenge could be addressed by applying modifications to well-known and highly accurate prokaryotic gene prediction tools, such as GeneMark [598] and GeneMark.Hmm [599], Prodigal [558] and Glimmer [557]. Applying such an approach to accept external peptide hints as evidence, during gene prediction, would accelerate bacterial proteogenomics annotation

pipelines, such as with the use of Augustus [101, 102], a predominantly eukaryotic gene prediction tool demonstrated within Chapters 5 to 7.

Another caveat of this analysis was the high proportion of overlapping genes, the high proportion of coding to non-coding genes, and relatively small intergenic spaces in bacterial genomes. This caveat, which was more commonly identified in this study, resulted in a number of novel gene events being misinterpreted as reverse strand events, as genes also present on the reverse strand, were not considered by Enosi when inferring annotation events. The high proportion of overlapping genes also made the two-pass search approach less effective in choosing the most appropriate stop-to-stop ORFs from the six-frame translation to reduce the database size and improve sensitivity. A way to improve upon the selection of ORFs was later demonstrated in Chapters 6 and 7; by accepting only relatively significant matches to ORFs and hence likely spurious PSMs were removed from the second-pass search. The use of a combined FDR strategy within this study reduced sensitivity and proved less effective at accurately discriminating true and false positives from within the known and novel search spaces. This was later rectified in Chapters 6 and 7 with the improved two-stage FDR strategy, but this case study could not be later revisited to utilise the new methodology due to time constraints. To account for the high proportion of overlapping genes in this type of study, in future the use of stranded RNA-seq analysis [600] could be performed during sampling. Possibly in combination with Ribo-seq [464], previously mentioned in Section 2.4.1, to confidently identify the expressed gene on either strand and to select the most appropriate ORF, if any, prior to peptide mapping and clustering.

Although a number of caveats were identified in the present study, the use of orthologous gene prediction, demonstrated with predictions sourced from NCBI, Prodigal and RAST, proved an effective way to compare the accuracy of different prediction approaches. This method also provided an orthogonal approach to validating

predictions from the reference, and assist with determining suitable event probability thresholds, given there was no means of determining the annotation event FDR at any chosen event probability.

8.1.2 Grape

In Chapter 5, proteogenomics analysis was applied to *Vitis vinifera* (grape), identifying 341 novel annotation events (103 exclusively) that led to 57 Augustus protein predictions (Figure 8.1). A number of major findings and challenges were identified, as detailed below.

1) ***Major findings***

One of the novel annotation events identified hinted at the possibility of an over-assembly of the genome, particularly on chromosome 7. The over-assembly was inferred after the reference protein, identified novel peptides, and the Augustus gene prediction all matched significantly to a RuBisCo protein in NCBI NR, which is normally found exclusively in the chloroplast genome. The presence of a RuBisCo gene on chromosome 7 is probably a result of reads derived from the chloroplast genome being incorporated into the assembly of chromosome 7. The conclusion of an over-assembled genome is also backed up by the already fragmented nature of the genome, with unresolved chromosome fragments, which can fail to resolve into complete chromosomes when scaffolds in the assembly contain either long repeat regions and/or an over-assembly. Future proteogenomics analysis of the grape genome should therefore not proceed until this over-assembly issue can be resolved to prevent such occurrences happening in future proteogenomics studies.

In addition, the original reference annotation was found to contain numerous CDS phase errors within the GFF file, which conflicted with the predicted reference protein sequences and required correction prior to Augustus gene prediction. This

brought into question the employed method used for the original reference annotation. Ultimately for the grape genome, many improvements will be required to improve the assembly and annotation. Therefore, an in-depth review of both the completeness of the assembly and annotation should be considered to identify all of the caveats moving forward. For example, the level of completeness of the genome assembly and annotation could be determined using tools such as BUSCO [588], which utilises single-copy orthologs to assess the completeness of conserved genomic regions, as was mentioned in Sections 5.4 and 7.4.

2) ***Challenges encountered***

A number of caveats were identified throughout this study, such as the use of low-accuracy MS/MS spectra, which limited the PSM identification rate and, as a result, required larger precursor mass tolerances. Additionally, only a combined FDR strategy was implemented, with no two-pass search approach with improved two-stage FDR strategy, which was later developed and employed in Chapters 6 and 7, but which could not be revisited in this case study due to time constraints. As a direct result from using the combined FDR strategy and the larger grape genome translated in six-frames, there was also a resultant loss of sensitivity, with 52% fewer known proteins identified when performing a proteogenomics search, which also equated to a loss in sensitivity with the identification of novel peptides to infer novel annotation events.

In an attempt to negate the impact of low-accuracy MS/MS spectra and improve the sensitivity of novel annotation event identifications, the results from two proteogenomics runs with two different precursor mass tolerances of 2.0 Da and 3.0 Da were used, which provided 3.3% peptide FDR and 3.9% peptide FDR at the known proteome level respectively. The identified annotation events from both of the proteogenomics analysis runs were aggregated thus improving the coverage. Given

additional sources of MS/MS spectra of higher accuracy and sourced from multiple proteases, this same approach towards aggregating proteogenomics results could be applied to further improve coverage, as was later demonstrated in Chapter 7.

In this study, annotation events with only a single unique and novel peptide which were identified outside the applied annotation event thresholds, were accepted based on the consideration of other evidence, such as sequence homology to proteins in NR with considerations of the genomic coordinates, as well as known peptides, spectral counts, and peptide length. This approach provided a means to validate the annotation events from other orthogonal protein evidence, not considered in the original annotation, as well as provided a means to improve sensitivity, at the same time retaining relatively good specificity, as no method to determine the annotation event FDR was available. Due to the requirement to manually check matches for entries in GenBank for genomic coordinates and orthogonal evidence, the throughput of the approach was negatively impacted, making the approach particularly unruly for much larger studies. To improve the throughput of annotation event screening a modification was later considered and applied in Chapters 6 and 7.

Utilizing N-terminal acetylated peptides in both the novel annotation events and known proteins to identify translation initiation start (TIS) sites was not a trivial task to accomplish in this study, particularly in a relatively larger genome, and required manual searching to interpret potential conflicts within the annotation. Results from this analysis identified numerous ambiguous identifications due to the protein inference problem. A method to resolve this issue to reduce ambiguity and improve throughput could have been to consider only unique N-terminal acetylated peptides, as was later applied in Chapters 6 and 7. To further assist in resolving the N-terminal end of proteins and annotation events, the use of N-terminomics in combination with multiple proteases and replicates, top-down proteomics and other more heuristic methods to resolve the

protein inference problem, from a proteomics-only context (Chapter 2, Table 2.6), could be applied.

Many of the identified novel and known N-terminal acetylated peptides appeared to have undergone N-terminal Methionine Excision (NME) [248], due to the lack of their 5' Methionine cap. In instances where the 5' Methionine cap is retained and an alternative translation initiation codon is used, the peptide would not be able to map to its genomic coordinates and would go undetected in the current proteogenomics methodology. The most practical approach to identifying such peptides would be to supply another protein sample, digest, perform N-terminal peptide enrichment and then add Methionine aminopeptidase *in vitro* to cleave off all N-terminal Methionine residues, and then map the resulting peptides. The locations of peptides with 5' Methionine caps and alternative translation initiation codons could then be identified by comparison to the same samples without treatment with Methionine aminopeptidase.

The processing of overly large result files during PSM FDR filtering, from combined FDR results, proved problematic, requiring the processing of subsets of the search results to achieve filtered results in a reasonable time-frame. But this led to reduced accuracy with PSM FDRs and event probabilities. This problem was later resolved in Chapters 6 and 7 with the two-pass search approach with improved two-stage FDR strategy, which essentially provided a multiple step filtering process to reduce the final dataset size prior to the final PSM FDR filtering while also improving sensitivity and discrimination between true and false positives.

8.1.3 Human

In Chapter 6, proteogenomics analysis was applied to *Homo sapiens* (human) and 617 novel annotation events (126 exclusively) were identified that led to 52 Augustus protein predictions (Figure 8.1). A number of major findings and challenges were identified, as detailed below.

1) ***Major findings***

This case study was the first to apply a new methodology in FDR filtering for proteogenomics, using a two-pass search approach by accepting only significant matches for the second-pass search and the use of an improved two-stage FDR strategy, by performing PSM FDR filtering on separated known and novel identified sequences. This method was demonstrated to outperform the combined FDR approach and conservative two-stage FDR strategies, by identifying 44% more known proteins and 35 more novel peptides. In addition, the new methodology resulted in reduced search result file sizes, allowing for a reduction in the overhead needed to process the files for PSM FDR filtering, which was problematic in Chapter 5. This strategy was also later employed in Chapter 7. In addition, as a direct result from the improved methodology, the identification of an additional 15,020 peptides compared to the ENCODE project [11] was realised.

2) ***Challenges encountered***

A number of caveats were identified throughout this study, such as the identified discrepancy between the proteome of cell line GM12878 and the proteome from GENCODE v19. The discrepancy was realised when an expressed pseudogene ENSG00000250933.1 also known as glyceraldehyde 3-phosphate dehydrogenase 66 (GAPDHP66) was identified. This pseudogene is known to be protein-coding in cell line GM12878 and non-protein coding in GENCODE v19. As a result, the novel peptides identified from the expressed pseudogene were misinterpreted as a translated UTR. This highlighted the effect that a lack of support by Enosi to identify expressed pseudogenes and non-coding RNAs has on the accuracy of annotation event identification, as well as the time it takes to manually and correctly interpret the annotation events. This delay could have been avoided by Enosi adding the additional

annotation event types by first parsing the known annotation. In addition, future studies could identify the known protein-coding pseudogenes and/or RNA genes from the different sources being studied, such as cell line GM12878 and supplement the known proteome to avoid further false identifications.

In comparison to the previous grape study in Chapter 5, in this study an improvement was made to the throughput of screening single peptide annotation events. The same event probability and peptide parsimony thresholds were applied, and instead of using NCBI NR for sequence homology searches, NCBI RefSeq protein was used instead at higher stringency to improve the specificity. Due to its impact on throughput, with minimal exception, no screening of each match in GenBank for genomic coordinates and orthogonal evidence was performed, as was done previously in Chapter 5 with grape. However, this could be resolved in later studies by automating the process by parsing GenBank entries downloaded from NCBI. Other similar concepts to quickly and efficiently validate annotation events in a highly specific manner could be via the use of spectral library searching or with the use of spectral archives [398].

The selection of only unique and known N-terminal acetylated peptides proved to be a viable approach to the identification of TIS sites to improve throughput by reducing the number to a manageable size, in contrast to Chapter 5 with the grape study where every N-terminal acetylated peptide was manually checked and validated.

The identification of one particular unique and novel N-terminal acetylated peptide indicated the TIS site of a new Augustus prediction inferred from a reverse strand annotation event, which identified a novel gene on the reverse strand (Section 6.3.7). This brought to the dissertation author's attention the fact that unique N-terminal acetylated peptides could be used to define the boundaries of annotation events by their specific position within the annotation event to avoid including the wrong peptides from

other clusters in close proximity. However, this would be highly dependent on the level of fragmentation of the genome, with a fragmented genome making the identification of any unique peptide highly ambiguous, due to missing genomic regions, and would not account for some genes encoding multiple protein isoforms, with multiple TIS sites, and which may be expressed concurrently. Further methodologies could also be applied to improve the identification of the N-terminal end of proteins, as mentioned previously in Section 8.1.2.

8.1.4 Wheat

In Chapter 7, a proteogenomics analysis was applied for *Triticum aestivum* (Bread wheat) and 189 novel annotation events (187 exclusively) were identified that led to 70 Augustus protein predictions (Figure 8.1). A number of major findings and challenges were identified, as detailed below.

1) ***Major findings***

The same methodology, using the two-pass search approach with improved two-stage FDR strategy and annotation event-screening method as employed in Chapter 6, was also used in this study, which provided the benefit of increasing the sensitivity of annotation event identification, given the much larger 17 Gbp genome. Compared to the initial study by the dissertation author with Mayer and colleagues [12], that only utilised a small wheat flour tryptic-only MS/MS spectral dataset, 156 additional novel annotation events were identified. The use of multiple sources of MS/MS spectra from different tissues and multiple different protease digests provided higher coverage than using a single protease, and after aggregation resulted in more peptides identified within annotation events with improvement of the event probabilities. In addition, the inclusion of multiple proteases probably offset some negative effects from not clustering or quality filtering each of the MS/MS spectral datasets.

2) ***Challenges encountered***

A number of caveats were identified throughout this study, such as the annotation event screening which was done previously in Chapter 6 using RefSeq protein, but in this study due to the lack of wheat protein entries, it was necessary to use NR, filtered at much higher stringencies to account for the higher redundant and un-curated protein repository. The requirement to use NR may have inadvertently introduced false positives and suffered from reduced sensitivity in identifying the correct corresponding proteins. The inaccuracies introduced from this approach could be negated if spectral library/archive approaches were developed, as was suggested in Section 8.1.3.

Another caveat that occurred in other studies but was more common in this study was the appearance of unique N-terminal acetylated peptides incorporated into the middle region of new Augustus predictions. This error was most likely a direct result of the highly fragmented genome causing a misidentification of the unique and shared status of all identified novel peptides. Further methodologies to improve the identification of N-terminal ends of proteins could be applied, as was outlined in Section 8.1.2, although what benefits, if any, will be gained by applying these methods to a highly fragmented genome in this study is unknown. In addition, due to the small size of many chromosome fragments the peptide linkage distance was often much larger causing some annotation events to be misidentified, such as novel genes, which may be gene boundary events to genes on another chromosome fragment.

Another caveat that was a result of a highly fragmented genome was the inflation in the number of predicted genes and proteins, due to many genes and proteins spanning across chromosome fragments. Additionally, many genes could not be predicted due to small scaffolds and missing translation initiation sites off the end of the

scaffold, even though proteogenomics evidence, as well as other evidence from the reference and orthogonal hints evidence was available.

Further improvements to the genome assembly and annotation are needed before any additional proteogenomics analysis is conducted, which could be assisted using tools such as BUSCO [588], as was previously highlighted in Section 8.1.2.

A unique issue that was only identified in this study was the misidentification of a translated UTR event, revealed as a probable exon boundary event for two different protein isoforms within the same gene. The exon boundary event was not primarily identified by the Enosi tool, due to two protein isoforms containing an ambiguous X amino acid region, where the novel peptides should have been identified as known but were instead identified as novel. Resolving the ambiguous sequences would have resulted in the novel annotation event not being identified, as the single unique peptide was identified within the 'known' ambiguous region of the protein. The dissertation author drew two conclusions from this observation: 1) if possible, ambiguous X amino acid protein sequences in protein predictions should not be included in analysis, however this is often difficult due to partial reference predictions with first draft genomes; and 2) using all mapped peptides during clustering with at least 1 unique peptide, followed by removal of all known peptides would identify many more annotation events at a high specificity. Using all peptides during peptide clustering would be superior to using only novel peptides, as many more valid peptide clusters could be identified containing a unique peptide. Even as the annotation improved the sensitivity of novel annotation event identification could remain high as the number of identified unique and novel peptides diminished.

8.2 FUTURE DIRECTIONS

Over the course of this thesis an improvement to how proteogenomics could be conducted was realised, in terms of further expanding on the case studies outlined within this thesis and the field of proteogenomics as a whole.

8.2.1 Future directions for case studies

In terms of how any specific case study could be improved in the future, the use of the two-pass search approach with improved two-stage FDR strategy, which was developed and utilised in the human (*Homo sapiens*) and wheat (*Triticum aestivum*) case studies in Chapters 6 and 7, respectively, could be applied to the nitrogen-fixing bacteria (*Bradyrhizobium diazoefficiens*) case study in Chapter 4 and the grape (*Vitis vinifera*) case study in Chapter 5.

To further expand on the case studies, more evidence could be utilised, such as larger MS/MS spectral datasets and of higher mass accuracy, preferably from multiple different proteases and also comparably suitable large RNA-seq datasets, all of which could be provided with high depth and breadth of coverage. In addition, the large datasets could also be tailored with other studies in mind, hence value-adding to its use. For example, once the datasets have contributed to improving the genome annotation, the new predictions could be used in a proteomics-only analysis, examining the differences between environmental conditions such as stress, time-points, tissues, protein expression levels, and peptide variants between individuals and/or closely related species.

An example of such a large-scale proteogenomics study could be the human genome. As one of the most studied research areas, there is now a plethora of additional datasets to choose from, one of the better known being the 1,000 Genome Project [65]. This large dataset could provide a resource of RNA-seq data that, when converted to a

splice graph, could help identify the complement of variant peptides within a population, in a similar manner to the cancer peptides in the study from [474], but from 1,092 individuals. However, for the study to be comprehensive, proteomics data would also need to be obtained representing these individuals. There is currently one large resource of MS/MS spectral data for human, from recent proteogenomics studies which amassed an impressive 25 million MS/MS spectra from a diverse range of tissues and cell lines [450, 451], and which could be further mined for information using the methodologies outlined within this thesis to identify more novel annotations which previously have been missed.

Considering the above-mentioned human studies, a much larger and more ambitious undertaking would be to conduct a systems biology study, where the genomes, transcriptomes and proteomes of 1,000 individuals would be sampled from a wide variety of tissues and multiple ethnicities. This would account for protein-coding variants across the global human population at the genomics, transcriptomics and proteomics levels, and could form a new standard for identifying variations across human populations. To take this one step further and to value-add to such a study, the whole epigenome and metabolome could also be mapped from the global population.

8.2.2 Methodology improvements

In general terms of future directions for proteogenomics methodologies, huge room exists for improvement, with many of these avenues for improvement identified throughout this thesis and which are now discussed in detail in the following eleven points.

1) ***Expanding on event types***

To reduce the occurrence of misinterpreted annotation events, further annotation event types could be added to the Enosi tool. The addition of further annotation events could

include expressed pseudogenes, for example, the two pseudogenes manually identified in Section 6.3.8 and long non-coding RNAs (lncRNAs), which are employed in tools such as PGTools [496] (previously discussed in Section 2.4.4). Other types of annotation events could include over-predicted genes, N-terminal methionine excision (NME) and signal peptides inferred from the presence of non-tryptic N-terminal peptide ends, and alternative TIS sites within the known proteins [492]. The identification of potential over-predicted genes and TIS sites was conducted manually throughout the thesis looking at only unique N-terminal acetylated peptides due to the large numbers and ambiguity from the protein inference problem. In addition, sequence variant events could also be considered, such as mutation, insertion and deletion, achieved through the use of RNA-seq alignments and common variant-calling tools such as the Genome Analysis Toolkit (GATK) [473], which could then be converted into a splice graph to supplement the standard splice graph. The variant splice graph could then be used to answer questions such as differences in the types of variants between samples, to identify potential causes of changes in protein function, and to account for physiological responses and/or underpin specific phenotypes.

2) *Resolving the negative impact of fragmented genomes*

During the proteogenomics analysis of grape (Chapter 5) and wheat (Chapter 7), the level of fragmentation of these respective genomes was quite high, with wheat being the most fragmented and grape appearing to be also over-assembled, with numerous identifications from chloroplasts and mitochondria. Future proteogenomics studies of these genomes could either implement a methodology to utilise sequences of close homology to account for the missing genomic regions, such as implementing template proteogenomics using a tool similar to GenoMS [481]. The alternative could be to wait until they have been assembled sufficiently and accurately, devoid of over-assemblies until re-analysis is conducted, to avoid any false positive identification. In order to

utilise template proteogenomics, another closely related genome could be added to the analysis. Regions of the genome that do not have homology to any of the fragmented or partially assembled target genome could be identified, and then selected to use with a tool such as GenoMS, which could be modified to handle larger genomic sequences. Any MS/MS spectra identified as unassigned to the target genome could be run against the homologous genomic fragments to identify missed protein-coding genes. The identified proteins could then possibly assist with validating the genome assembly as it further improves.

3) ***Impact of inaccurate annotation events and filtering prior to gene prediction***

Many novel peptides could not be included into Augustus gene models, as was first pointed out in Chapter 5, Section 5.3.4, and found to be the case throughout Chapters 6 and 7. This had been confirmed to be the case in a previous study [81], even with an applied annotation event FDR of 5%, as was communicated to the dissertation author (S. Payne, personal communication, September 29, 2012). Based on this information, it is not unreasonable to assume that numerous novel peptides, which were not incorporated into predictions by chance, could have also been incorporated into the predictions, leading to false positives. This was observed to be the case throughout the thesis, with some probable spurious peptides within annotation events leading to what appeared to be false positive predictions, and in some cases conflicting with other evidence. This was at odds with what was suggested in the original study that led to the development of Enosi [482]. According to the author of that study, acceptance of novel peptides at the PSM and annotation event identification levels was allowed to be more tolerant for final filtering of false positives at the gene prediction level with Augustus (N. Castellana, personal communication, April 15, 2014).

However, as observed from examples throughout this thesis, using Augustus as a final false positive peptide-screening tool was not without its problems, particularly when the actual FDR at the gene prediction level is not known. To improve the quality and throughput of genome annotation the quality of the final gene products need to be improved, and therefore the quality of the proteogenomics data provided to Augustus prior to gene prediction needs to be improved to reduce the occurrence of false positive predictions.

A number of methods could be employed to reduce the occurrence of false positives at the peptide, annotation event and prediction level, as further detailed below.

4) ***Defining the peptide cluster prior to annotation event inference***

A method to reduce the occurrence of false positives reaching the annotation event inference and gene prediction stage is to change the way peptide clusters, and therefore annotation events, are defined by considering other evidence during clustering. One variable which most influences the peptide cluster is the peptide linkage distance, which defines how peptides are clustered, and the distance between a gene and a cluster, and which by extension, directly influences the assignment of events. In particular, the novel gene, gene boundary and reverse strand events.

In Enosi, the peptide linkage distance is currently a fixed value and, at least for eukaryotic genomes, is determined based on the majority of gene sizes across the genome, unlike in reality where gene sizes often vary widely and can often appear in close proximity or with large intergenic distances. A more intuitive approach would be to define the peptide linkage distance for each peptide cluster in a dynamic way, looking at additional evidence such as: 1) the distribution of all peptides across the genome; 2) considering evidence from aligned EST, protein and RNA-seq sequences within the region; and 3) the average size of genes in the local region. All of this evidence could be

used to build a model of likely gene distribution across the genome using a machine-learning approach. However, it is unreasonable to assume that the correct peptide linkage distance would be chosen every time for every peptide cluster, as it may not always be easy to determine due to the level of fragmentation of the genome and the available evidence in proximity to the identified peptide clusters. In cases where evidence is lacking to build a reliable model, a default peptide linkage distance, determined by the user, could be assigned.

Another method would be to redefine the way peptide clusters are generated. Currently in Enosi, all known peptides are removed and then the remaining novel peptides are clustered together based on the peptide linkage distance, which may include other spurious peptides or peptides that should belong to other nearby clusters. Each unique or shared novel peptide is identified in the context of the proteogenomics search space (six-frame translated genome and if applicable, splice graph) and whether the peptide is absent from the known proteome and the respective genomic coordinates. The known peptides are only labeled as either unique or shared in the context of the known proteins. Since only novel peptides are used in peptide clustering, with 1 unique peptide per cluster, prior to annotation event inference, this highlights a potential oversight for the use of the known peptides to identify valid annotation events.

As was briefly introduced in Section 8.1.4, and will now be expanded on here, a better suggestion would be to first perform clustering prior to discriminating between known and novel peptides, which would improve the sensitivity of annotation event identification when selecting for at least 1 unique peptide per cluster.

Since the genome annotation will improve over time, the pool of unique novel peptides will ultimately be reduced until most peptide clusters only consist of shared novel peptides. To avoid this and improve on the sensitivity of annotation event

identification, all peptides, including those that are identified to map to known proteins, could also be mapped to the six-frame translated genome, and if applicable, splice graph and used for peptide clustering, with subsequent removal to leave only novel peptide clusters. This would ensure that the sensitivity of annotation event identification would not diminish, even as the genome annotation improved. After filtering out all known peptides, the remaining peptide clusters with at least 1 unique novel peptide, as well as only shared novel peptide clusters (which previously contained at least 1 unique known peptide), could be used for annotation event inference. This approach came to light when considering the annotation event identified in Section 7.3.8, where a unique novel peptide was identified in an ambiguous protein sequence region. Such an approach could be implemented into the Enosi tool in future versions, however it would require some major reworking of a number of core functions.

To further expand on this approach, after peptide clustering the included peptides could be screened based on other evidence. For example, peptides could be excluded to their own peptide clusters if they to do not have the same frame as the majority of other peptides (if any) and do not cluster within the same exon/CDS and ORF(s). Considerations such as these, if neglected, could lead to ambiguous and likely false positive frame-shift events and other proximal events such as translated UTRs and exon boundaries, as was identified to be the case in Section 6.3.7.

Further evidence, which could also be considered, includes the location of unique and novel N-terminal acetylated peptides. For example, in Section 6.3.7 a unique and novel N-terminal acetylated peptide was identified on the boundaries of a peptide cluster, which led to a new prediction in agreement with the N-terminal acetylated peptide.

As was demonstrated in Section 6.3.7, peptide clusters may sometimes inadvertently incorporate other groups of peptides that should not belong together. Improvements could be achieved by considering any identified unique N-terminal acetylated peptides in the peptide cluster, and then splitting the peptide cluster at the location of the N-terminal acetylated peptide. This could be done provided that the unique status of the N-terminal acetylated peptide was not incorrectly assigned due to a fragmented genome and that multiple protein isoforms have not concurrently been expressed containing multiple different TIS sites across the same gene, which would be ambiguous due to the protein inference problem and lead to incorrectly splitting peptide clusters. However, by identifying the expression of multiple isoforms for a single gene, the larger peptide cluster assigned to that gene could be retained to avoid the ambiguous splitting of the peptide cluster or peptides could be assigned to the different protein isoforms and their respective peptide clusters more coherently. The identification of multiple protein isoforms for a gene and the intelligent assignment of peptides to different split peptide clusters could be achieved by first using traditional proteomics means, using tools such as those listed in Table 2.6, prior to proteogenomics analysis to identify the most likely protein isoforms being expressed. Such an approach would, however, only be of use for the known proteins used in the proteomics analysis, with later interpretation of any novel gene protein isoforms performed post-proteogenomics analysis.

The use of N-terminomics [452] in combination with multiple proteases and sample replicates could also be used to improve coverage and assist with the resolution of ambiguity by assigning N-terminal acetylated peptides within peptide clusters and annotation events, to help define their boundaries and also validate the TIS sites of known proteins. This approach could also include the identification of non-AUG translation initiation codons and a variety of non-acetylated N-terminal peptides.

5) ***Filtering peptide clusters and annotation events***

Throughout this thesis, during annotation event filtering it was observed that many smaller annotation events that were filtered out with lower event thresholds often overlapped with larger annotation events with higher event thresholds and were erroneously retained as a result. Which contributed to false positives at the peptide, annotation event and prediction levels. An example of one such annotation event where this occurred was identified in Section 6.3.7, where a single novel peptide was inferred as a frame-shift event. It was previously filtered out with a lower event probability, but was retained in a larger reverse strand event with a higher event probability, and ultimately led to a probable false gene prediction that conflicted with RNA-seq evidence in GenBank.

Many of these caveats could be addressed to improve the final peptide clusters and annotation events by identifying and then filtering out these erroneously included peptides and annotation events. This would keep the filtering of all annotation events consistent, improve specificity and reduce false positives, as well as removing much of the manual annotation and screening.

A further method to improve how novel annotation events can be defined is by being more selective in how a PSM and, by extension, the identified peptide is defined as known or novel prior to peptide clustering and inference. This approach would further improve the discrimination between true and false positives. Currently the Enosi tool simply identifies a peptide as known by parsing the known proteome coordinates, with all peptides not identified considered as novel. This, like the peptide linkage distance, is overly simplistic in its approach. In many instances a spectrum matches the proteogenomics search space as well as the known search space due to overlap with their respective genomic coordinates, particularly in the combined FDR approach,

which makes the search space redundant and reduces sensitivity when applying PSM FDR to all known and novel results. The ‘conservative’ two-stage FDR approach applied from the study in [474], and examined in detail in Chapter 6, goes to the other extreme by completely removing any PSMs identified as known. These problems were addressed in this thesis with the improved two-stage FDR strategy, by creating separate known and novel search spaces prior to PSM FDR filtering. However, even though this was appropriately addressed, there are still a number of spectra which can match both the novel and known search spaces and which can be further confounded when the search performed by Enosi accepts the top 10 matches, increasing the potential overlap between known and novel identifications.

To further differentiate the novel and known PSMs, the results between the known identifications and novel identifications could be compared. Any PSMs identified in one search space which have a lower spectral E-value compared to the corresponding PSM in the other search space should be retained, while the other PSM is discarded. In cases where the spectral E-value is identical in both novel and known search spaces, both PSMs should be discarded to avoid ambiguous identifications. This method to utilize the spectral E-value provides a convincing approach, as the spectral E-value is independent of the database size and would allow a simple comparison between the likelihood of one spectral interpretation over another.

6) ***Increase stringency and specificity at cost of sensitivity***

Lacking the means to apply any other methodology improvements as has been detailed elsewhere, an alternative means to reduce the incidence of false positives could be achieved by applying much higher stringencies on the quality of the MS/MS spectra by only using large clustered MS/MS spectral datasets which could tolerate quality filtering at the highest stringencies to reduce false positive rates. In addition, only the highest of

event thresholds with the higher event probability (>99.9%) with ≥2 unique peptides per peptide cluster could be accepted across all annotation event types. Another approach could be to increase the stringency at the PSM level by using a PSM FDR lower than 1% or by filtering the search results directly using a spectral E-value, which would be more suitable for smaller genomes [492]. Multiple different PSM level filtering methods could be integrated into Enosi to allow for more flexible approaches for filtering. Although these methods would significantly reduce the incidence of false positives, they may also negatively impact sensitivity with the loss of many real annotation events. The impact of increasing stringencies at the PSM level and/or annotation event level could be determined in future studies to ascertain acceptable thresholds across different studies using the best balance between sensitivity and specificity to achieve the lowest possible false positive rate.

7) ***Defining further search spaces for refined control on FDR***

To improve control on FDR the final filtered peptide clusters and annotation events as well as the different types of annotation events need to have their boundaries well defined, then they can be used to split the proteogenomics search space based on their propensity for false positives, for example annotation events from intergenic spaces versus intragenic spaces. Referring back to point 1) above, by expanding on the repertoire of annotation event types the proteogenomics database could be further split into annotation event-specific sequences to define a search space for each annotation event and thus improve the identification rate of novel PSMs and the distinction between true and false positives. This approach would prove highly beneficial when using a higher diversity of proteomics and RNA-seq data.

The idea to apply PSM FDR filtering on each of the different types of annotation events was first mentioned by Nesvizhskii [439], who suggested that each class of novel

PSM (i.e. novel PSMs belonging to different annotation event types), could have their class or annotation event-specific PSM FDR calculated. For example, the false positive rate of the novel gene and distal events, such as gene boundary, translated UTR and reverse strand, is much higher than other annotation event types. The number of false positives in the intergenic space for novel genes would be highest, and lower for translated UTR which is very close to the reference coding region. Furthermore, for proximal events the false positive rate would tend to be lower than novel gene and distal events, with novel exons and frame-shifts likely to have higher false positive rates than novel splice, which would be higher still, compared to exon boundary events. In consideration of these varying false positive rates, a multi-stage FDR strategy could be applied, as opposed to a two-stage FDR strategy which only implements FDR filtering on the known and entire novel search space.

Based on the above information, a correlation can be seen with the number of expected false positives within certain novel annotation events among the three categories of annotation events; novel gene, distal events, and proximal events, and the degree of change required for re-annotation. Further work could map the false positive rates of each annotation event in a range of different case studies and datasets with a variety of different genomic sizes. After improvements to the proteogenomics methodology, as previously envisioned, the annotation event FDR could be determined within a number of case studies and across the various types of annotation events. From such a study, a general rule could be deduced for an applied event probability for each annotation event type depending on the size of the genome (within a given range of magnitudes), allowing for more appropriate control on the proteogenomics analysis and to reduce the false identifications of annotation events and predictions in the future. A side-benefit afforded by applying PSM FDR filtering to each of the annotation events

would be a further reduction in the processing overhead for FDR filtering, which could be particularly advantageous when the datasets used in a study are particularly large.

8) ***Improving proteogenomics throughput***

The analysis of large genomes in proteogenomics is often hampered by poor throughput, however a means around this issue would be the utilization of high-performance computing (HPC) resources. There are currently two throughput bottlenecks in a proteogenomics workflow: 1) the MS/MS database search and 2) applying PSM FDR filtering and local FDR calculations on proteogenomics results.

As was detailed in Section 2.4.2, a number of methods can be employed to improve the throughput of an MS/MS database search, with one such method being the splitting of the database into smaller parts, which is employed by the Enosi tool and which was employed throughout this thesis. Another method that could efficiently utilize HPC resources would be an MS/MS database search tool developed with Message Passing Interface (MPI), allowing for searches to be performed across multiple nodes in a cluster, such as MassMatrix [601]. However, the large majority of search tools, like MS-GF+ [292] and others, only use multi-threading, forcing the user to either split databases or use search tools with faster but less sensitivity algorithms, such as the sequence tag approach employed by InsPecT [326], and other approaches as highlighted in Section 2.4.2. Ultimately, this lack of broad support for MPI in the proteomics community relegates MPI to only specialty cases.

To the best of the dissertation author's knowledge there is no way to improve the throughput of PSM FDR filtering and local FDR calculation. However, by splitting the result file and processing each file separately the throughput can be improved at the cost of reduced PSM FDR accuracies and conversely to improve accuracies iterative searches can be applied to reduce the database size using the two-stage FDR strategy, at

the cost of throughput. Therefore, for this stage of the proteogenomics workflow, a trade-off between throughput and FDR accuracy needs to be decided, until more efficient algorithms are developed or which could possibly implement strategies utilizing MPI.

9) ***Added functionality to proteogenomics***

Another viable improvement to proteogenomics would be further functionality, by adding multiple other pre-processing, post-processing and analysis tools to the workflow. For example, PGTools [496] has multiple functionality, such as: 1) allowing switching between proteomics- and proteogenomics-only analysis; 2) merging multiple search results from multiple search tools to improve sensitivity; 3) a file conversion module for handling conversion of multiple spectral data formats into MGF format for processing; 4) visualization modules for the generation of Venn diagrams to display unique and overlapping peptides; 5) chromosome distribution and Circos plots of identified mapped novel peptides; 6) the generation of a treemap to show protein grouping; 7) protein annotation; 8) customization of genome databases to suit the type of analysis; 9) the ability to port into genome browsers such as the UCSC Genome Browser and IGV; and finally 10) generating a summary report.

These types of additional functions could be integrated into tools, like Enosi, by the developers or more easily and much more quickly added by the user with little or no assistance by a developer, within a workflow environment such as Yabi, previously discussed in Section 2.5. Yabi could run an entire proteogenomics or proteomics workflow from beginning to end, with multiple branching inputs and outputs, allowing for very powerful workflows that could run on HPC resources.

Other tools which could be utilized in a workflow environment to enable further functionality for proteogenomics and proteomics analysis include pre-processing tools

for deconvolution, such as Zscore [293], deisotoping, such as LASSO [294], quality filtering, such as PepNovo [302] and/or clustering, such as MS-Cluster [312]. In addition, post-processing tools to automate the filtering of annotation events using additional information, such as was manually conducted throughout this thesis, as well as further post-processing tools such as Cytoscape [602] and STRING [603] for functional protein-protein interaction studies, and functional annotation analysis using tools like DAVID [604] and AutoFACT [139], could process the final gene and protein products for further genomic annotation.

10) ***Post-proteogenomics: Integration of multi-stage peptide identification***

In addition, methodologies such as the multi-stage peptide identification strategy [426], could also be integrated into such workflow environments. The use of multi-stage peptide identification was outlined previously in Section 2.3.9, and aims to assign each and every spectrum to a peptide, by iteratively searching conventional databases, spectral libraries, identifying common post-translational modifications (PTMs) through a "blind" search and finally any remaining unassigned MS/MS spectra searched against a translated genome database, as is done in proteogenomics analysis. This method is akin to the 'conservative' two-stage FDR strategy [474], which was outlined and compared in Section 6.3.4, identifying its pitfalls and as previously discussed in terms of the ambiguities which can result with the overlap between MS/MS spectra identified as both known and novel.

A modification to the above mentioned multi-stage peptide identification strategy could be to first apply a proteogenomics analysis (with splice graph if applicable) using the methodologies demonstrated within this thesis, as well as the above suggested improvements, followed by gene prediction and validation. Following this strategy, any unassigned MS/MS spectra not matching the known proteome or

leading to an annotation event, which identified a new or improved gene prediction, could then be used in multi-stage peptide identification. This strategy could identify further peptide sequences and possibly their genomic coordinates; 1) with unaccounted for PTMs; 2) from different database; 3) identified from raw translated reads or transcripts *de novo* assembled from any un-aligned RNA-seq reads not included in the initial splice graph; and 4) spectral libraries. Any remaining spectra, by process of elimination, may likely be false positives due to poor quality/partial fragmentation. The unassigned MS/MS spectra, which would now be accounted for and identified, could then re-enter the proteogenomics analysis, with new considerations for adding additional PTMs and/or further sequences such as missing genomic regions, splicing regions from RNA-seq evidence or derived from particularly small ORFs in the six-frame translation, missed in the first proteogenomics run. A strategy of this nature could be implemented within a workflow environment, such as Yabi, which would enable the user to capture meta-data, share workflows and parameters with other collaborators to assist within large national or international research efforts.

11) ***Integration of ortho-, meta and comparative proteogenomics***

Another improvement to proteogenomics could be an automated approach towards applying comparative proteogenomics [446, 448], ortho-proteogenomics [79, 445] and metaproteogenomics [503], previously discussed in Section 2.4.3. In comparative proteogenomics, given two or more closely related species with proteomics, genomics and transcriptomic datasets for each, the identification of annotation events could be validated in each of the parallel analyses within the same genomic coordinates. Such an approach could help confirm that the identification is real, particularly for single peptide annotation events, often referred to as “one-hit-wonders”, which would also do away with much of the effort needed to validate novel peptides against orthogonal protein repositories.

Although the use of event probabilities is touted as a viable solution to salvaging single hit peptides for the identification of proximal events [82], the most appropriate thresholds needed to be applied to the various annotation events to achieve a high level of sensitivity and specificity while keeping the false positive rate low is still unclear. A more direct form of validation with the use of two or more confirmed annotation events in similar parallel comparative proteogenomics analyses could resolve the ambiguities.

In ortho-proteogenomics any identifications from a single proteogenomics study could be used to further identify genes and proteins in other highly similar genomes [79, 445], which was highlighted with a few examples in the bacterial study from Chapter 4, particularly for the frame-shift event identified in Section 4.3.10.

Given the complexity of higher eukaryotes, analyses such as these using comparative proteogenomics may only be viable for closely related species of bacteria, or sub-species or different varieties of higher eukaryotes, which may only differ with subtle variations in sequence. Such an approach could be automated to perform analyses across different proteogenomics runs in a workflow environment, which when comparing the analyses could apply a scoring scheme for the identification of annotation events across species. Such an approach would vastly improve sensitivity and resolve ambiguous annotation events, as well as fast-track analysis.

In metaproteogenomics, since many of the species within the bacterial communities are likely unknown, an initial step towards a comprehensive analysis could be a metaproteomics step for identification of suspected known proteins in the sample already represented in public repositories such as GenBank. Through the proteogenomics pipeline, these identified MS/MS spectra could then be searched against the six-frame translation of the metagenome, mapped, genomic coordinates determined, peptides clustered and novel gene events inferred. All the identified novel

protein-coding regions could then be annotated and gene models predicted, followed with functional annotation based on comparative genomics analysis. The full set of MS/MS spectra and the now identified "known" proteome could then be used during a full and more thorough metaproteogenomics analysis of the metagenome using the two-pass approach with improved two-stage FDR strategy, as outlined within this thesis, and with potentially some additional improvements suggested throughout this chapter. Given sufficient depth of metagenome sequencing, coverage and assembly, the identification of unique peptides within a metaproteogenomics analysis would likely indicate the identification of proteins unique to a particular bacterial species, and which may indicate a unique role within the bacterial community.

8.2.3 Application of new MS technologies

Up until now, much has been discussed on how to improve upon the current implemented methodologies for proteogenomics, with adjustments to already established methods, inherited from developments in proteomics over the last few decades. Many of these methods are reflected in the proteogenomics tools listed in Section 2.4.4. A quick PubMed search reveals an explosion of new proteogenomics tools within the last 12 months of this thesis, all with slight variations compared to the Enosi tool used throughout this thesis, which provides a good indication that the future of proteogenomics will evolve more rapidly with time. However, there are three new technologies into which proteogenomics has yet to tap and predictably will evolve to become the gold standard in proteogenomics in the coming years. These tools are: top-down proteomics, multiplexed data-independent acquisition (DIA) and spectral archives.

Top-down proteomics, as outlined previously in Section 2.3.2, would be suitable for complementing bottom-up proteomics, due to its limitation to provide global coverage but powerful ability to provide full coverage of a limited number of proteins,

including a full range of their PTMs, due to solubility and ionization problems with some large proteins. The technique has improved in throughput over the last decade making it a viable tool for large-scale studies, where it is important to identify the N-terminal end of proteins and to identify sites of potential signal peptide cleavage and full coverage of PTMs. This makes the technology a suitable candidate for implementation into proteogenomics, allowing further coverage to infer many more annotation events both confidently and without ambiguity, resolving the protein inference problem, resolving multiple co-expressed protein isoforms, and identifying annotation events amongst other paralogous genes and proteins. There are not to the dissertation author's knowledge, currently any proteogenomics tools or strategies available that utilize top-down proteomics data. Such tools would require algorithmic improvements for handling and interpreting the data, over traditional bottom-up approaches. As the mass spectrometry technology improves to generate, interpret and handle such data, this will become an important resource to complement current bottom-up proteogenomics approaches and in combination with approaches such as N-terminomics [452].

Multiplexed DIA, as outlined previous in Section 2.3.3, is another technology gaining much more traction in recent years, as the rate of technological advancement of mass spectrometry has improved, allowing for better speeds of data acquisition, interpretation and management. This technology allows for an unprecedented level of depth and breadth of coverage by considering all precursor ions (including multiple precursor ions from multiplexed MS/MS spectra) and their MS/MS spectra in a sample instead of pre-selecting the precursor ions in DDA. Multiplexed DIA also allows for all the data from the sample to be stored for further data mining without the need to re-sample. By incorporating multiplexed DIA into proteogenomics, much more coverage would be attainable, and with sufficient control on FDR would allow much higher rates

of annotation event discovery. Coupling this technology with top-down proteomics would require reconsideration on proteogenomics algorithmic design to accommodate the data. This arrangement would ultimately provide near-complete coverage of the proteome for proteogenomics analysis, resolving ambiguities with assigning annotation events and significantly reducing the occurrence of annotation events with single peptides, as well as resolving ambiguities when trying to identify TIS sites.

Spectral archives, as outlined previously in Section 2.3.8, is another technology that promises to change the landscape for spectral interpretation in MS-based proteomics completely, and predictably will have wide sweeping implications for proteogenomics in the years to come. In essence, the use of spectral archives would allow a higher level of specificity in identification and also discrimination between known and novel spectra by matching to a large archive of clustered spectra, both identified and unknown. Each spectrum in the archive would improve and be further validated each time further spectra are added. This level of enhancement would eliminate the ambiguity of assigning matches, since a SSM compared to a PSM, provides a much higher sensitivity and specificity. Any spectra from a proteogenomics study could contribute to the archive and in return the archive could assist with validating known proteins and identifying novel annotation events with confidence. The spectral archive would also remove the need to cross-validate matches against curated protein databases, as performed manually throughout this thesis, which was identified as an important consideration when performing proteogenomics analysis [439].

A possible method which could be used to integrate spectral archives with proteogenomics, to identify the type of annotation events and their genomic locations, would be to first cluster the spectra by adding it to the spectral archive and contribute towards the growing archive. Following this step, sequences from the stop-to-stop ORFs of the six-frame genomic translation and sequences from a splice graph could be

split into known and novel sequences, based on the known reference annotation. The set of known and novel sequences could then be digested *in silico* to peptide sequences, based on different proteases and cleavage specificities and converted into simulated spectra using tools such as Mspire [605], all the while keeping track of their genomic coordinates. The known simulated spectra derived from the reference annotation could be validated against the spectral archive, to see if it matches only known spectra identified as belonging to a curated protein within the same genomic coordinates. The novel simulated spectra derived from the novel search space could be validated against the spectral archive to determine if they happen to match any known spectra already identified from a curated protein or un-identified spectra currently not assigned any annotation to validate its novelty. All the SSMs could be processed by the proteogenomics pipeline with both the identified novel and known spectra (with any PTMs) interpreted back into peptide sequences with retained genomic coordinates. The peptide sequences could then be clustered and annotation events inferred in context to the reference annotation.

This approach is in contrast to previous methods of spectral identification, as the query and target have been inverted during the search stage. However, using this approach can provide an unparalleled level of specificity and sensitivity and, if event probabilities could be incorporated, the large spectral support provided by the spectral archive could be leveraged to improve the discrimination between true and false positive annotation events further, with few or no ambiguous annotation events.

Overall, there is much promise for proteogenomics moving forward, with many possible approaches in its application, or a combination of approaches, such as the latest considerations with top-down proteomics, multiplexed DIA and spectral archives, to truly bring proteogenomics on a par with the latest next generation sequencing

technologies. The time when proteogenomics will herald the extinction of the non-model organism is fast approaching [3].

8.2.4 Guidelines for proteogenomics

To bring proteogenomics into the future, based on the findings from this thesis, some or all of the following seven points (Figure 8.1), should be considered in order to better contribute to genome annotation efforts:

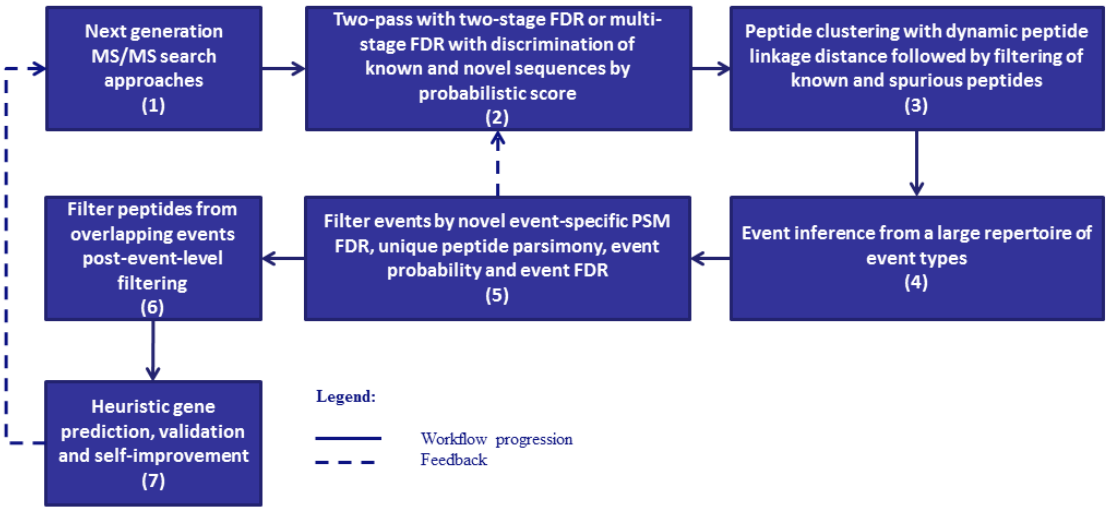
1) The use of the latest "next-generation" fast, accurate and sensitive MS/MS search tools such as MS-GF+ [407] should be used, or improved proteogenomics approaches utilizing multiplexed-DIA spectra and associated search tools, top-down proteomics methods and/or highly specific approaches, such as spectral library searching e.g. with tools like Tremolo [397] or the spectral archives approach [398, 399]. In addition, these tools should include methods to improve throughput, such as multi-threading and/or MPI support.
2) The workflow should include a two-pass search approach using significant matches in the second-pass and some form of two-stage PSM FDR for the known and then all novel sequences. Alternatively, multi-stage PSM FDR to the known sequences and then novel sequence from each annotation event type (novel annotation event-specific PSM FDR) could be applied. To improve the discrimination between the known and novel identifications, a probabilistic approach that is independent of the database (e.g. spectral E-value) should be used.
3) Regardless of the proteogenomics approach adopted, to improve the rate and accuracy of the assignment of annotation events the workflow should accurately cluster peptides using all identified peptides and use of a dynamically and machine-learned peptide linkage distance, considering other evidence. Peptide clusters should have known peptides removed and the novel peptide clusters should be filtered to remove any ambiguity introduced from incorrectly clustered peptides, based on

other evidence, such as the frame of the majority of peptides within exon/CDS and ORF(s). These steps would achieve consistency and reduce ambiguity across all identified peptides and annotation events.

4) In any proteogenomics approach there should be a very broad variety of novel annotation event types to account for all different types of search spaces that have varying levels of false positive rates. Not only those annotation event types identified by Enosi throughout this thesis, but also annotation events such as over-predicted genes, NME, signal peptides, expressed pseudogenes, expressed non-coding RNAs, and variants such as insertions, deletions and mutations. Only then can better control be applied to the false positive rate throughout the analysis, as each annotation event will have its own specific false positive rate.
5) The identified genomic and RNA-seq derived sequences from the different novel annotation events should then feed back into point 2) above, to define the novel annotation event-specific search space to apply an annotation event-specific PSM FDR and/or apply filtering with unique peptide parsimony and event probability thresholds with an annotation event level FDR determined for each annotation event type.
6) After the annotation events have been accepted from point 5) above, they should be further filtered for any incorrectly included overlapping peptide clusters, to keep the applied thresholds consistent and unambiguous at both the peptide and annotation event level, which could otherwise lead to false positive annotation events and predictions.
7) Finally, heuristic methods of gene prediction, such as Augustus should be used to improve the predictions given the proteogenomics evidence as well as any other orthogonal evidence. The predictions should be annotated and validated through a variety of means, such as an automated annotation pipeline as well as through 'wet'

lab validation. Any unassigned spectra could be validated through a modified template proteogenomics approach and/or a multi-stage peptide identification strategy to be identified in a second adjusted run of the proteogenomics pipeline to cater for the missed PSMs. All identified predicted novel and modified gene products should assist with the identification of orthologous genes in closely related species, where applicable, in an ortho-proteogenomics approach. In addition, the whole proteogenomics run could be performed in parallel with a closely related species, where applicable, with a cross-validation of identified annotation events and novel PSMs to improve sensitivity and specificity. Additionally, filtering stringencies at the PSM level and annotation event level should be optimized to reduce false positives and any identified peptides that were not incorporated into validated gene predictions should be investigated to see how to further improve the methods employed within points 1) through 6).

Figure 8.1 Seven guidelines for proteogenomics



8.3 CONCLUSIONS

This thesis identified and developed a bioinformatics framework for proteogenomics, defined within a new methodology, which was demonstrated in a number of case studies through which its viability was tested. A number of caveats were identified as

the methodology was implemented and, as a result, the methodology evolved through a series of modifications and future improvements. In addition, significant contributions were made towards the genomic annotations of bacteria, grape, human and wheat, highlighting important discoveries and caveats. Overall, the thesis identifies and brings to light considerations for future proteogenomics strategies, by suggesting the integration of new methods and technologies.

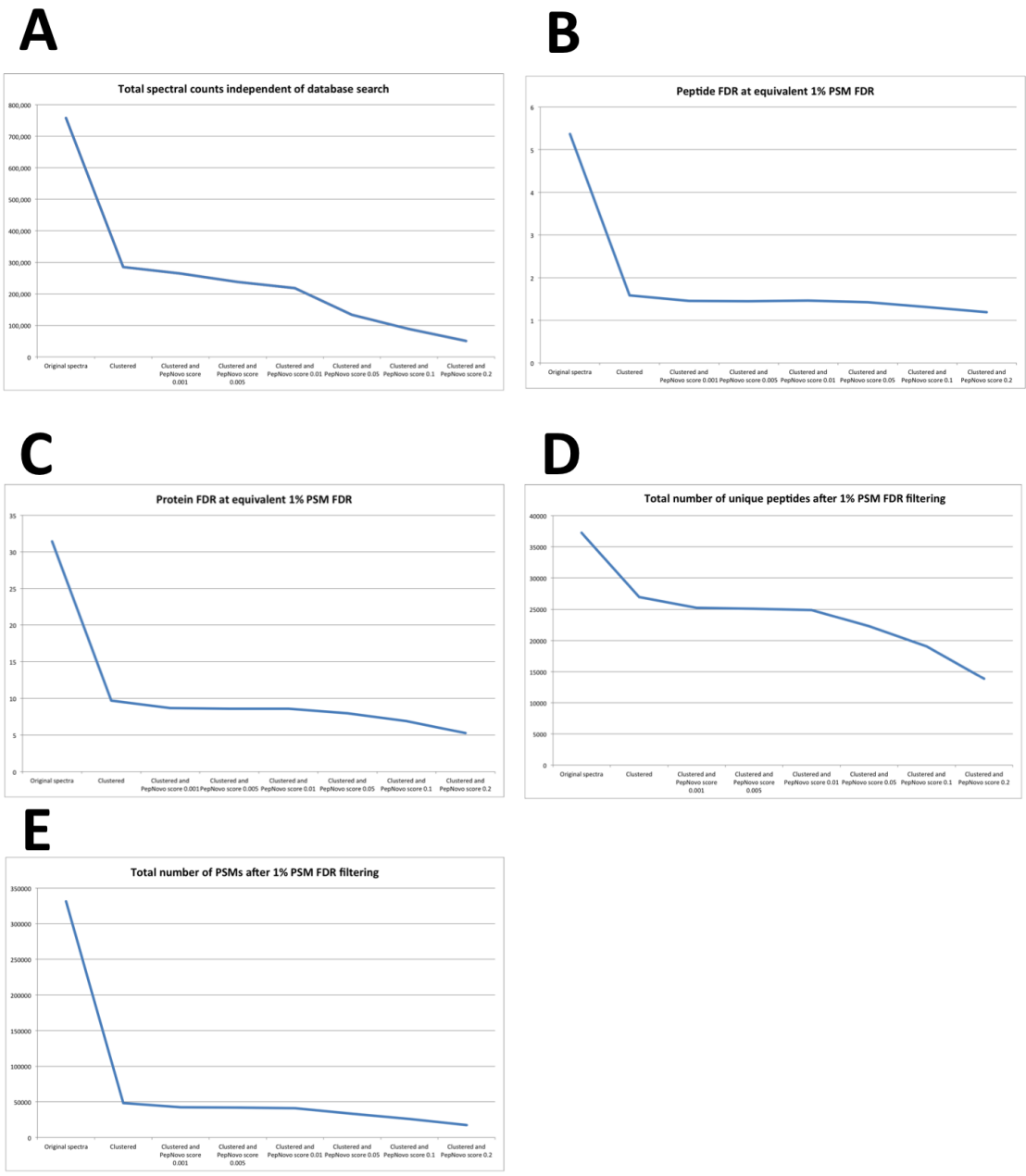
APPENDIX

Appendix File 4.1 Reference predictions are in zip file 'AppendixFile4.1.zip' on the DVD provided.

Appendix File 4.2 Clustering, quality filtering and precursor mass tolerance optimization results are in excel file 'AppendixFile4.2.xlsx' on the DVD provided.

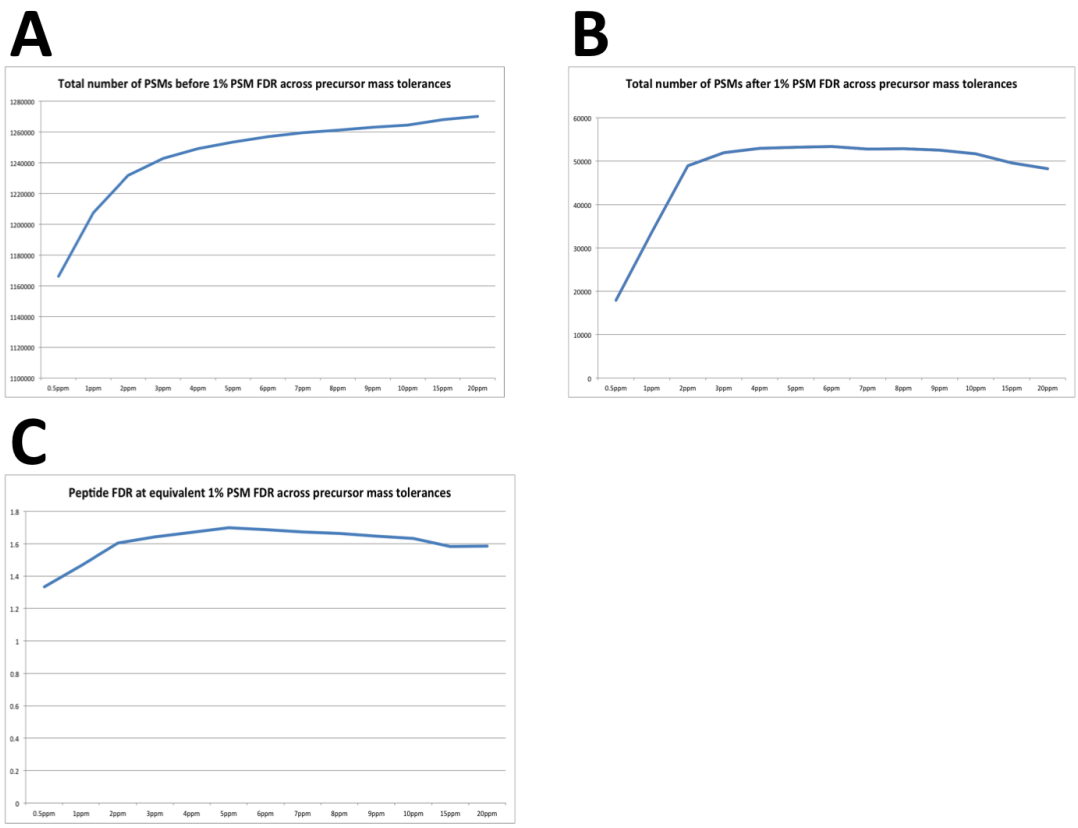
Appendix File 4.3 Processed proteogenomics results are in excel file 'AppendixFile4.3.xlsx' on the DVD provided.

Appendix File 4.4 Raw proteogenomics results are in zip file 'AppendixFile4.4.zip' on the DVD provided.



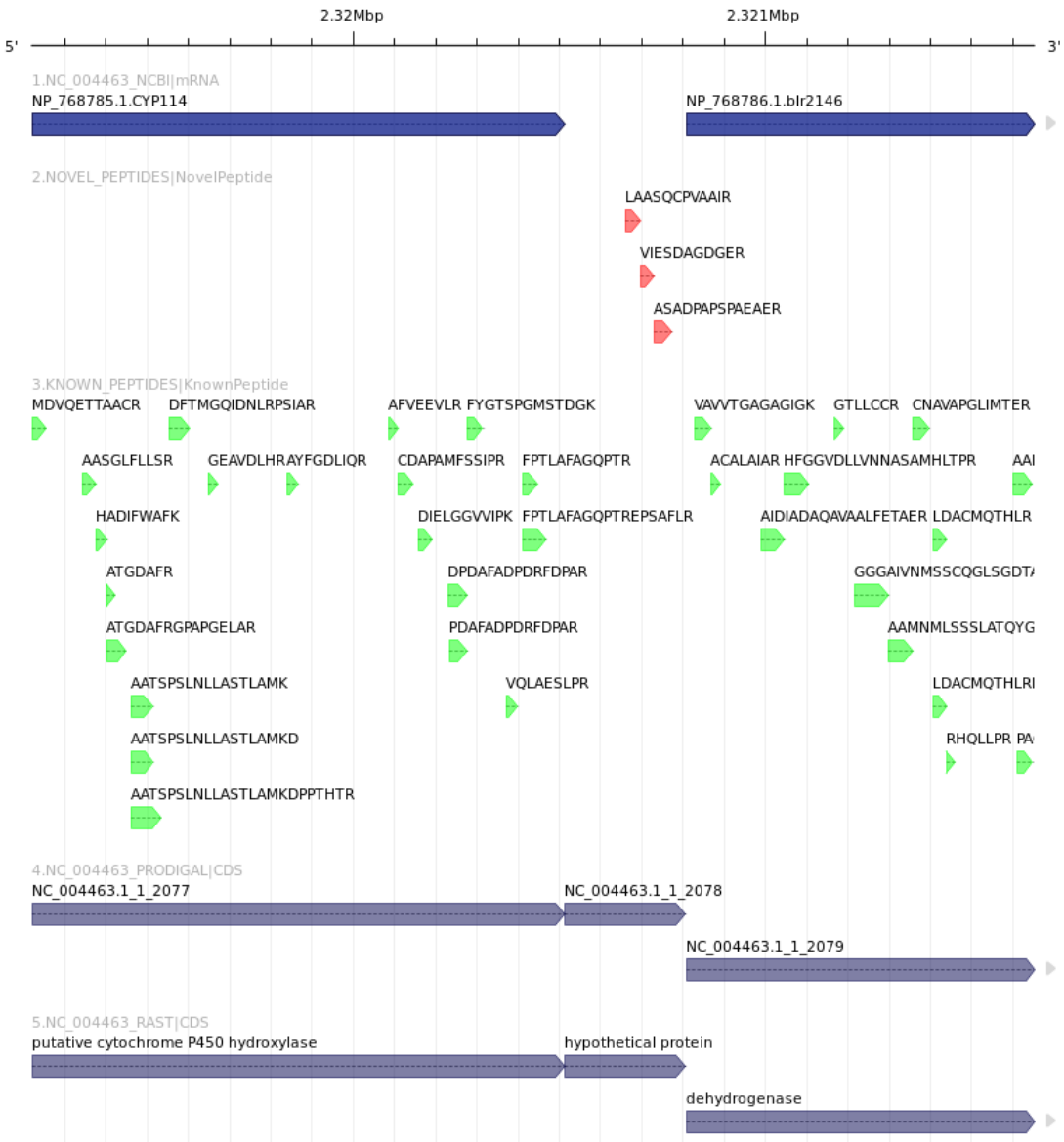
Appendix Figure 4.1 Pre and post clustering with and without PepNovo quality filtering

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



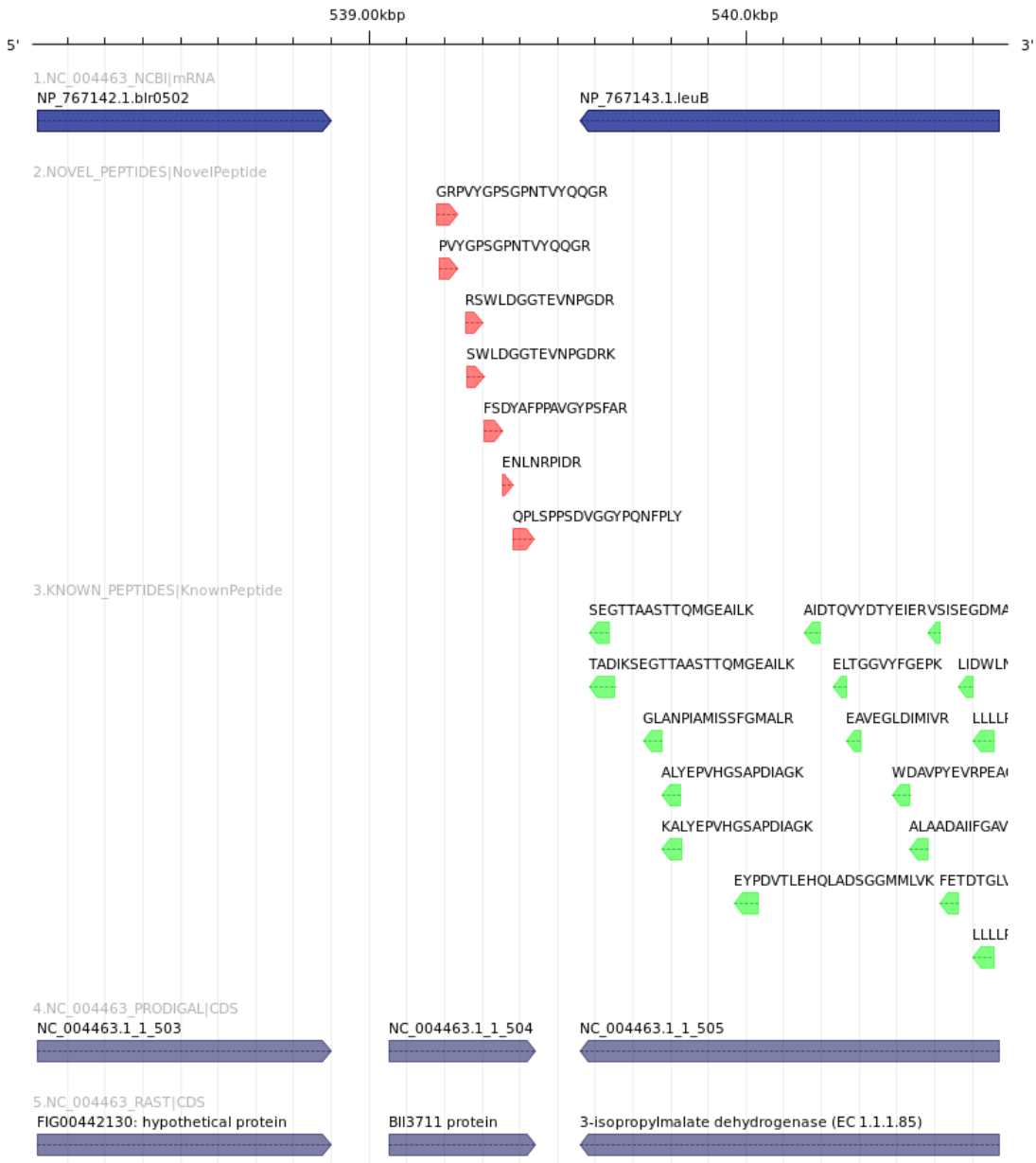
Appendix Figure 4.2 Precursor mass tolerance optimisation

(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.



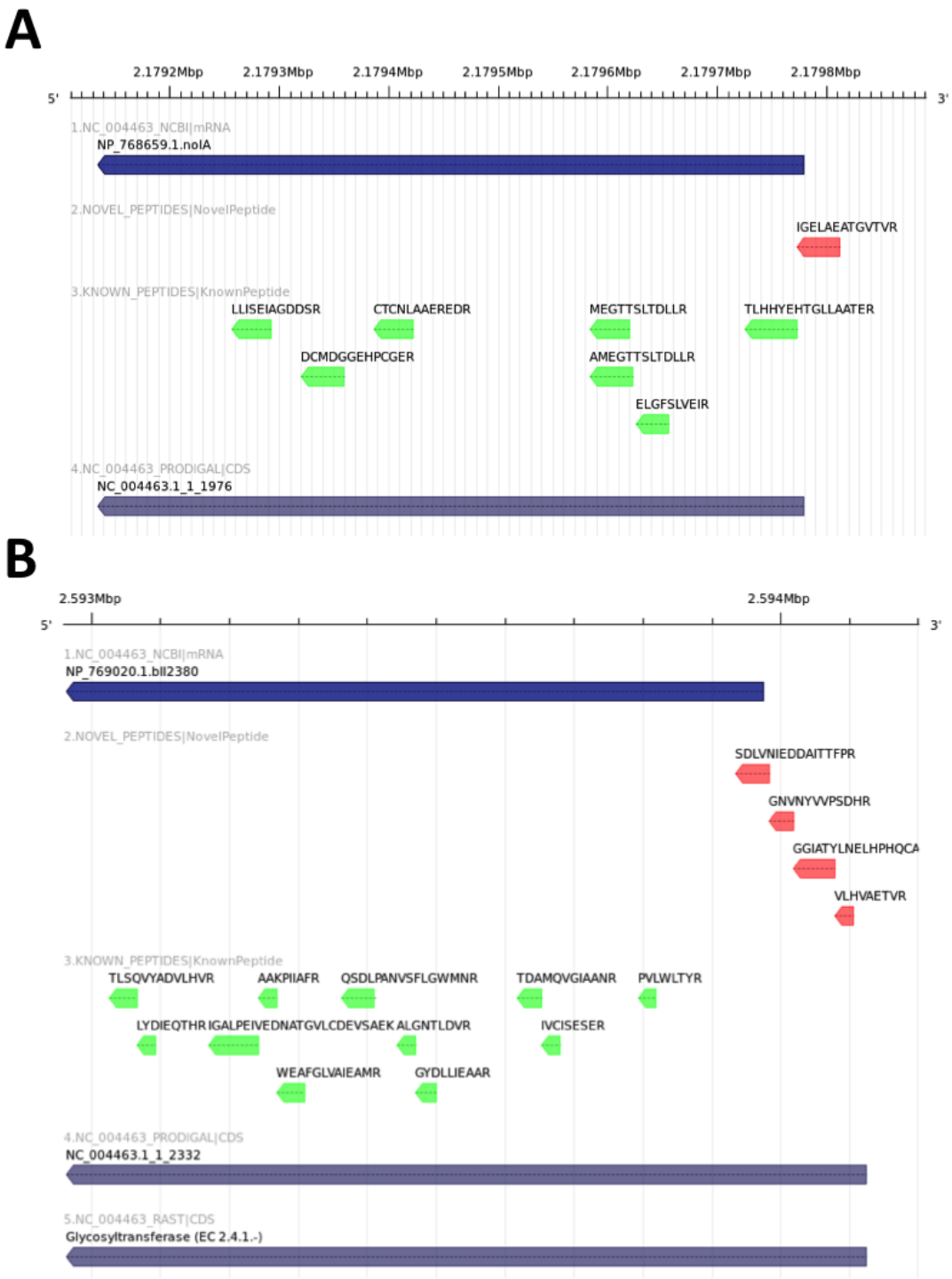
Appendix Figure 4.3 Gene boundary of blr2146 or novel gene

The gene blr2146 had a gene boundary event, while the study in [502], Prodigal and RAST suggested a novel gene annotation event.



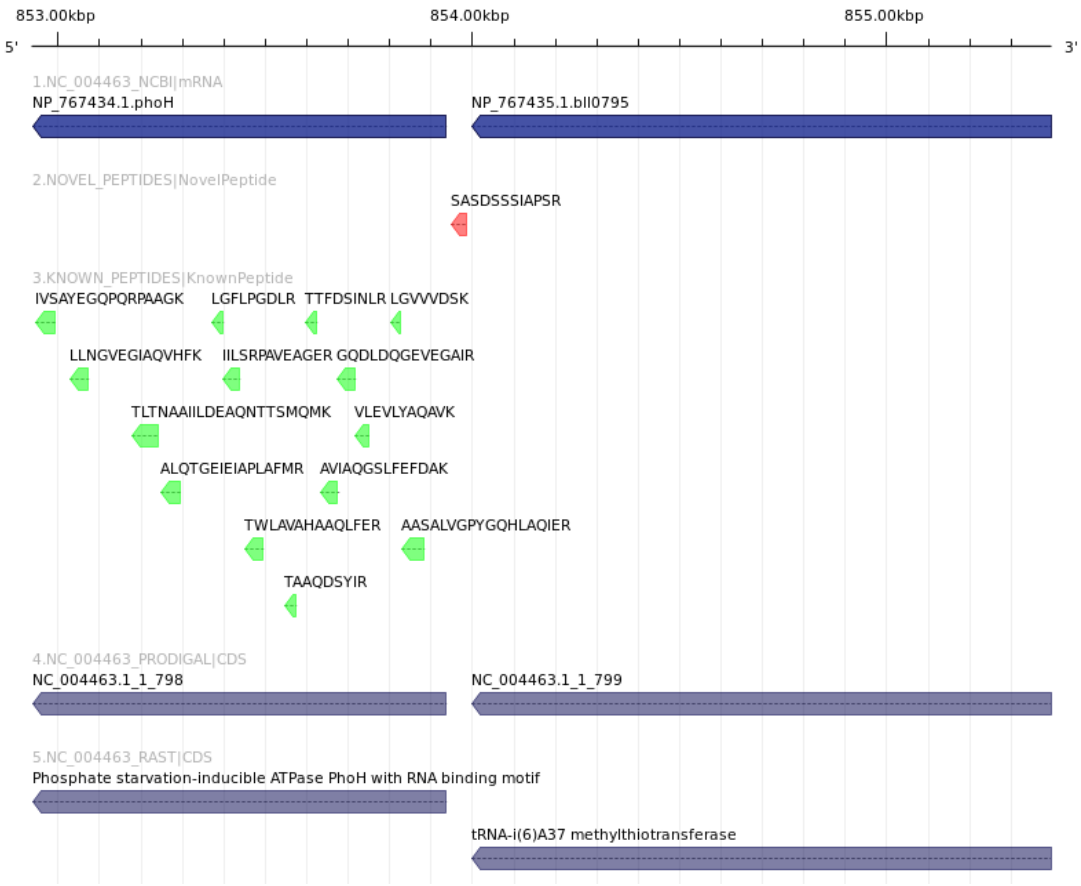
Appendix Figure 4.4 High confidence novel gene annotation

A novel gene annotation event in agreement with the study from [502], RAST annotation and Prodigal predictions.



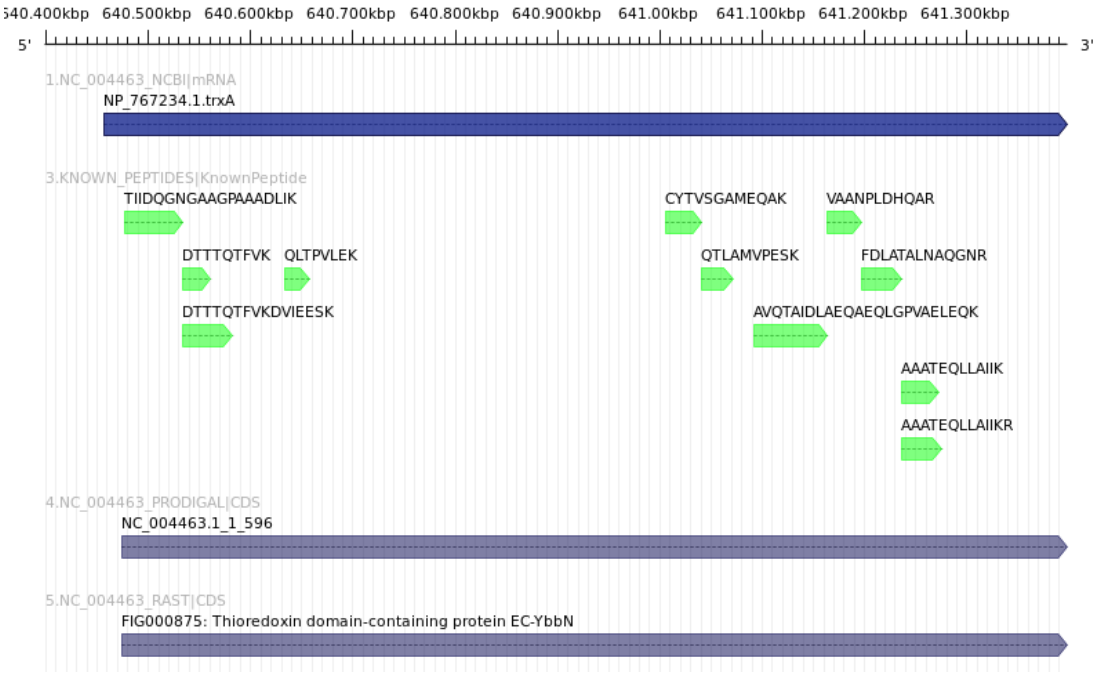
Appendix Figure 4.5 Exon boundaries of gene bll2019 and bll2380.

(A) The gene bll2019 (*NolA*) had an exon boundary event, in agreement with the study from [502]. RAST did not predict bll2019, resulting in a novel gene event, while Prodigal predicted bll2019, and is in agreement with the exon boundary annotation event. (B) The gene bll2380 had an exon boundary event, which was in agreement with the study from [502], RAST annotation and Prodigal predictions.



Appendix Figure 4.6 Gene boundary annotation

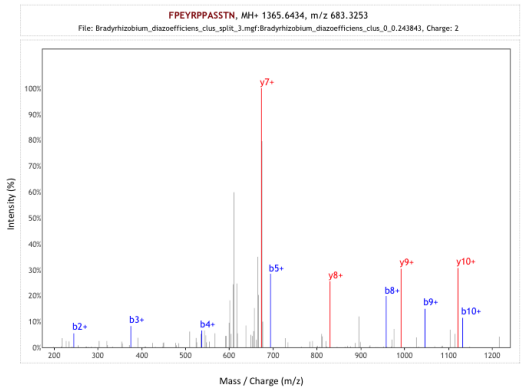
The gene bll0794 (*PhoH*) and bll0795 both had gene boundary annotations. The evidence from known peptides, suggested a gene extension to bll0794, however there was no supporting evidence from the RAST annotation and Prodigal predictions.



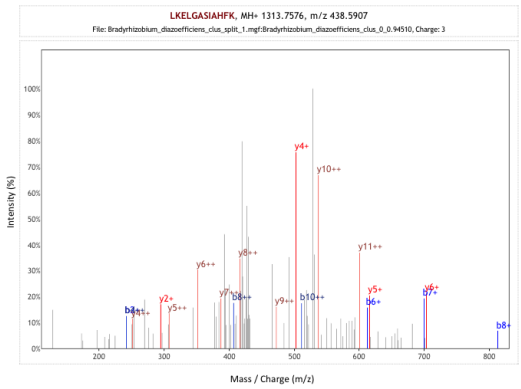
Appendix Figure 4.7 Known peptide mapping to blr0594 (*trxA*)

The unique peptide "TIIDQGNGAAGPAAADLIK" mapped to gene blr0594 (*trxA*). The study from [502] identified the unique peptide as being N-terminal acetylated indicating an alternative translated initiation start (TIS) site. However, according to the MS-GF+ search results in this study, the peptide was not N-terminal acetylated.

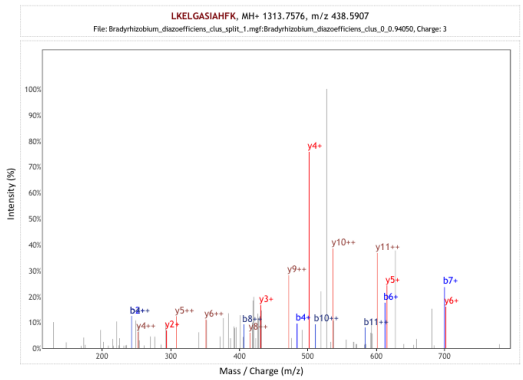
A


FPEYRPPASSTN, MH+ 1365.6434, m/z 683.3253
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.243843, Charge: 2
Intensity (%)
Mass / Charge (m/z)

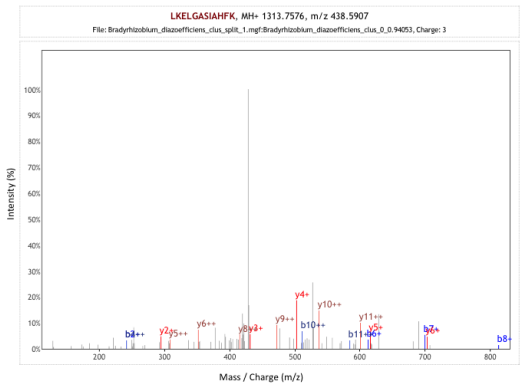
B


LKELGASIAHFK, MH+ 1313.7576, m/z 438.5907
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.94510, Charge: 3
Intensity (%)
Mass / Charge (m/z)

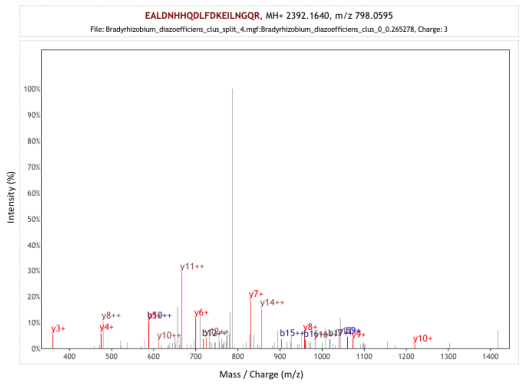
C


LKELGASIAHFK, MH+ 1313.7576, m/z 438.5907
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.94050, Charge: 3
Intensity (%)
Mass / Charge (m/z)

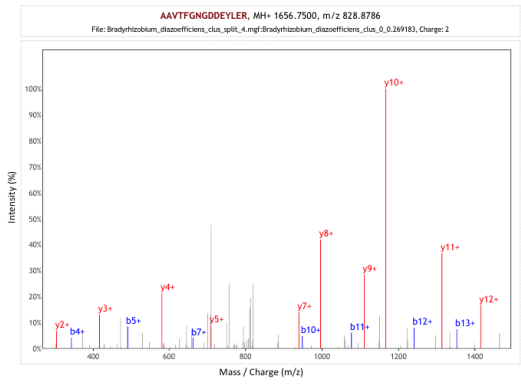
D


LKELGASIAHFK, MH+ 1313.7576, m/z 438.5907
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.94053, Charge: 3
Intensity (%)
Mass / Charge (m/z)

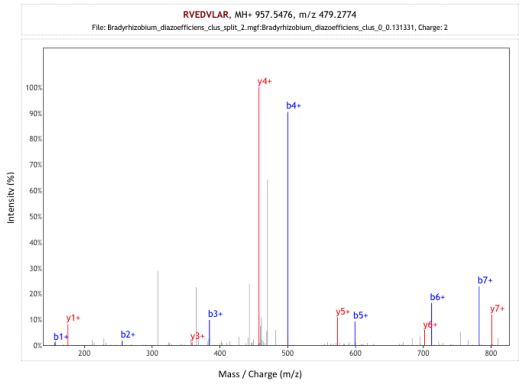
E


EALDNHHQDLFDKEILNGQR, MH+ 2392.1640, m/z 798.0595
File: Bradyrhizobium_diazoefficiens_clus_split_4.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.265278, Charge: 3
Intensity (%)
Mass / Charge (m/z)

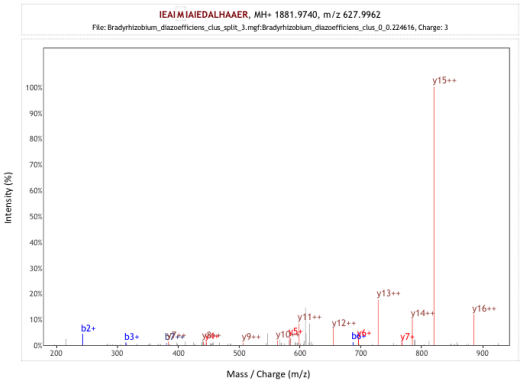
F


AAVTFGNGDDEYLER, MH+ 1656.7500, m/z 828.8786
File: Bradyrhizobium_diazoefficiens_clus_split_4.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.269183, Charge: 2
Intensity (%)
Mass / Charge (m/z)

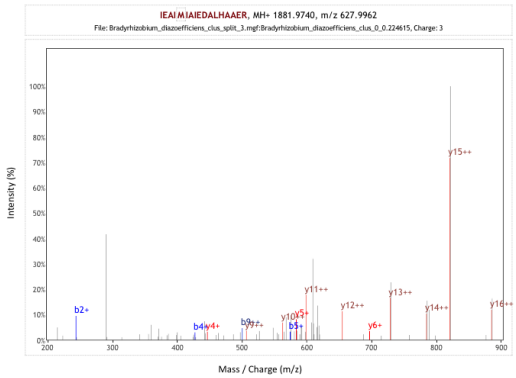
G


RVEDVLAR, MH+ 957.5476, m/z 479.2774
File: Bradyrhizobium_diazoefficiens_clus_split_2.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.131331, Charge: 2
Intensity (%)
Mass / Charge (m/z)

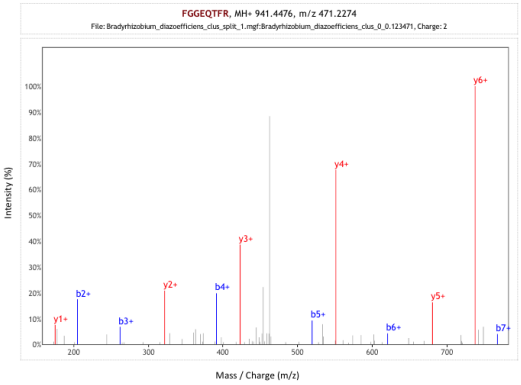
H


IEAIMIAIEDALHAAER, MH+ 1881.9740, m/z 627.9962
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.224616, Charge: 3
Intensity (%)
Mass / Charge (m/z)

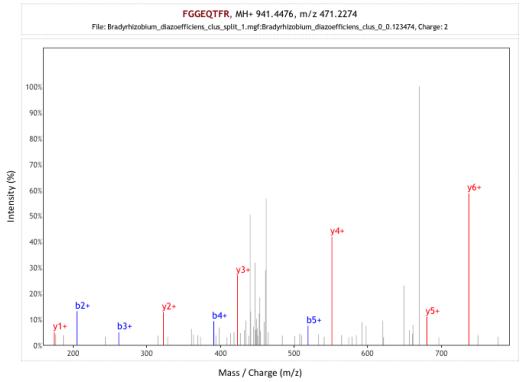
I


IEAIMIAIEDALHAAER, MH+ 1881.9740, m/z 627.9962
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.224615, Charge: 3
Intensity (%)
Mass / Charge (m/z)

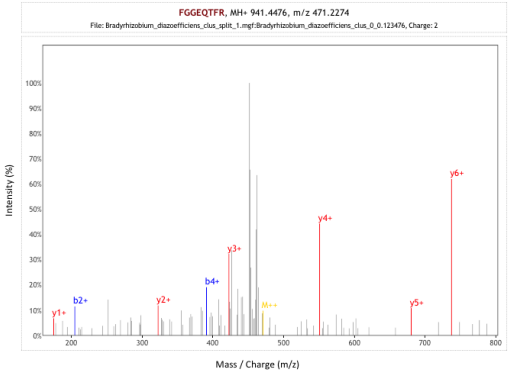
J

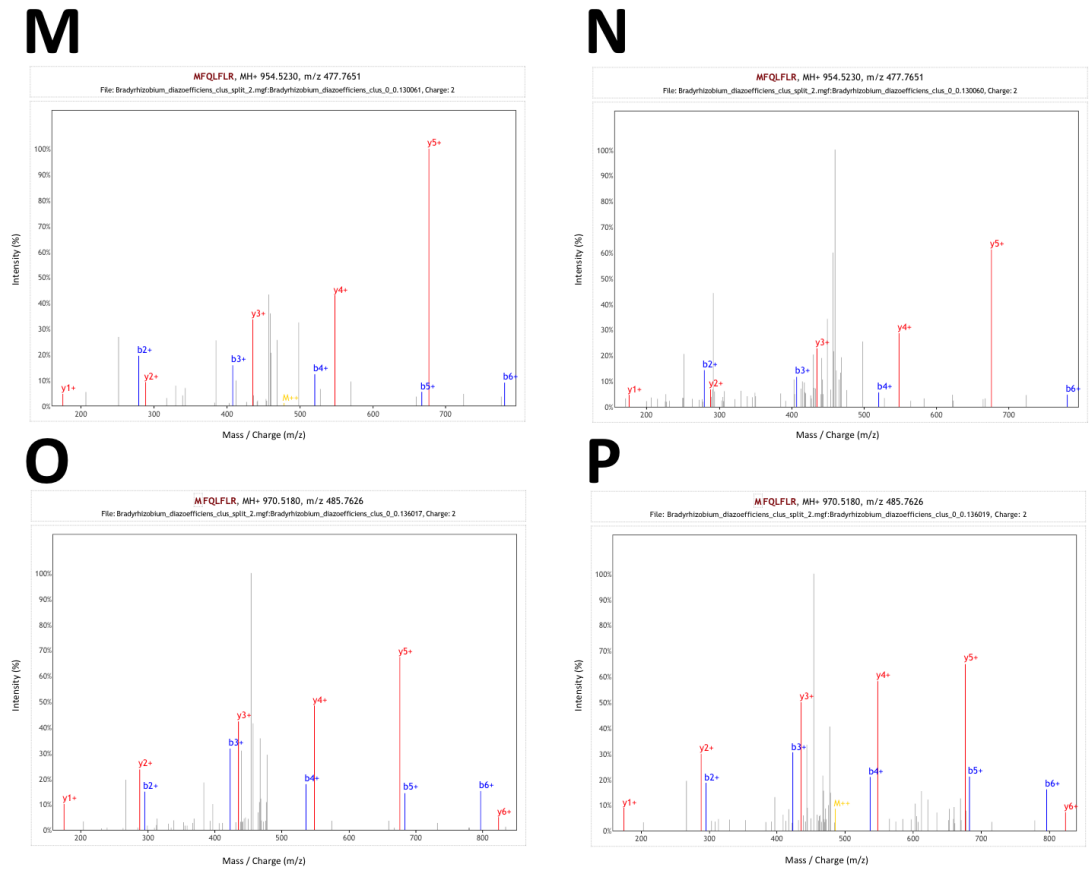

FGGEQTFR, MH+ 941.4476, m/z 471.2274
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.123471, Charge: 2
Intensity (%)
Mass / Charge (m/z)

K


FGGEQTFR, MH+ 941.4476, m/z 471.2274
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.123474, Charge: 2
Intensity (%)
Mass / Charge (m/z)

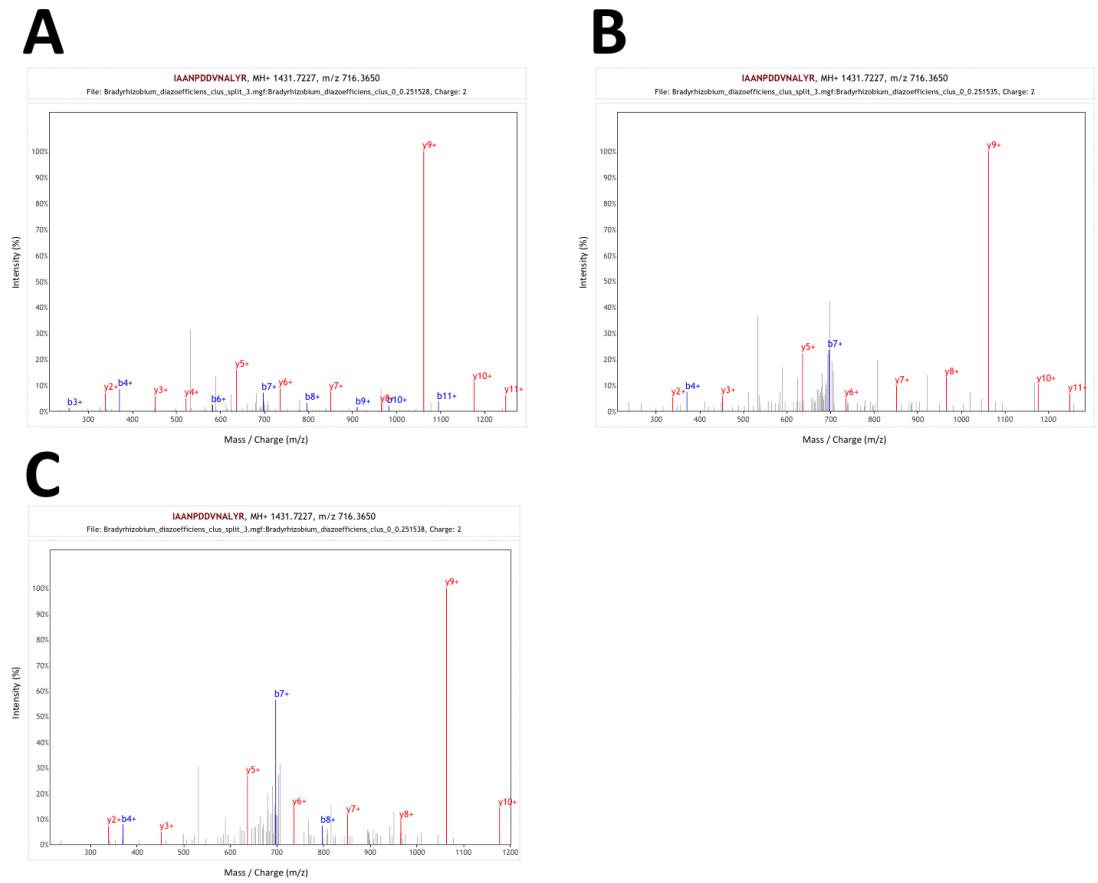
L


FGGEQTFR, MH+ 941.4476, m/z 471.2274
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.123476, Charge: 2
Intensity (%)
Mass / Charge (m/z)



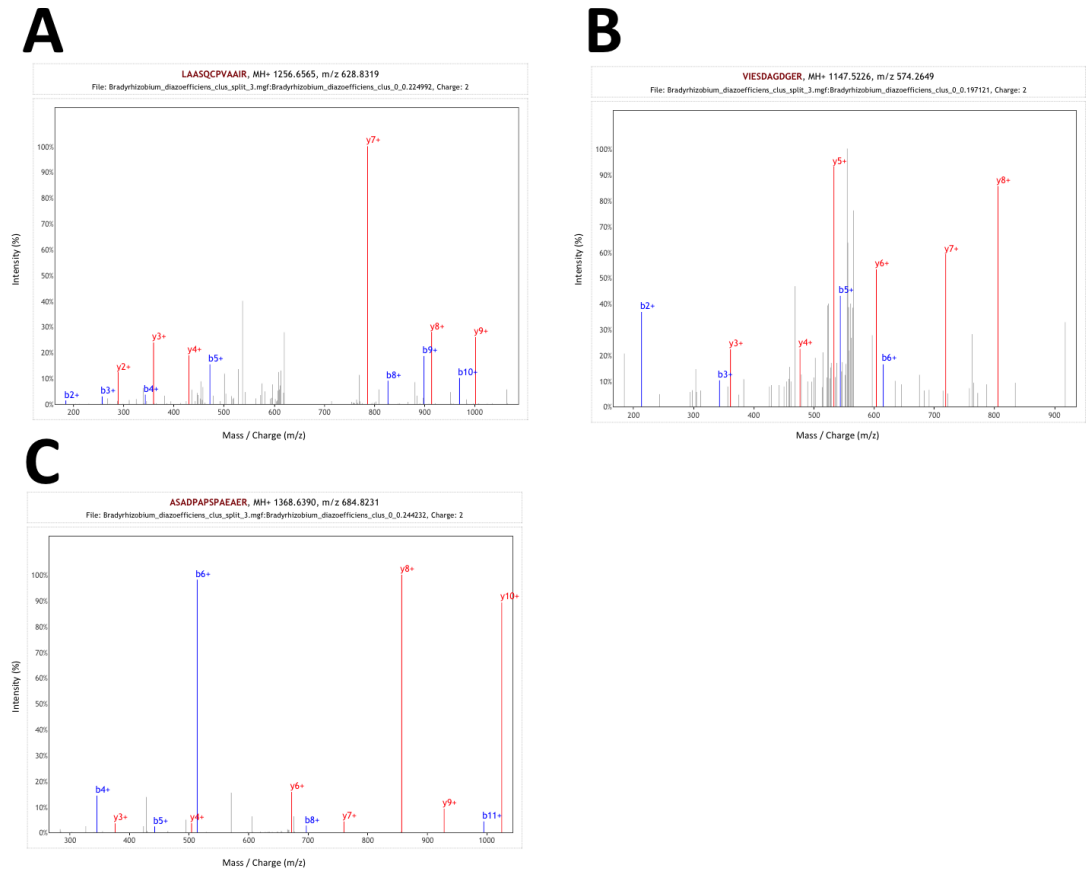
Appendix Figure 4.8 Supporting MS/MS spectra for reverse strand or novel gene event

Sixteen MS/MS spectra (A-P) supporting the eight novel peptides annotating the reverse strand or novel gene annotation event illustrated in Figure 4.1.



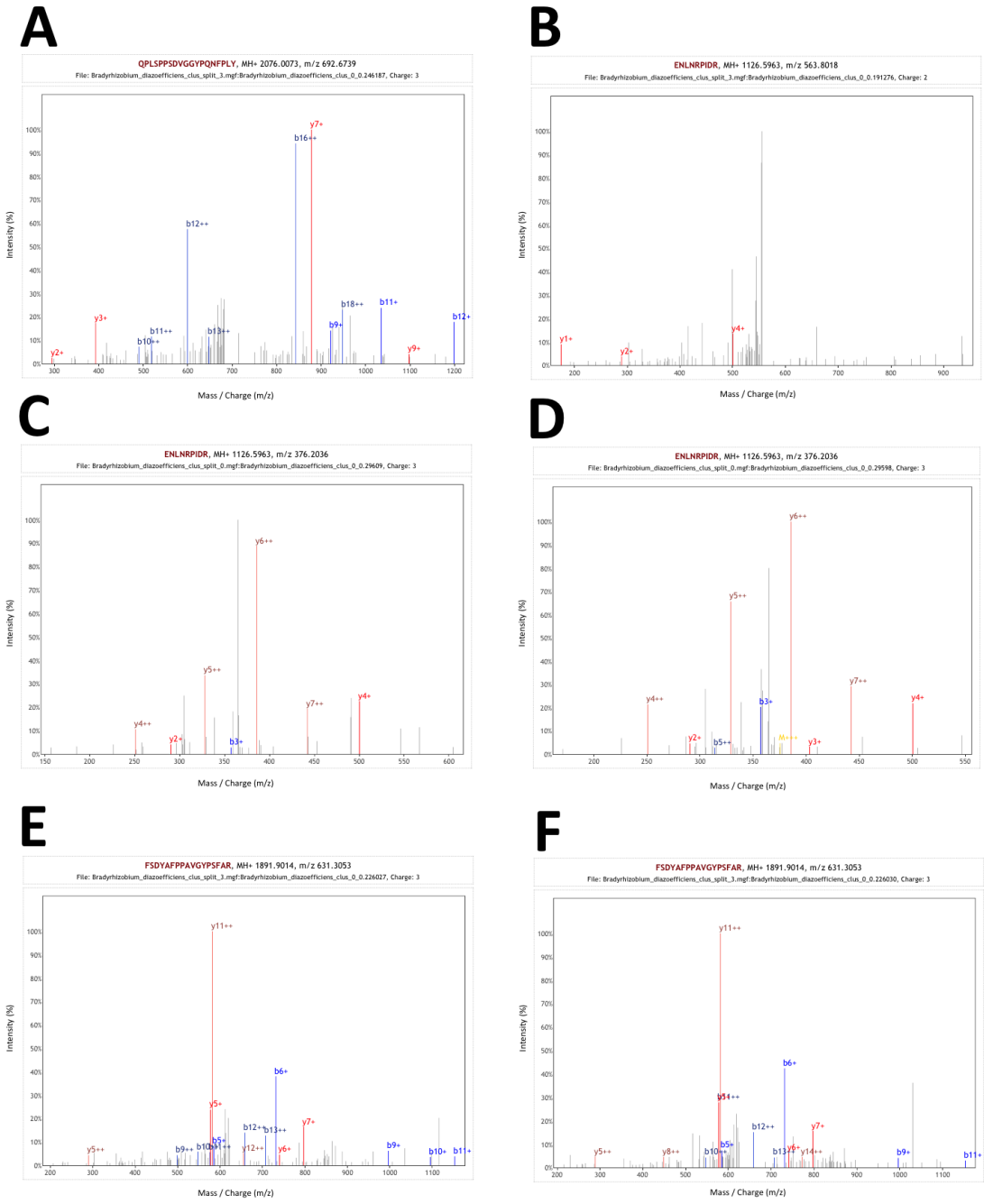
Appendix Figure 4.9 Supporting MS/MS spectra for exon boundary and frame-shift annotation event or sequencing error

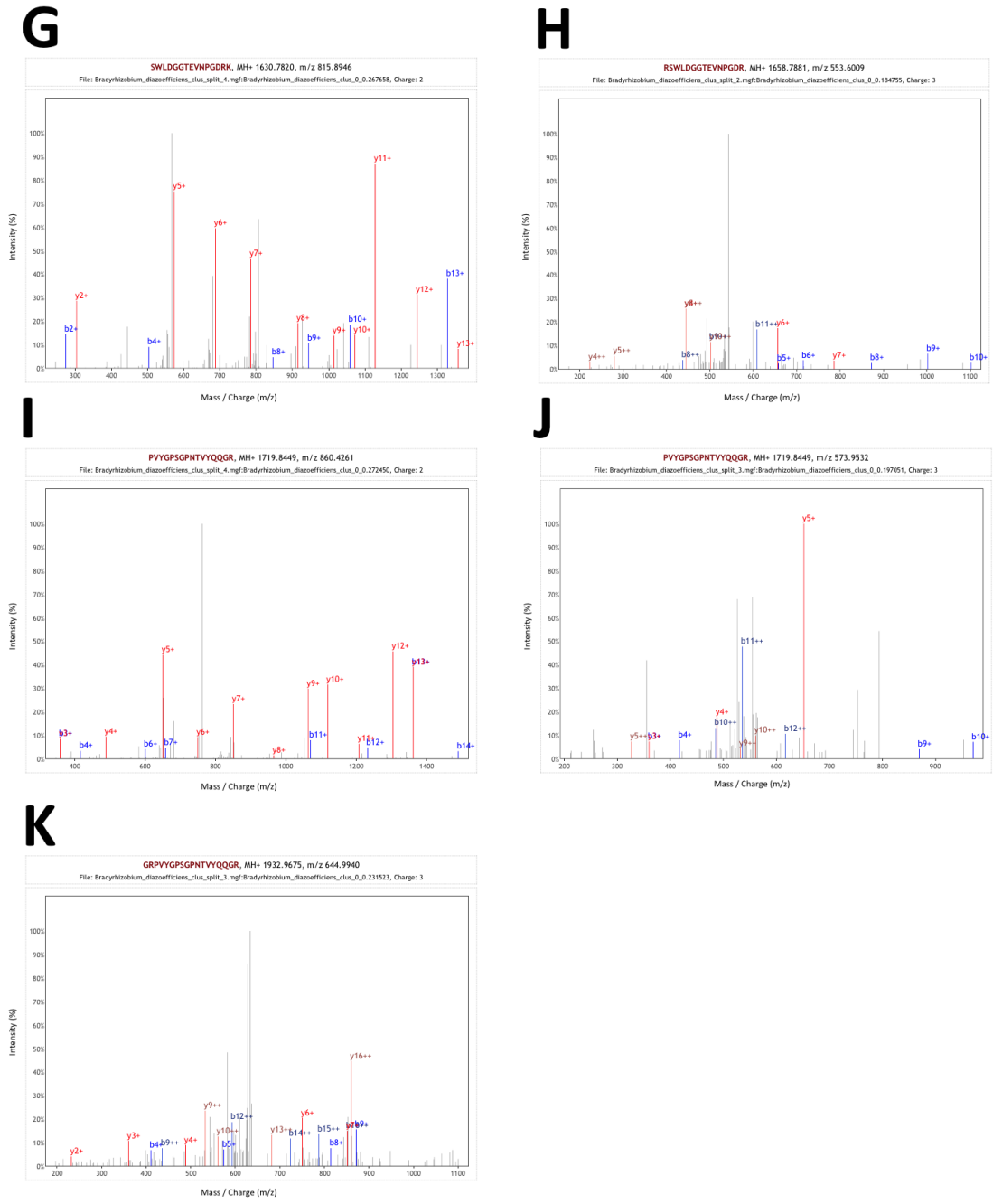
Three MS/MS spectra (A-C) supporting the novel peptide annotating the exon boundary and frame-shift annotation event or sequencing error illustrated in Figure 4.2.



Appendix Figure 4.10 Supporting MS/MS spectra for gene boundary or novel gene annotation event

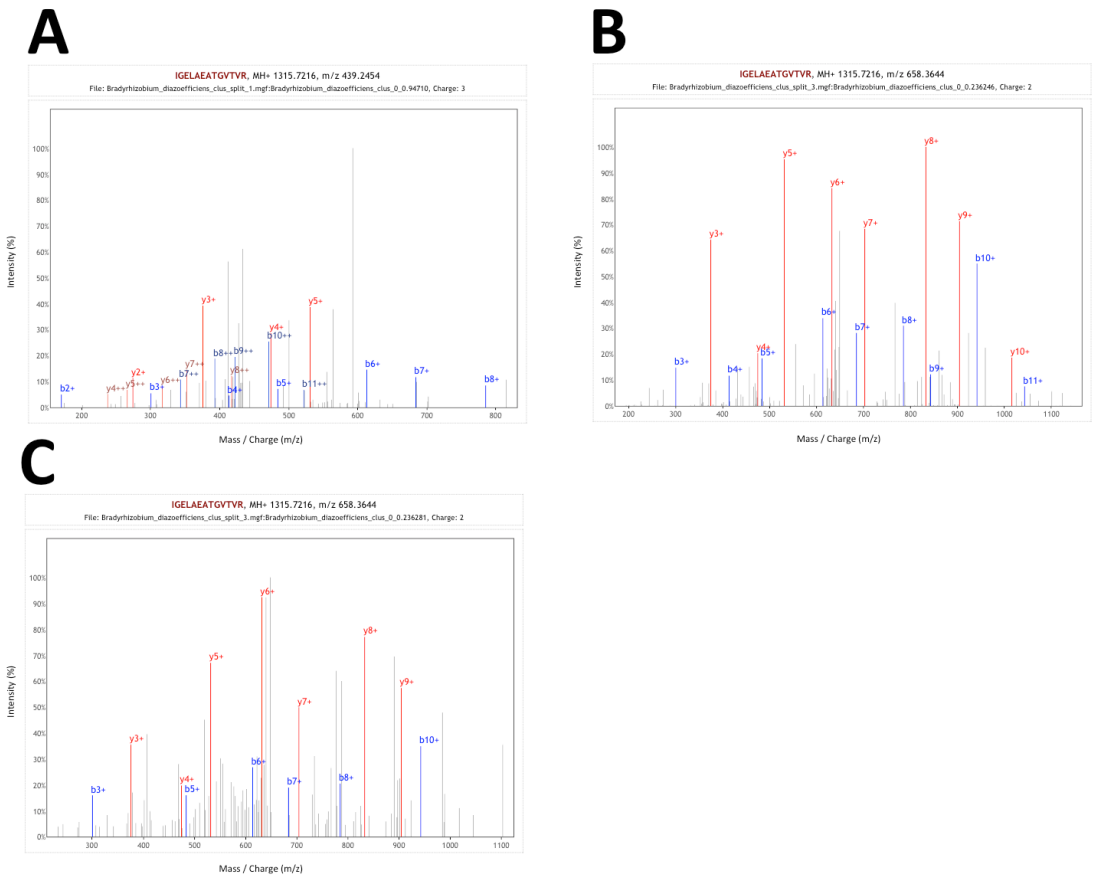
Three MS/MS spectra (A-C) supporting the three novel peptides annotating the gene boundary event of blr2146 or novel gene illustrated in Appendix Figure 4.3.


A
QPLSPPSDVGGYPQNFPLY, MH+ 2076.0073, m/z 692.6739
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.246187, Charge: 3
B
ENLNRPIDR, MH+ 1126.5963, m/z 563.8018
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.191276, Charge: 2
C
ENLNRPIDR, MH+ 1126.5963, m/z 376.2036
File: Bradyrhizobium_diazoefficiens_clus_split_0.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.29609, Charge: 3
D
ENLNRPIDR, MH+ 1126.5963, m/z 376.2036
File: Bradyrhizobium_diazoefficiens_clus_split_0.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.29598, Charge: 3
E
FSDYAFPPAVGYPSFAR, MH+ 1891.9014, m/z 631.3053
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.226027, Charge: 3
F
FSDYAFPPAVGYPSFAR, MH+ 1891.9014, m/z 631.3053
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.226030, Charge: 3
Intensity (%)
Mass / Charge (m/z)



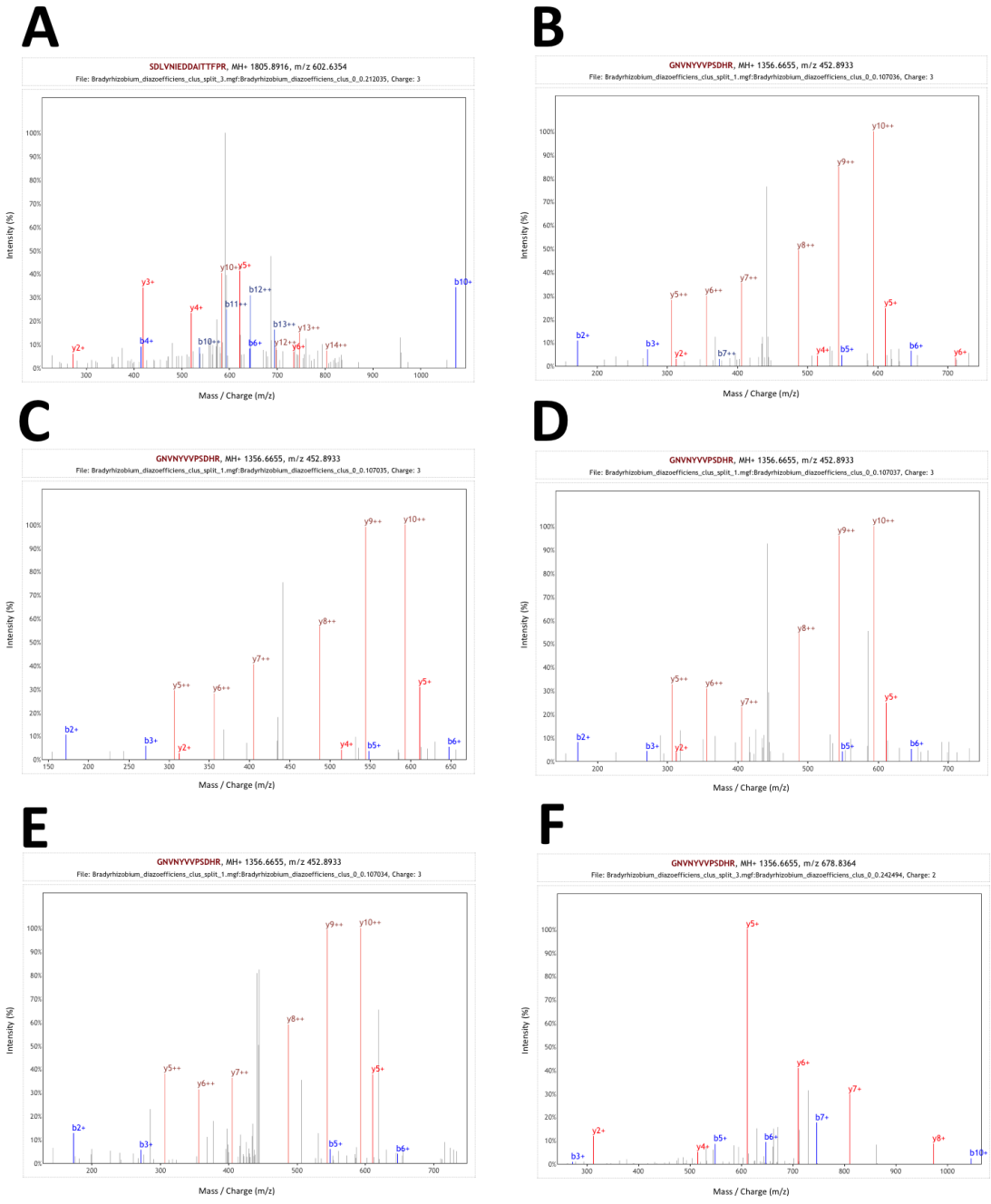
Appendix Figure 4.11 Supporting MS/MS spectra for a novel gene event

Eleven MS/MS spectra (A-K) supporting the seven novel peptides annotating a novel gene event illustrated in Appendix Figure 4.4.

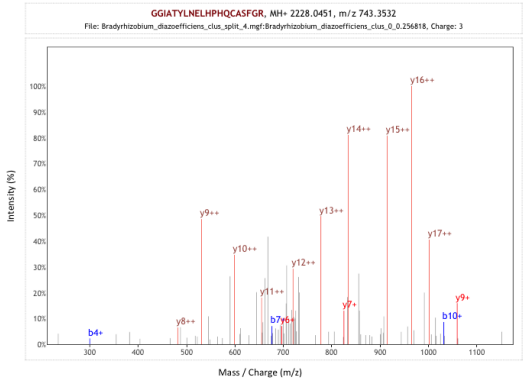


Appendix Figure 4.12 Supporting MS/MS spectra for exon boundary event of gene bll2019 (*NoIA*)

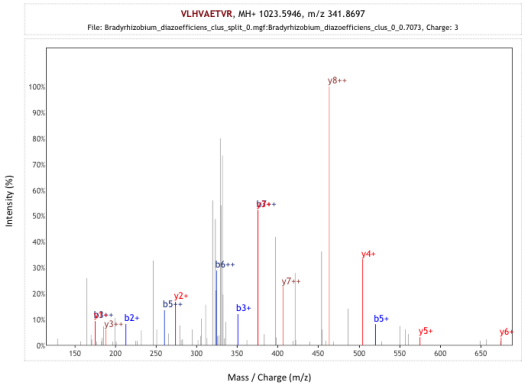
Three MS/MS spectra (A-C) supporting the novel peptide annotating the exon boundary event of bll2019 (*NoIA*) illustrated in Appendix Figure 4.5A.


A
SDLVNIEDDAITTFPR, MH+ 1805.8916, m/z 602.6354
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.212035, Charge: 3
B
GNVNYVVPSDHR, MH+ 1356.6655, m/z 452.8933
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.107036, Charge: 3
C
GNVNYVVPSDHR, MH+ 1356.6655, m/z 452.8933
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.107035, Charge: 3
D
GNVNYVVPSDHR, MH+ 1356.6655, m/z 452.8933
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.107037, Charge: 3
E
GNVNYVVPSDHR, MH+ 1356.6655, m/z 452.8933
File: Bradyrhizobium_diazoefficiens_clus_split_1.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.107034, Charge: 3
F
GNVNYVVPSDHR, MH+ 1356.6655, m/z 678.8364
File: Bradyrhizobium_diazoefficiens_clus_split_3.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.242494, Charge: 2
Intensity (%)
Mass / Charge (m/z)

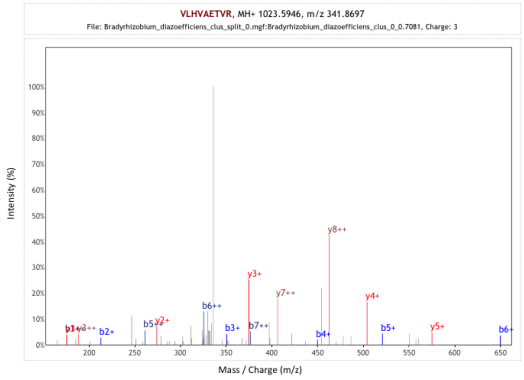
G


GGIATYLNELHPHQCASFGR, MH+ 2228.0451, m/z 743.3532
File: Bradyrhizobium_diazoefficiens_clus_split_4.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.256818, Charge: 3
Intensity (%)
Mass / Charge (m/z)

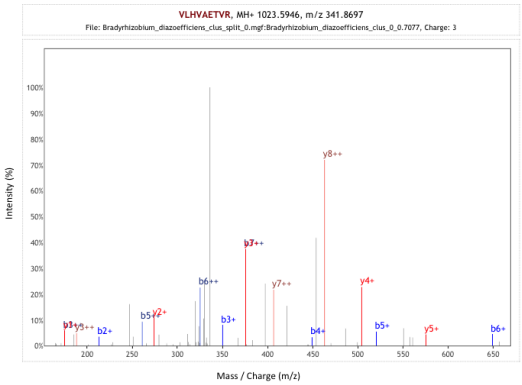
H


VLHVAETVR, MH+ 1023.5946, m/z 341.8697
File: Bradyrhizobium_diazoefficiens_clus_split_0.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.7073, Charge: 3
Intensity (%)
Mass / Charge (m/z)

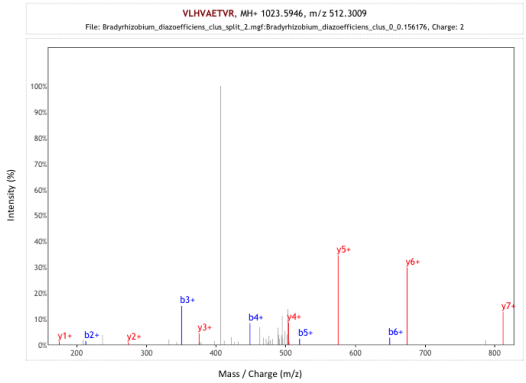
I


VLHVAETVR, MH+ 1023.5946, m/z 341.8697
File: Bradyrhizobium_diazoefficiens_clus_split_0.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.7081, Charge: 3
Intensity (%)
Mass / Charge (m/z)

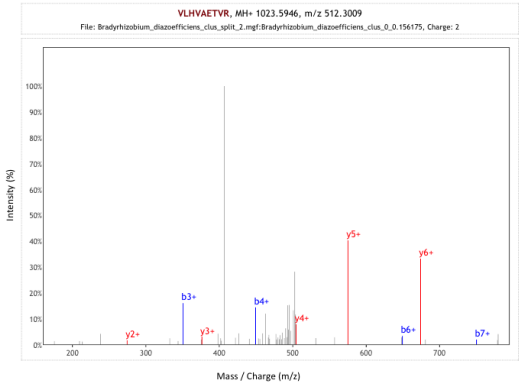
J

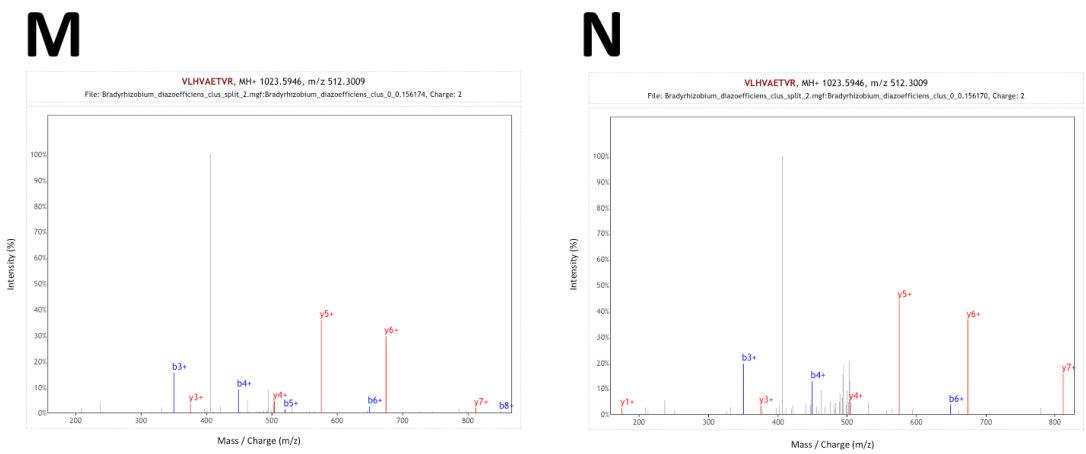

VLHVAETVR, MH+ 1023.5946, m/z 341.8697
File: Bradyrhizobium_diazoefficiens_clus_split_0.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.7077, Charge: 3
Intensity (%)
Mass / Charge (m/z)

K


VLHVAETVR, MH+ 1023.5946, m/z 512.3009
File: Bradyrhizobium_diazoefficiens_clus_split_2.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.156176, Charge: 2
Intensity (%)
Mass / Charge (m/z)

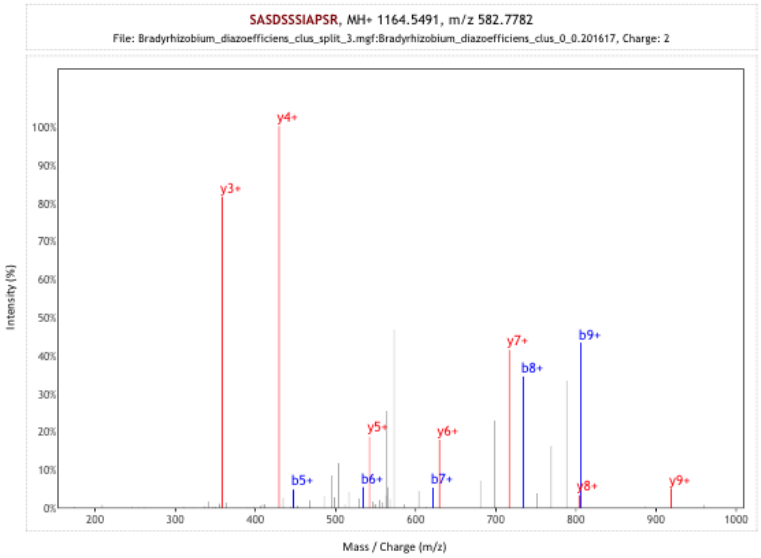
L


VLHVAETVR, MH+ 1023.5946, m/z 512.3009
File: Bradyrhizobium_diazoefficiens_clus_split_2.mgf:Bradyrhizobium_diazoefficiens_clus_0_0.156175, Charge: 2
Intensity (%)
Mass / Charge (m/z)



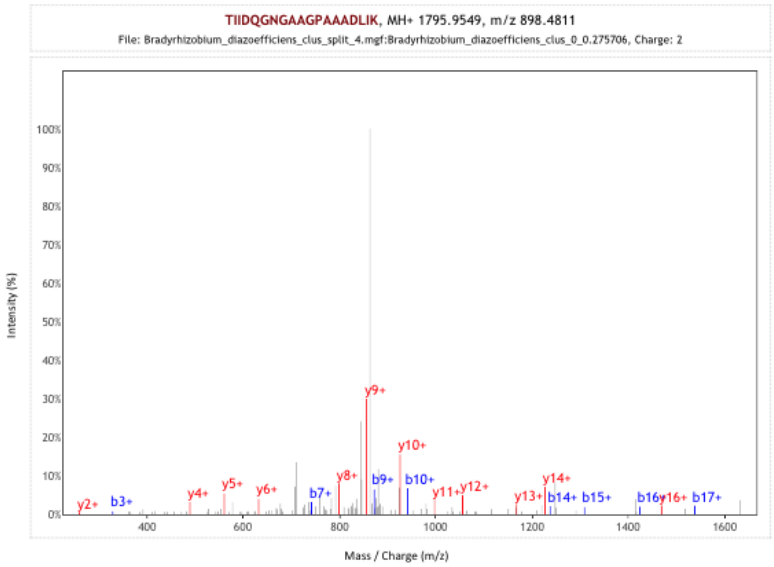
Appendix Figure 4.13 Supporting MS/MS spectra for exon boundary event of gene bll2380

Fourteen MS/MS spectra (A-N) supporting the four novel peptides annotating the exon boundary event of bll2380 illustrated in Appendix Figure 4.5B.



Appendix Figure 4.14 Supporting MS/MS spectrum for gene boundary event of gene bll0794 (*PhoH*)

One MS/MS spectrum supporting the novel peptide annotating the gene boundary event of bll0794 (*PhoH*) illustrated in Appendix Figure 4.6.



Appendix Figure 4.15 Supporting MS/MS spectrum for peptide "TIIDQGNGAAGPAAADLIK", indicating no N-terminal acetylation

One MS/MS spectrum supporting a known peptide with no N-terminal acetylation, in contrast to the study from [502], illustrated in Appendix Figure 4.7.

Appendix File 5.1 Reference predictions are in zip file 'AppendixFile5.1.zip' on the DVD provided.

Appendix File 5.2 Clustering, quality filtering and precursor mass tolerance optimization results are in excel file 'AppendixFile5.2.xlsx' on the DVD provided.

Appendix File 5.3 Processed proteogenomics results are in excel file 'AppendixFile5.3.xlsx' on the DVD provided.

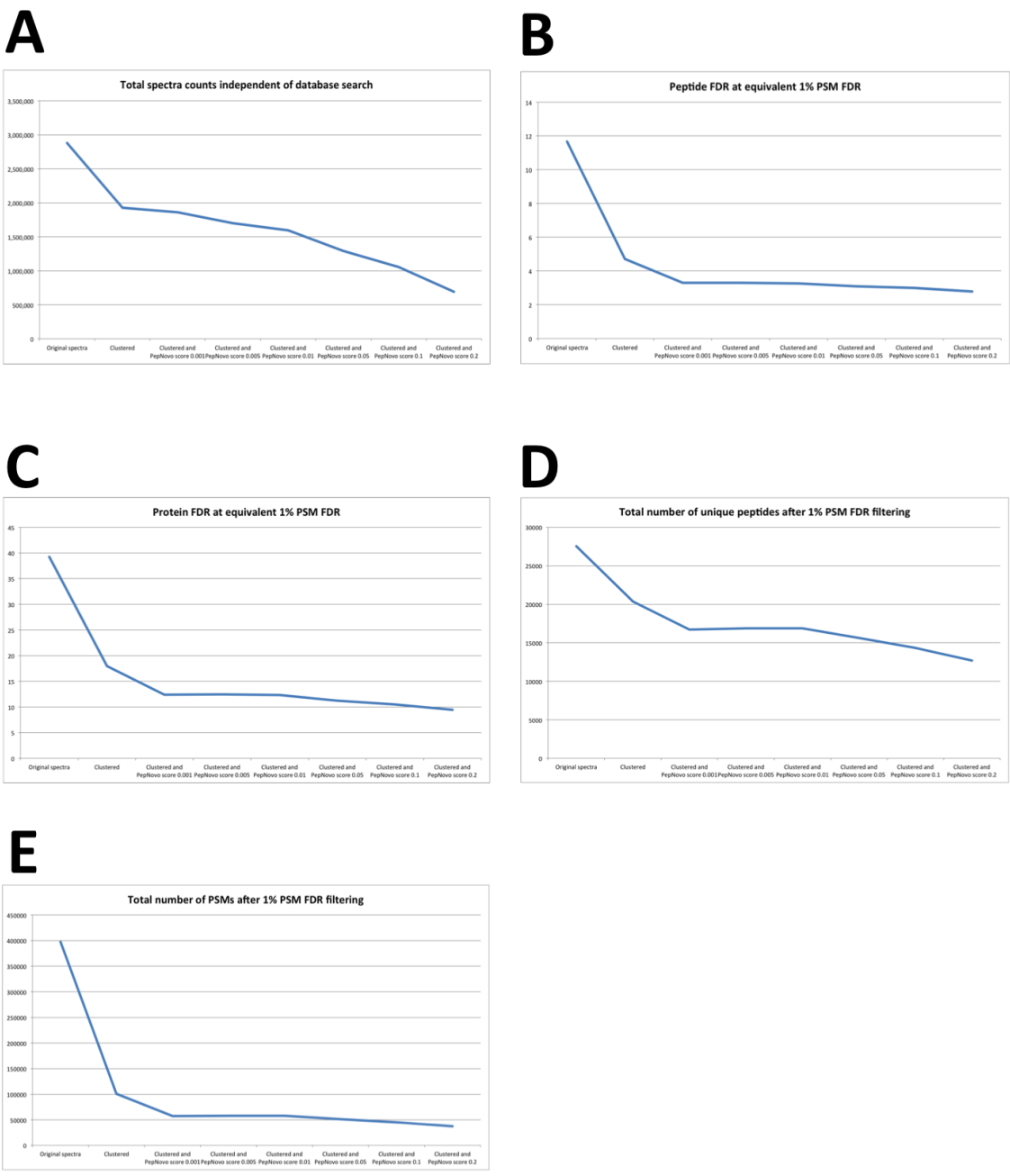
Appendix File 5.4 Raw proteogenomics results are in zip file 'AppendixFile5.4.zip' on the DVD provided.

Appendix File 5.5 Augustus gene predictions are in zip file 'AppendixFile5.5.zip' on the DVD provided.

Appendix File 5.6 Augustus gene predictions with incorporated novel peptides are in excel file 'AppendixFile5.6.xlsx' on the DVD provided.

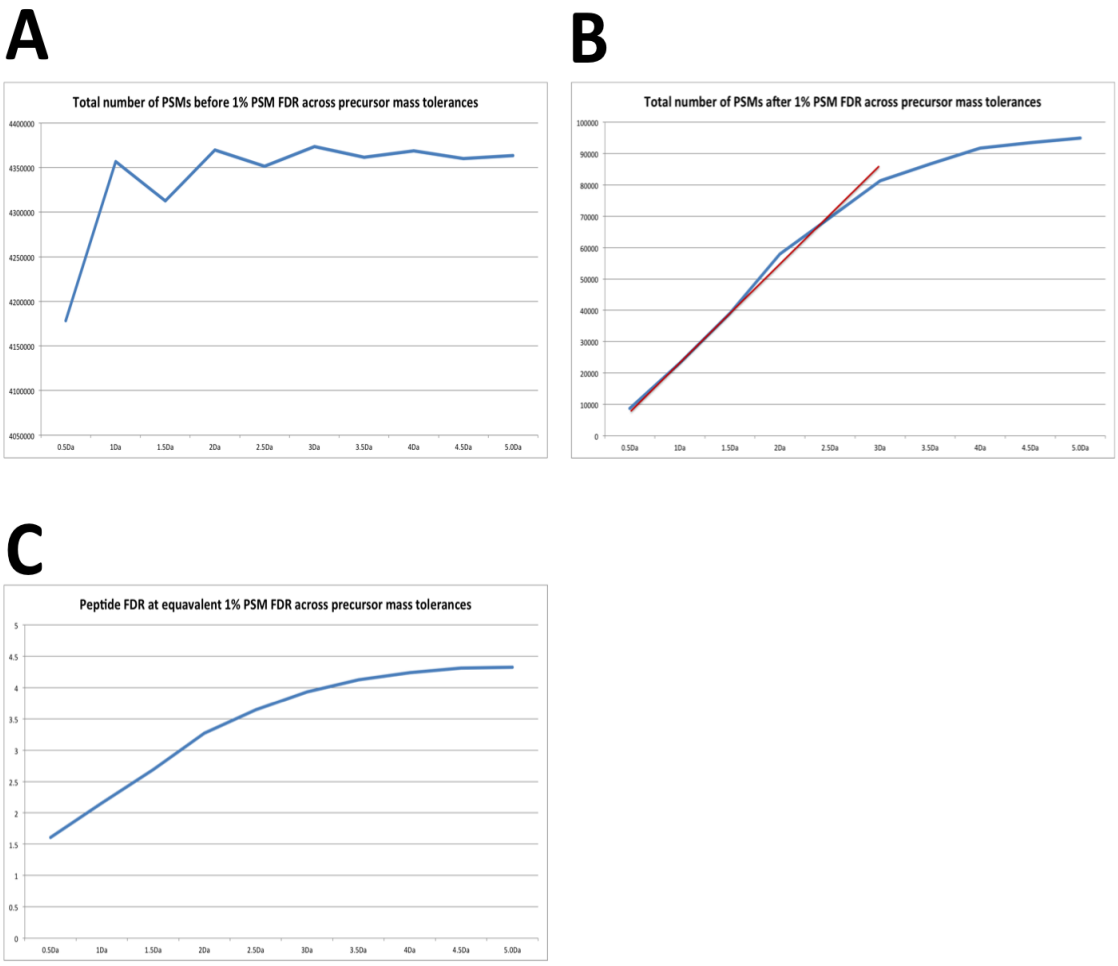
Appendix File 5.7 NetGene2 splice site prediction results are in zip file 'AppendixFile5.7.zip' on the DVD provided.

Appendix File 5.8 Known N-terminal acetylated peptides are in excel file 'AppendixFile5.8.xlsx' on the DVD.



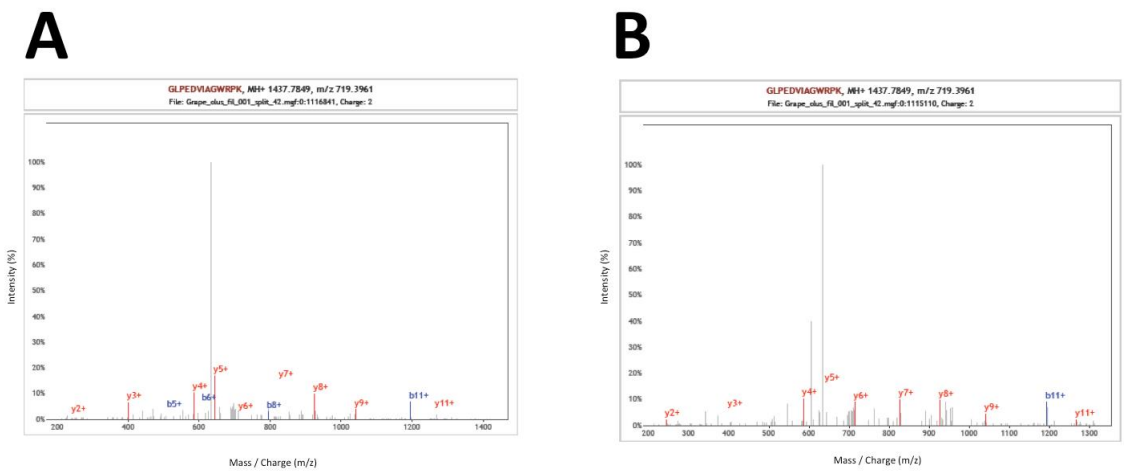
Appendix Figure 5.1 Pre and post clustering with and without PepNovo quality filtering

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



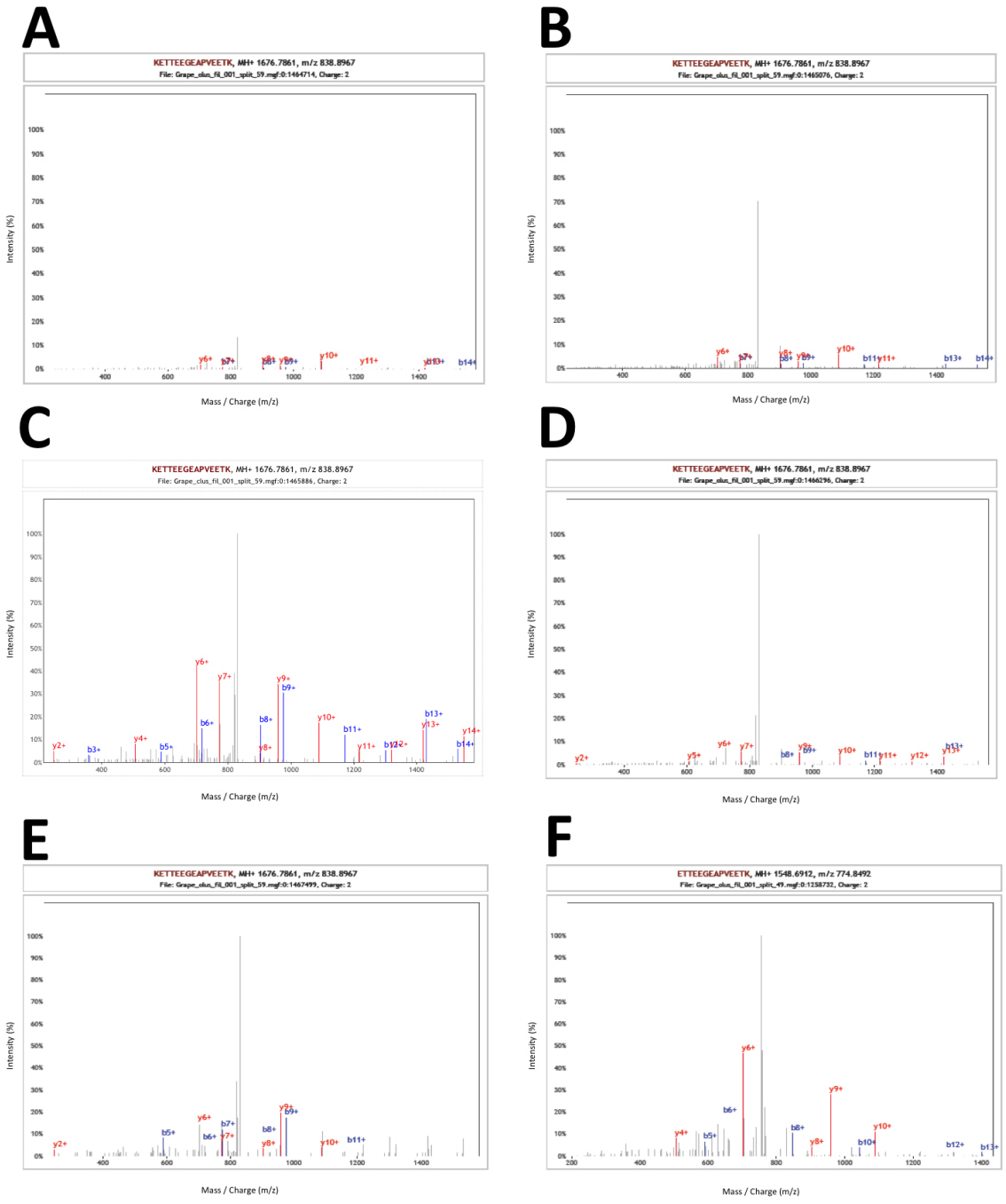
Appendix Figure 5.2 Precursor mass tolerance optimisation

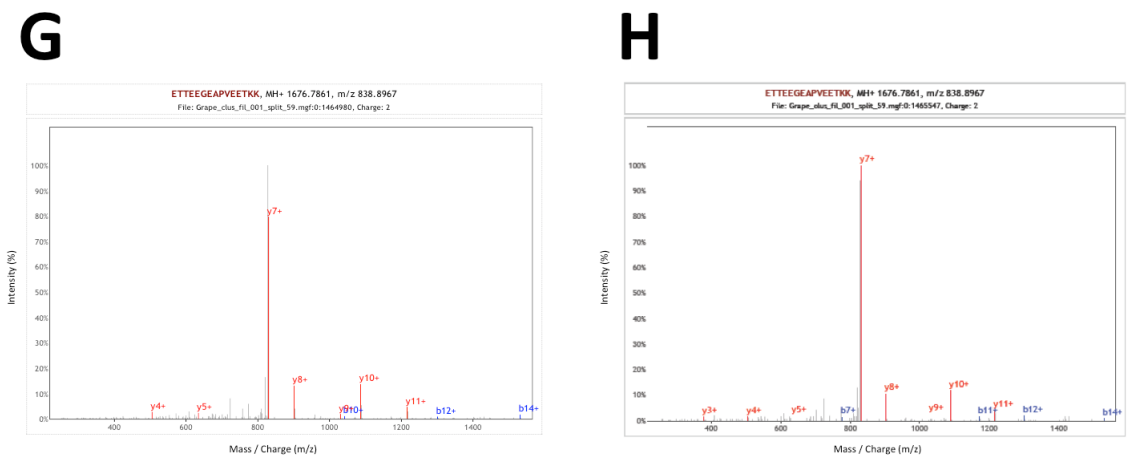
(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.



Appendix Figure 5.3 Supporting MS/MS spectra for a novel gene event

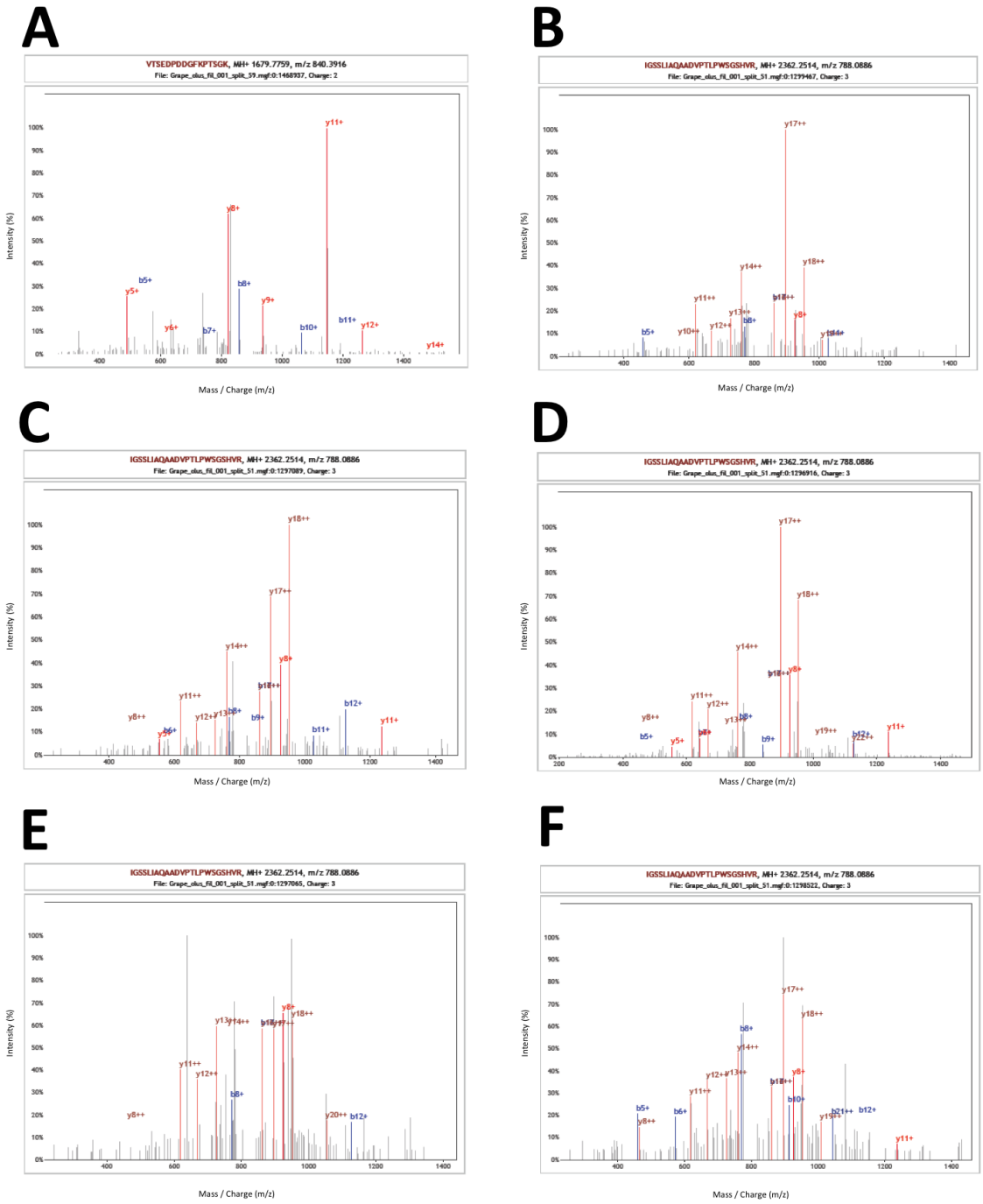
Two MS/MS spectra (A-B) supporting a novel peptide annotating a novel gene event, illustrated in Figure 5.1.


A
KETTEEGEAPVEETK, MH+ 1676.7861, m/z 838.8967
File: Grape_clus_fil_001_split_59.mgf:0:1464714, Charge: 2
B
KETTEEGEAPVEETK, MH+ 1676.7861, m/z 838.8967
File: Grape_clus_fil_001_split_59.mgf:0:1465076, Charge: 2
C
KETTEEGEAPVEETK, MH+ 1676.7861, m/z 838.8967
File: Grape_clus_fil_001_split_59.mgf:0:1465886, Charge: 2
D
KETTEEGEAPVEETK, MH+ 1676.7861, m/z 838.8967
File: Grape_clus_fil_001_split_59.mgf:0:1466296, Charge: 2
E
KETTEEGEAPVEETK, MH+ 1676.7861, m/z 838.8967
File: Grape_clus_fil_001_split_59.mgf:0:1467499, Charge: 2
F
ETTEEGEAPVEETK, MH+ 1548.6912, m/z 774.8492
File: Grape_clus_fil_001_split_49.mgf:0:1258732, Charge: 2
Intensity (%)
Mass / Charge (m/z)

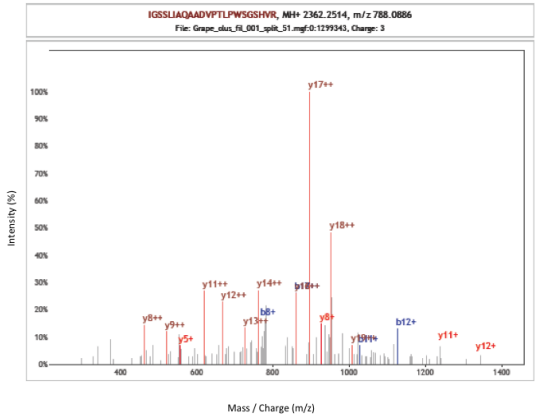


Appendix Figure 5.4 Supporting MS/MS spectra for a novel gene annotation misidentified as a reverse strand event

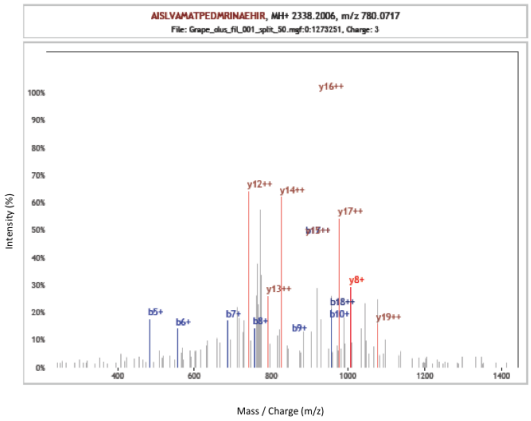
Eight MS/MS spectra (A-H) supporting novel peptides annotating a novel gene event via a reverse strand event, illustrated in Figure 5.2.


A
VTSEDPDDGFKPTSGK, MH+ 1679.7759, m/z 840.3916
File: Grape_olus_Fil_001_split_59.mgf:0:1468937, Charge: 2
B
IGSSLIAQAADVPTLPWSGSHVR, MH+ 2362.2514, m/z 788.0886
File: Grape_olus_Fil_001_split_51.mgf:0:1299467, Charge: 3
C
IGSSLIAQAADVPTLPWSGSHVR, MH+ 2362.2514, m/z 788.0886
File: Grape_olus_Fil_001_split_51.mgf:0:1297089, Charge: 3
D
IGSSLIAQAADVPTLPWSGSHVR, MH+ 2362.2514, m/z 788.0886
File: Grape_olus_Fil_001_split_51.mgf:0:1296916, Charge: 3
E
IGSSLIAQAADVPTLPWSGSHVR, MH+ 2362.2514, m/z 788.0886
File: Grape_olus_Fil_001_split_51.mgf:0:1297065, Charge: 3
F
IGSSLIAQAADVPTLPWSGSHVR, MH+ 2362.2514, m/z 788.0886
File: Grape_olus_Fil_001_split_51.mgf:0:1298522, Charge: 3
Intensity (%)
Mass / Charge (m/z)

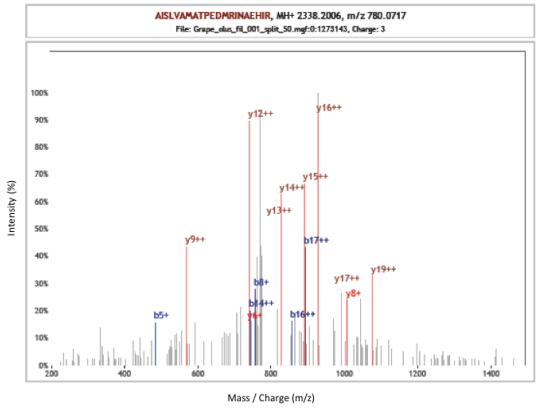
G


IGSSLIAQAADVPTLPWSGSHVR, MH+ 2362.2514, m/z 788.0886
File: Grape_clus_fil_001_split_51.mgf:0:1299343, Charge: 3
Intensity (%)
Mass / Charge (m/z)

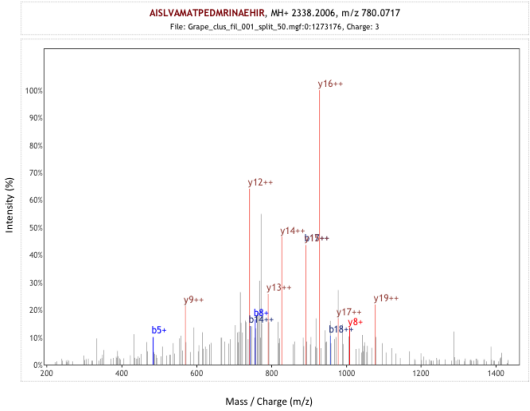
H


AISLVAMATPEDMRINAEHIR, MH+ 2338.2006, m/z 780.0717
File: Grape_clus_fil_001_split_50.mgf:0:1273251, Charge: 3
Intensity (%)
Mass / Charge (m/z)

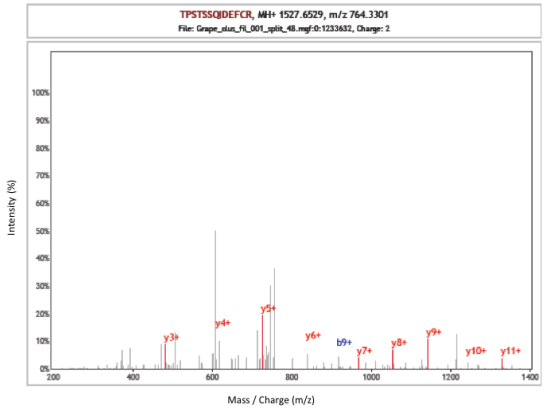
I


AISLVAMATPEDMRINAEHIR, MH+ 2338.2006, m/z 780.0717
File: Grape_clus_fil_001_split_50.mgf:0:1273143, Charge: 3
Intensity (%)
Mass / Charge (m/z)

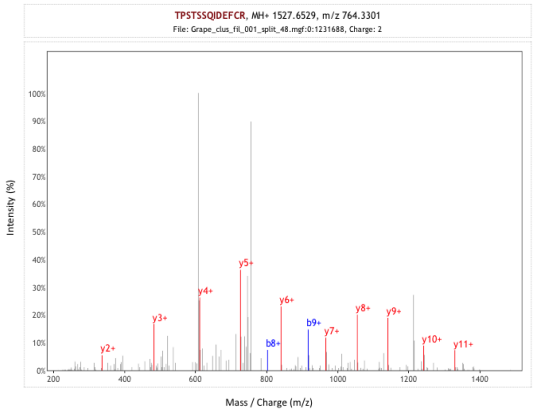
J

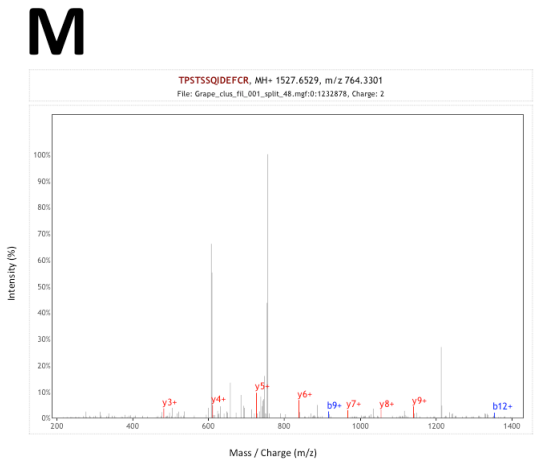

AISLVAMATPEDMRINAEHIR, MH+ 2338.2006, m/z 780.0717
File: Grape_clus_fil_001_split_50.mgf:0:1273176, Charge: 3
Intensity (%)
Mass / Charge (m/z)

K


TPSTSSQIDEFCR, MH+ 1527.6529, m/z 764.3301
File: Grape_clus_fil_001_split_48.mgf:0:1233632, Charge: 2
Intensity (%)
Mass / Charge (m/z)

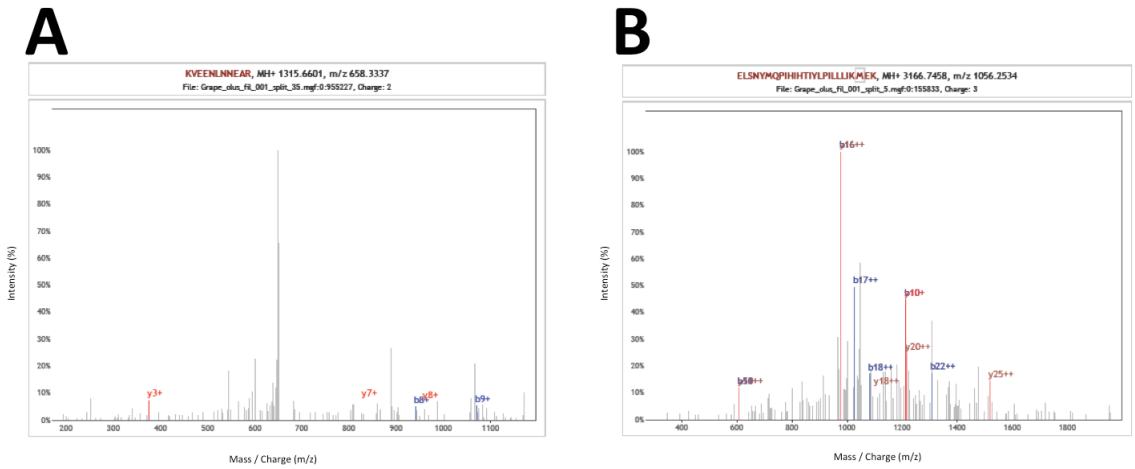
L


TPSTSSQIDEFCR, MH+ 1527.6529, m/z 764.3301
File: Grape_clus_fil_001_split_48.mgf:0:1231688, Charge: 2
Intensity (%)
Mass / Charge (m/z)



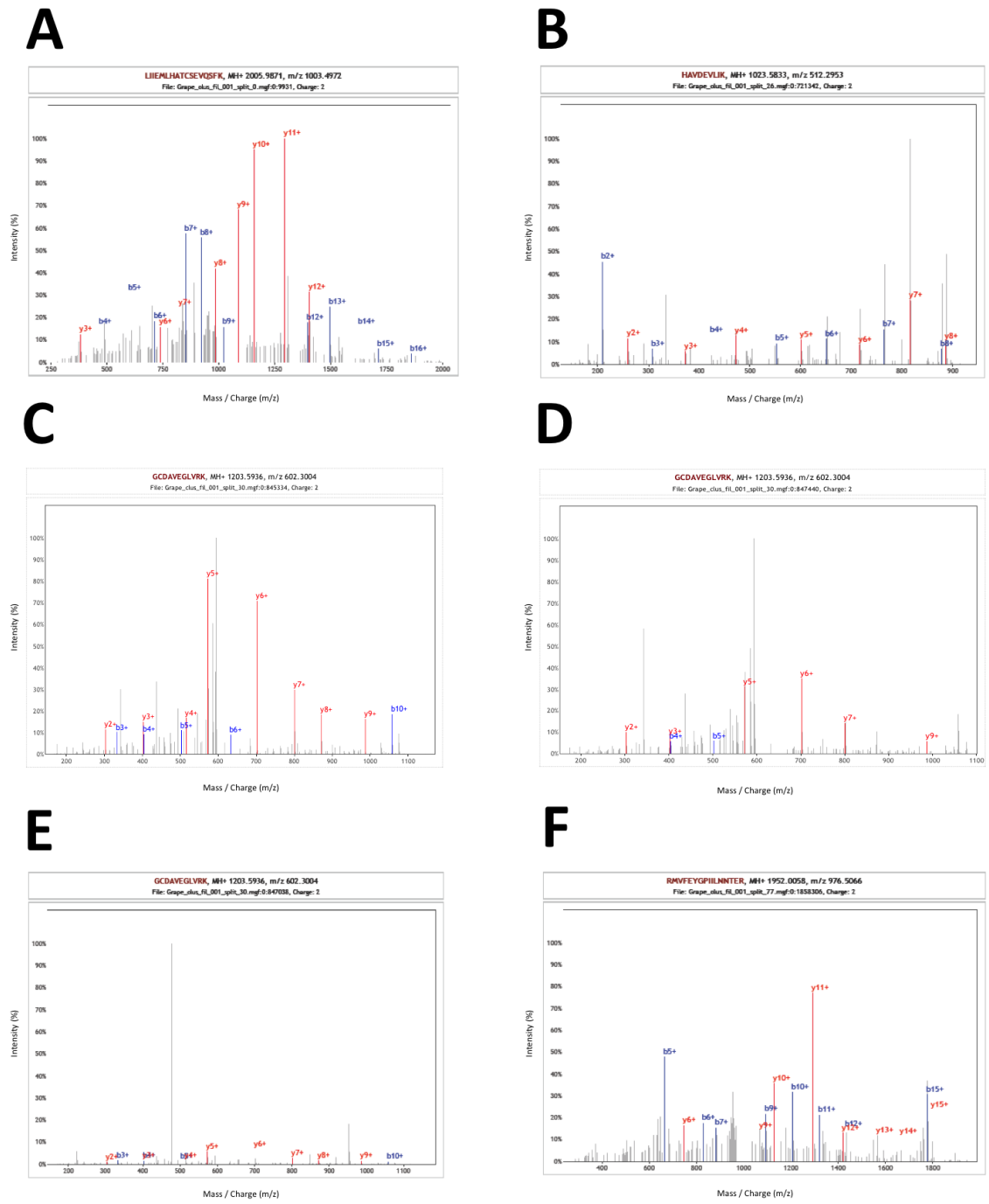
Appendix Figure 5.5 Supporting MS/MS spectra for a gene boundary annotation event

Thirteen MS/MS spectra (A-M) supporting novel peptides annotating a gene boundary event illustrated, in Figure 5.3.



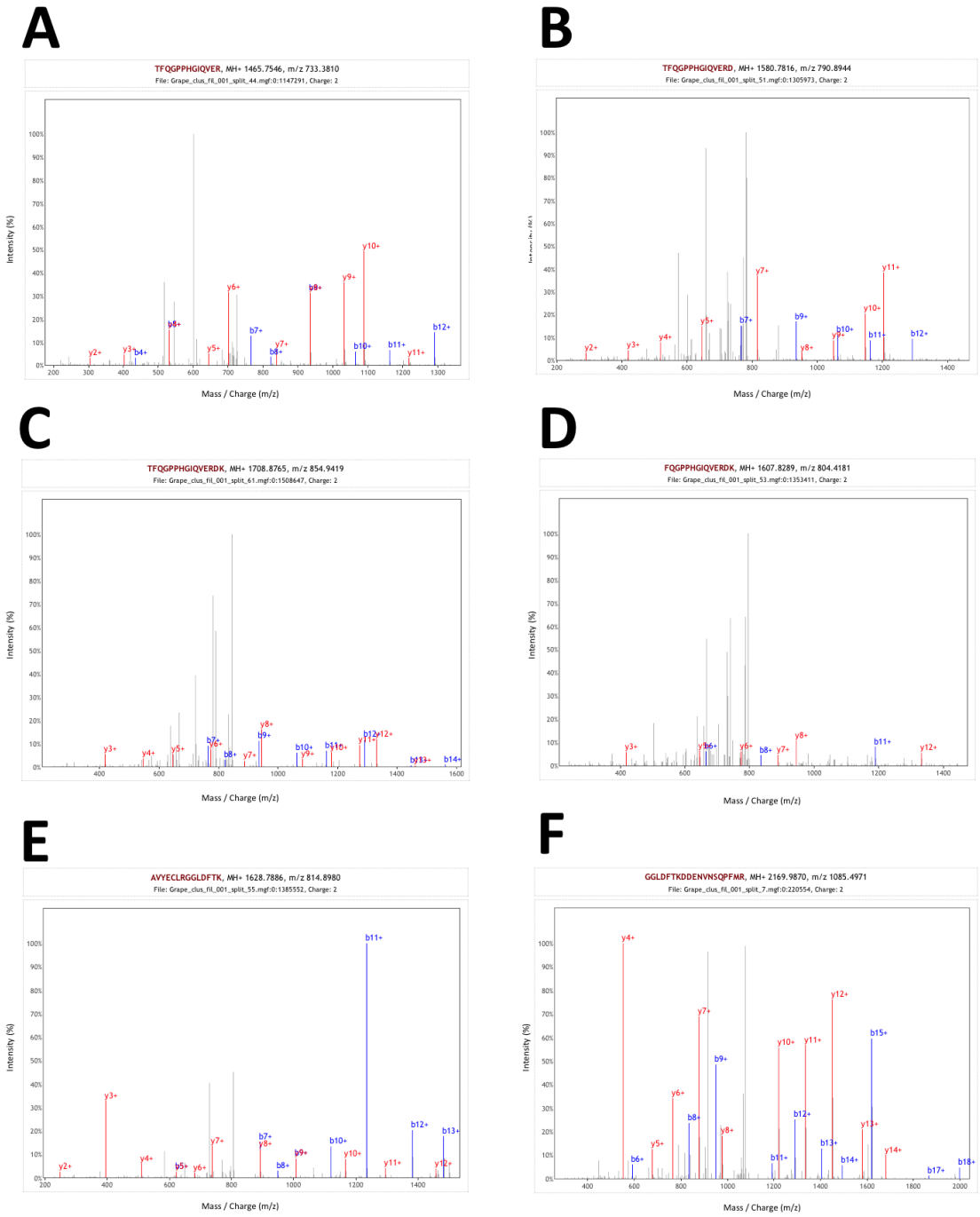
Appendix Figure 5.6 Supporting MS/MS spectra for a novel gene annotation via a reverse strand annotation event

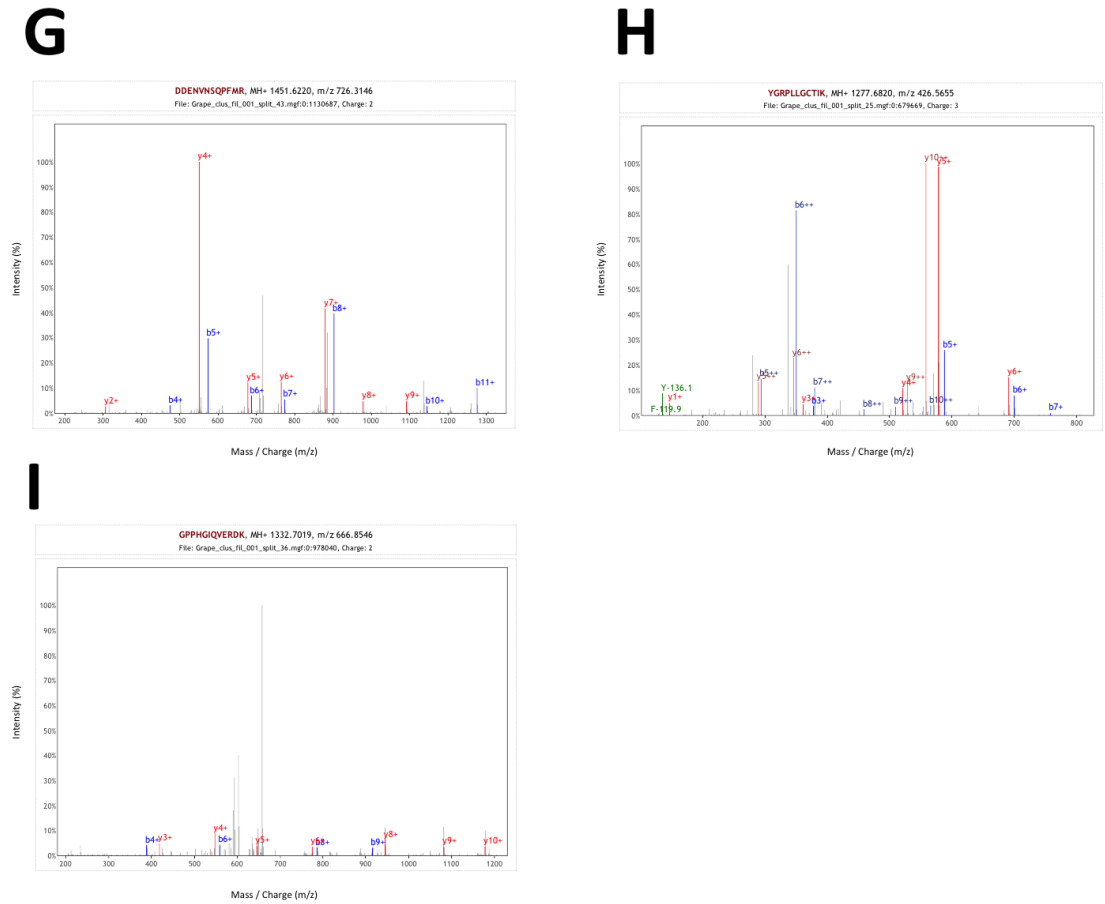
Two MS/MS spectra (A-B) supporting novel peptides annotating a novel gene event via a reverse strand event, illustrated in Figure 5.4.



Appendix Figure 5.7 Supporting MS/MS spectra for a translated UTR annotation event

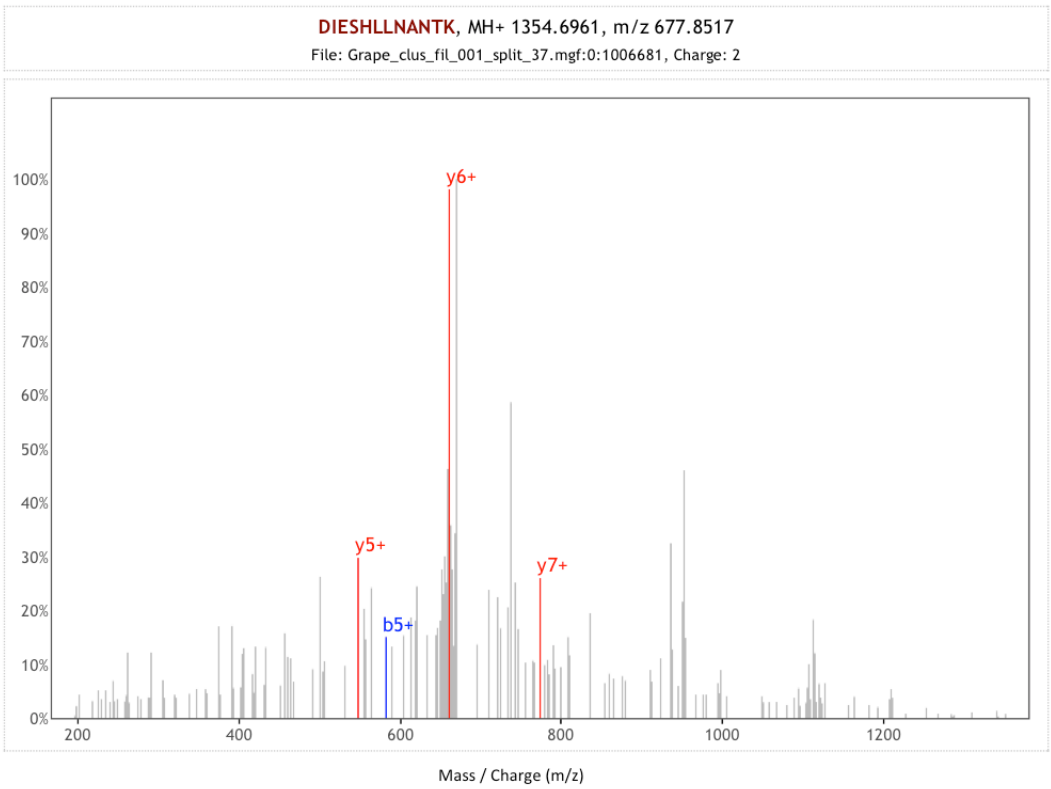
Six MS/MS spectra (A-F) supporting novel peptides annotating a translated UTR event, illustrated in Figure 5.5.


A
TFQGPPHGIQVER, MH+ 1465.7546, m/z 733.3810
File: Grape_clus_fil_001_split_44.mgf:0:1147291, Charge: 2
B
TFQGPPHGIQVERD, MH+ 1580.7816, m/z 790.8944
File: Grape_clus_fil_001_split_51.mgf:0:1305973, Charge: 2
C
TFQGPPHGIQVERDK, MH+ 1708.8765, m/z 854.9419
File: Grape_clus_fil_001_split_61.mgf:0:1508647, Charge: 2
D
FQGPPHGIQVERDK, MH+ 1607.8289, m/z 804.4181
File: Grape_clus_fil_001_split_53.mgf:0:1353411, Charge: 2
E
AVYECLRGGLDFTK, MH+ 1628.7886, m/z 814.8980
File: Grape_clus_fil_001_split_55.mgf:0:1385552, Charge: 2
F
GGLDFTKDDENVNSQPFMR, MH+ 2169.9870, m/z 1085.4971
File: Grape_clus_fil_001_split_7.mgf:0:220554, Charge: 2
Intensity (%)
Mass / Charge (m/z)



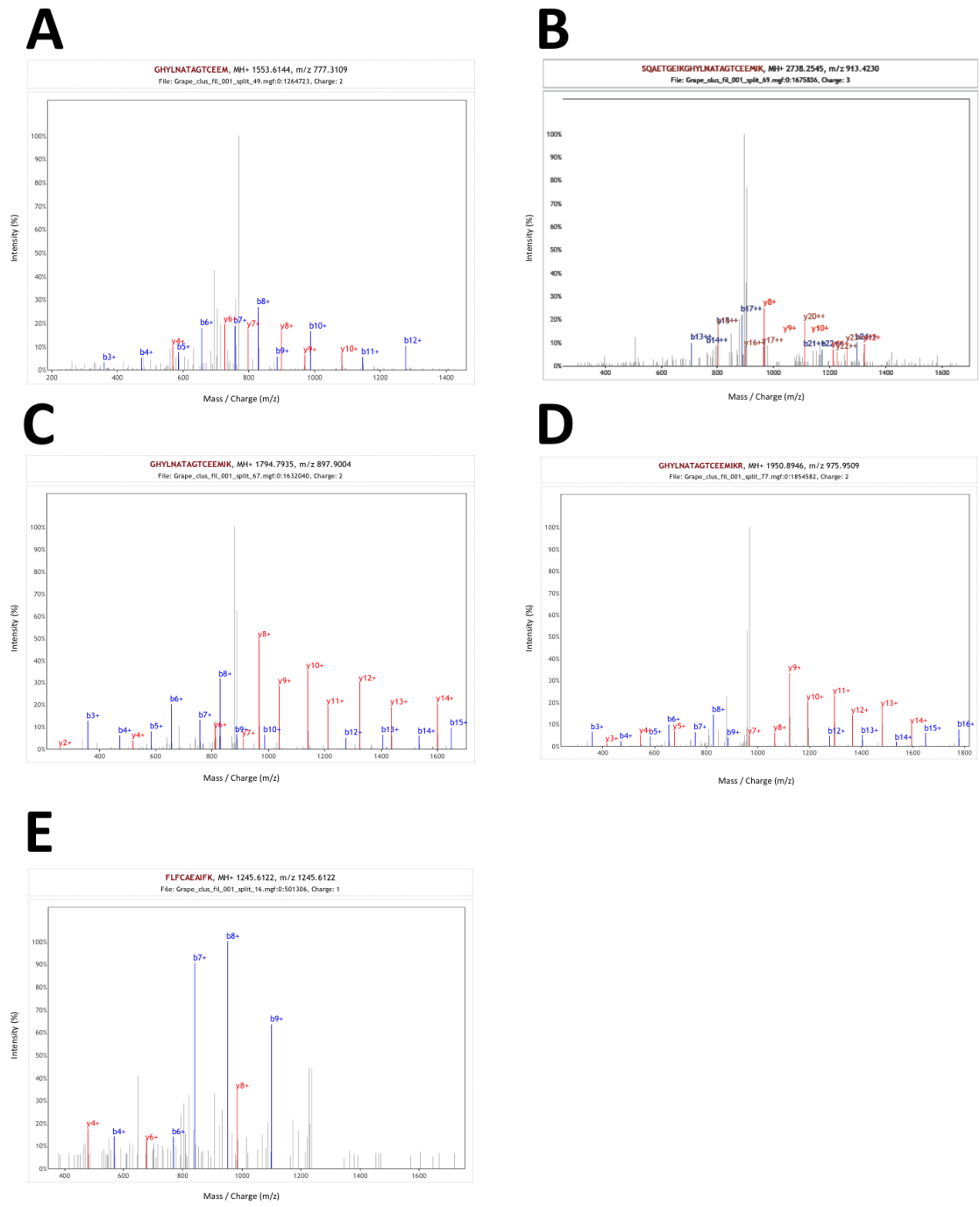
Appendix Figure 5.8 Supporting MS/MS spectra for a translated UTR annotation event

Nine representative MS/MS spectra (A-I; from a total of 90), supporting novel peptides annotating a translated UTR event, illustrated in Figure 5.6.



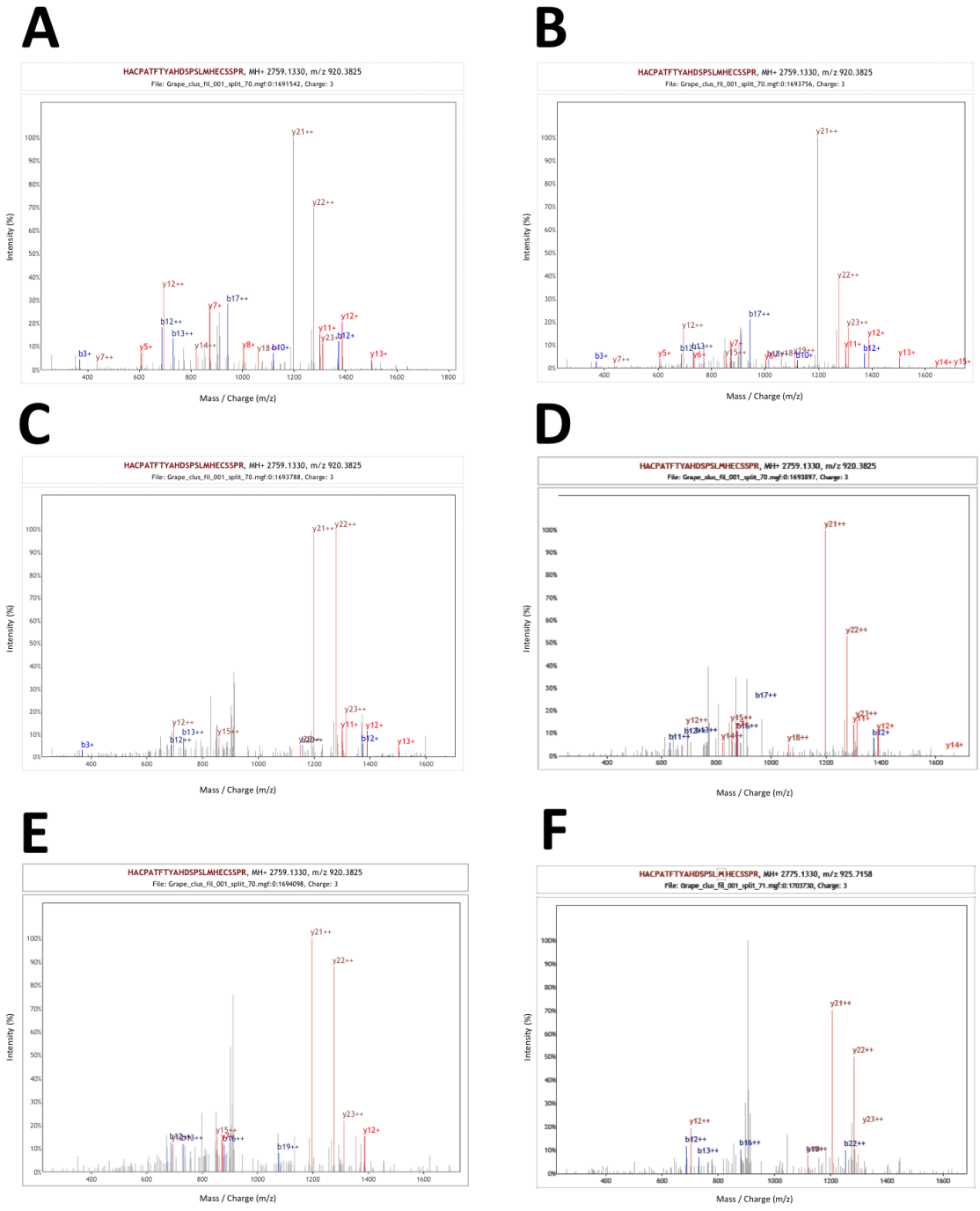
Appendix Figure 5.9 Supporting MS/MS spectra for a novel splice annotation event

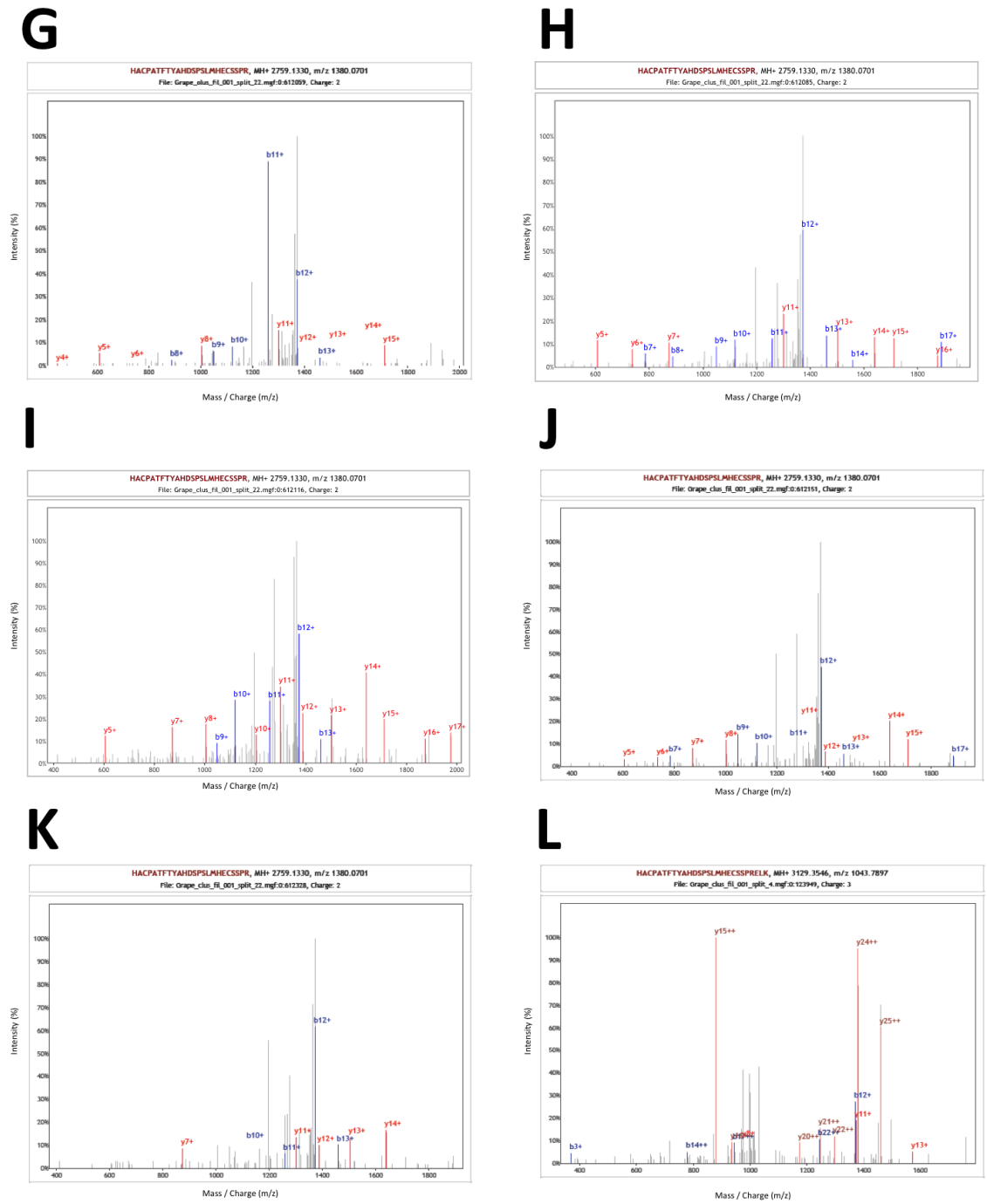
A single MS/MS spectrum supporting the novel peptide annotating a novel splice event, illustrated in Figure 5.7.



Appendix Figure 5.10 Supporting MS/MS spectra for an exon boundary annotation event

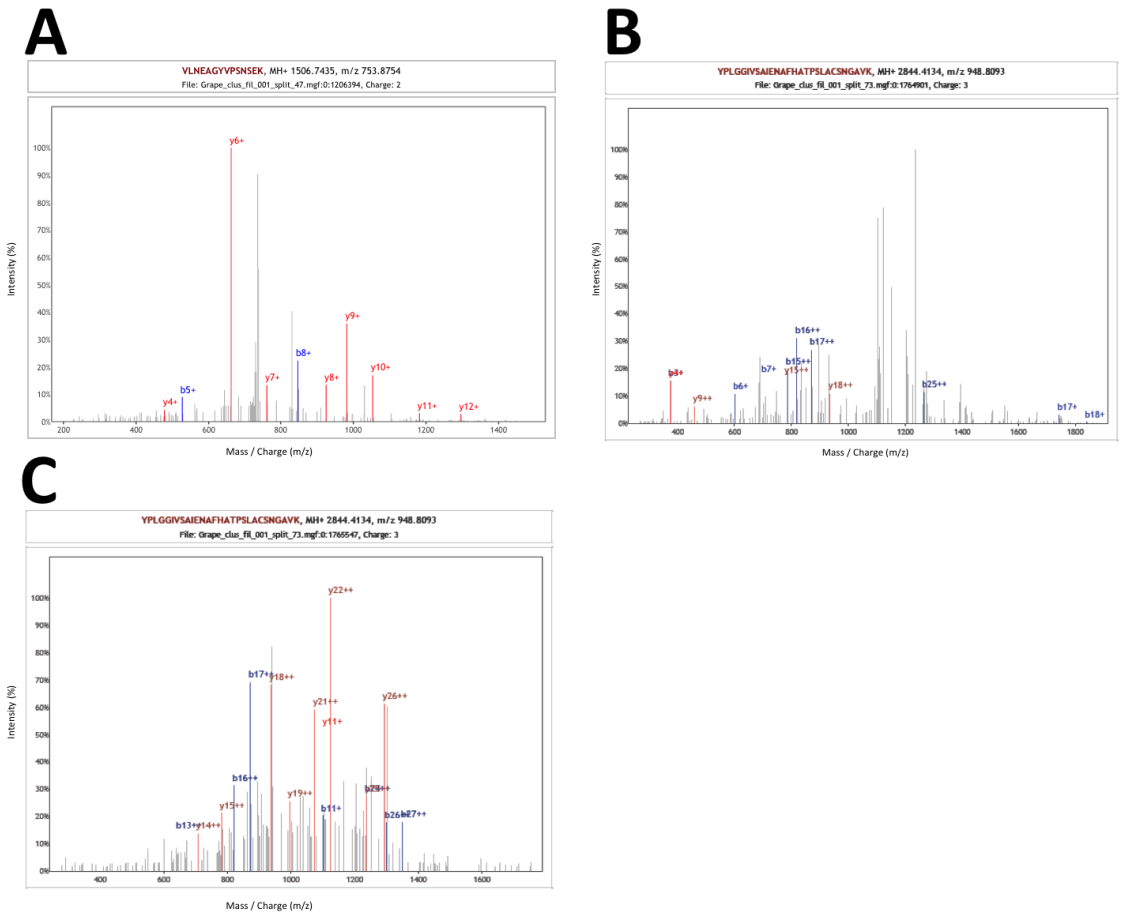
Five representative MS/MS spectra (A-E; from a total of 59) supporting novel peptides annotating an exon boundary event, illustrated in Figure 5.6.


A
HACPATFTYAHDSPSLMHECSSPR, MH+ 2759.1330, m/z 920.3825
File: Grape_clus_fil_001_split_70.mgf:0:1691542, Charge: 3
B
HACPATFTYAHDSPSLMHECSSPR, MH+ 2759.1330, m/z 920.3825
File: Grape_clus_fil_001_split_70.mgf:0:1693756, Charge: 3
C
HACPATFTYAHDSPSLMHECSSPR, MH+ 2759.1330, m/z 920.3825
File: Grape_clus_fil_001_split_70.mgf:0:1693788, Charge: 3
D
HACPATFTYAHDSPSLMHECSSPR, MH+ 2759.1330, m/z 920.3825
File: Grape_clus_fil_001_split_70.mgf:0:1693897, Charge: 3
E
HACPATFTYAHDSPSLMHECSSPR, MH+ 2759.1330, m/z 920.3825
File: Grape_clus_fil_001_split_70.mgf:0:1694098, Charge: 3
F
HACPATFTYAHDSPSLMHECSSPR, MH+ 2775.1330, m/z 925.7158
File: Grape_clus_fil_001_split_71.mgf:0:1703730, Charge: 3
Intensity (%)
Mass / Charge (m/z)



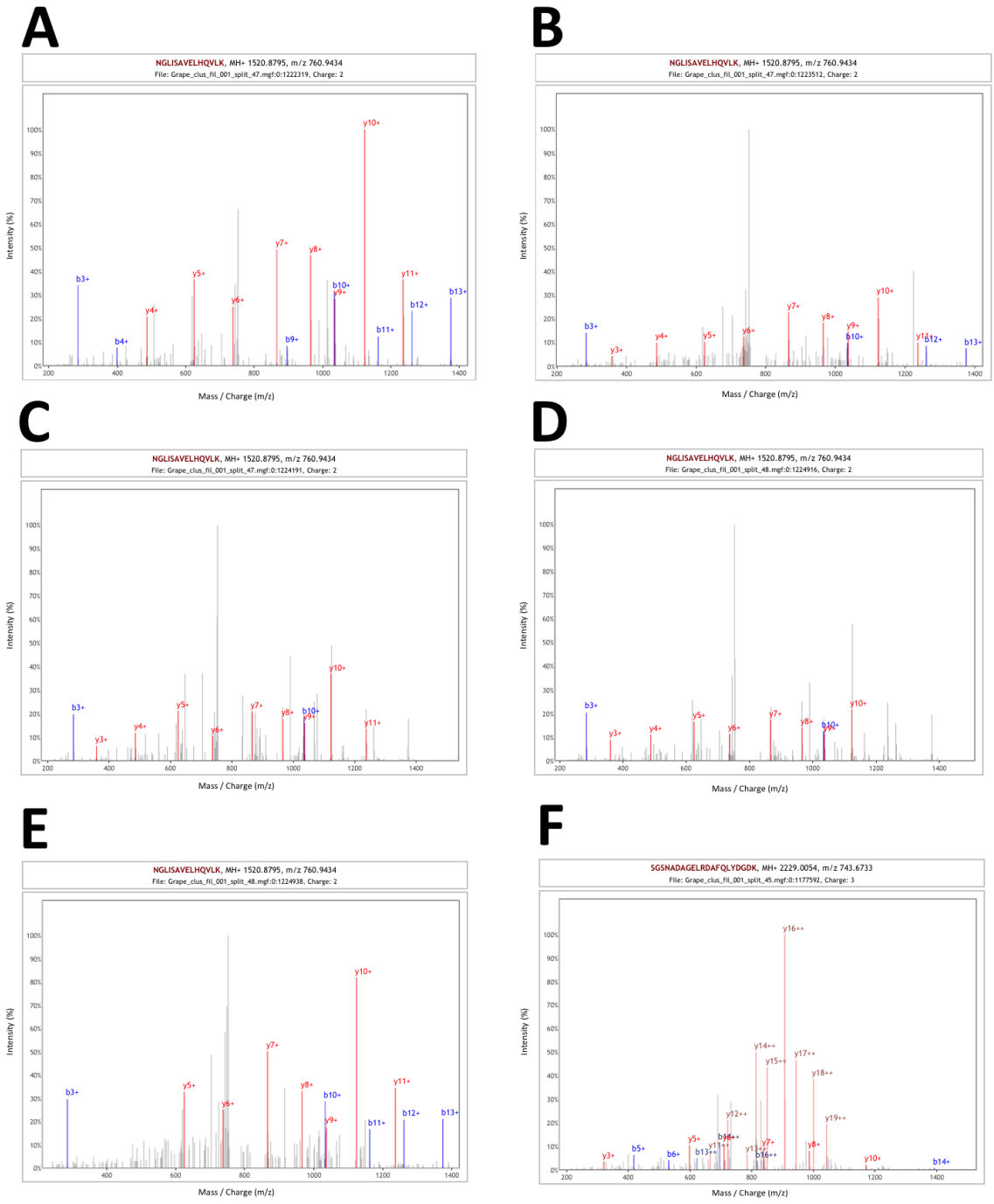
Appendix Figure 5.11 Supporting MS/MS spectra for an exon boundary annotation event

Twelve MS/MS spectra (A-L) supporting novel peptides annotating an exon boundary event, illustrated in Figure 5.8.

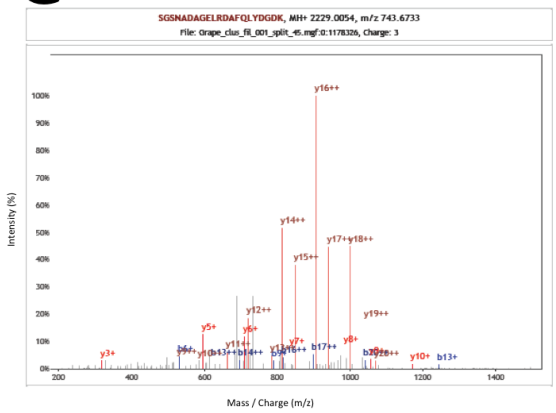


Appendix Figure 5.12 Supporting MS/MS spectra for a frame-shift annotation event

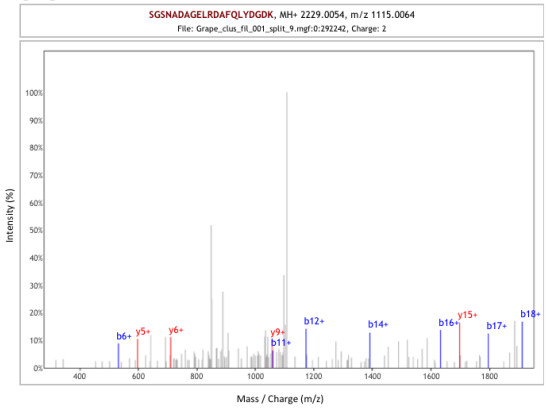
Three MS/MS spectra (A-C) supporting novel peptides annotating a frame-shift event, illustrated in Figure 5.9.


A
NGLISAVELHQVLK, MH+ 1520.8795, m/z 760.9434
File: Grape_clus_fil_001_split_47.mgf:0:1222319, Charge: 2
B
NGLISAVELHQVLK, MH+ 1520.8795, m/z 760.9434
File: Grape_clus_fil_001_split_47.mgf:0:1223512, Charge: 2
C
NGLISAVELHQVLK, MH+ 1520.8795, m/z 760.9434
File: Grape_clus_fil_001_split_47.mgf:0:1224191, Charge: 2
D
NGLISAVELHQVLK, MH+ 1520.8795, m/z 760.9434
File: Grape_clus_fil_001_split_48.mgf:0:1224916, Charge: 2
E
NGLISAVELHQVLK, MH+ 1520.8795, m/z 760.9434
File: Grape_clus_fil_001_split_48.mgf:0:1224938, Charge: 2
F
SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 743.6733
File: Grape_clus_fil_001_split_45.mgf:0:1177592, Charge: 3
Intensity (%)
Mass / Charge (m/z)

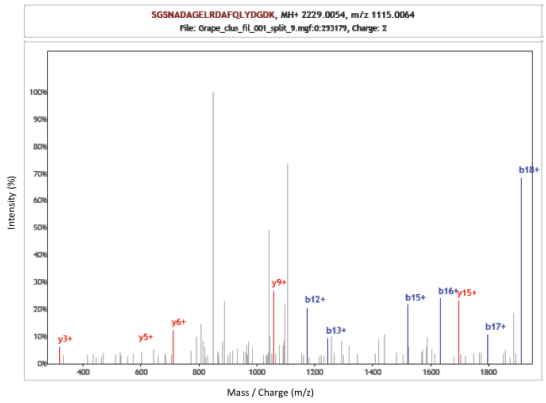
G


SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 743.6733
File: Grape_clus_fil_001_split_45.mgf:0:1178326, Charge: 3
Intensity (%)
Mass / Charge (m/z)

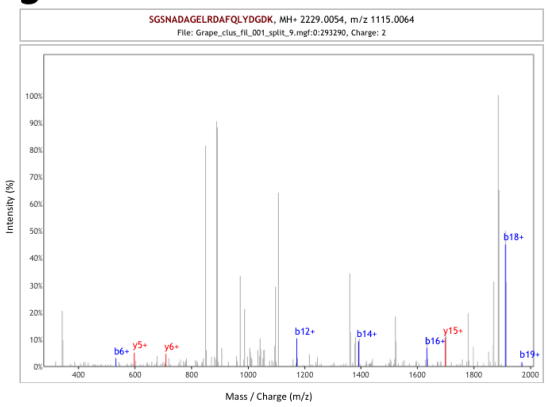
H


SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 1115.0064
File: Grape_clus_fil_001_split_9.mgf:0:292242, Charge: 2
Intensity (%)
Mass / Charge (m/z)

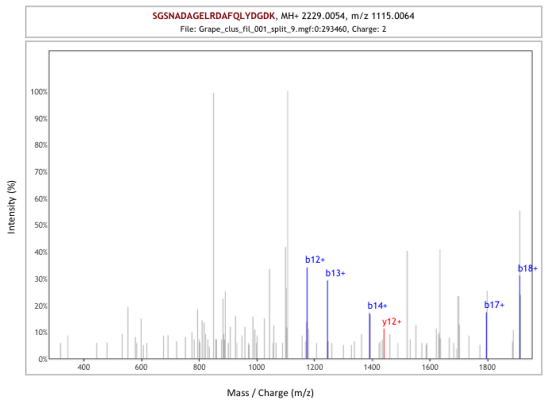
I


SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 1115.0064
File: Grape_clus_fil_001_split_9.mgf:0:293179, Charge: 2
Intensity (%)
Mass / Charge (m/z)

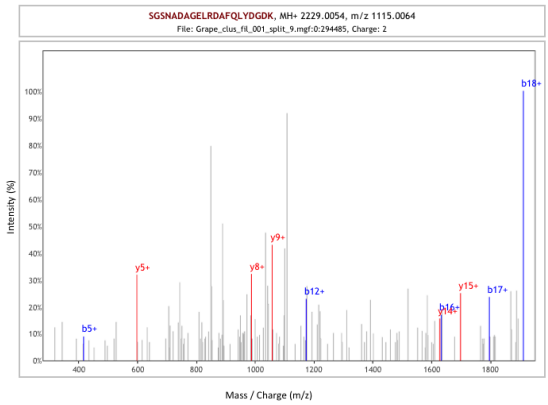
J

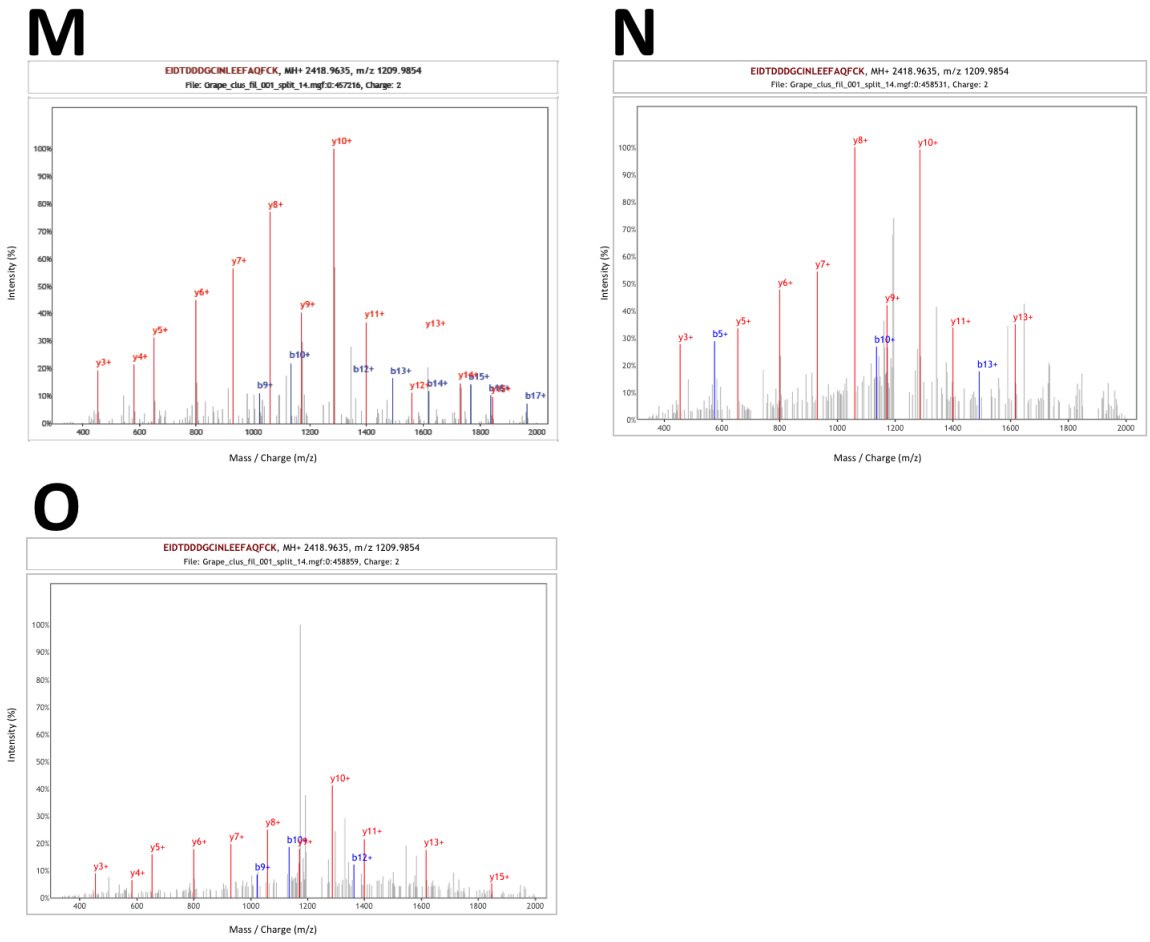

SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 1115.0064
File: Grape_clus_fil_001_split_9.mgf:0:293290, Charge: 2
Intensity (%)
Mass / Charge (m/z)

K


SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 1115.0064
File: Grape_clus_fil_001_split_9.mgf:0:293460, Charge: 2
Intensity (%)
Mass / Charge (m/z)

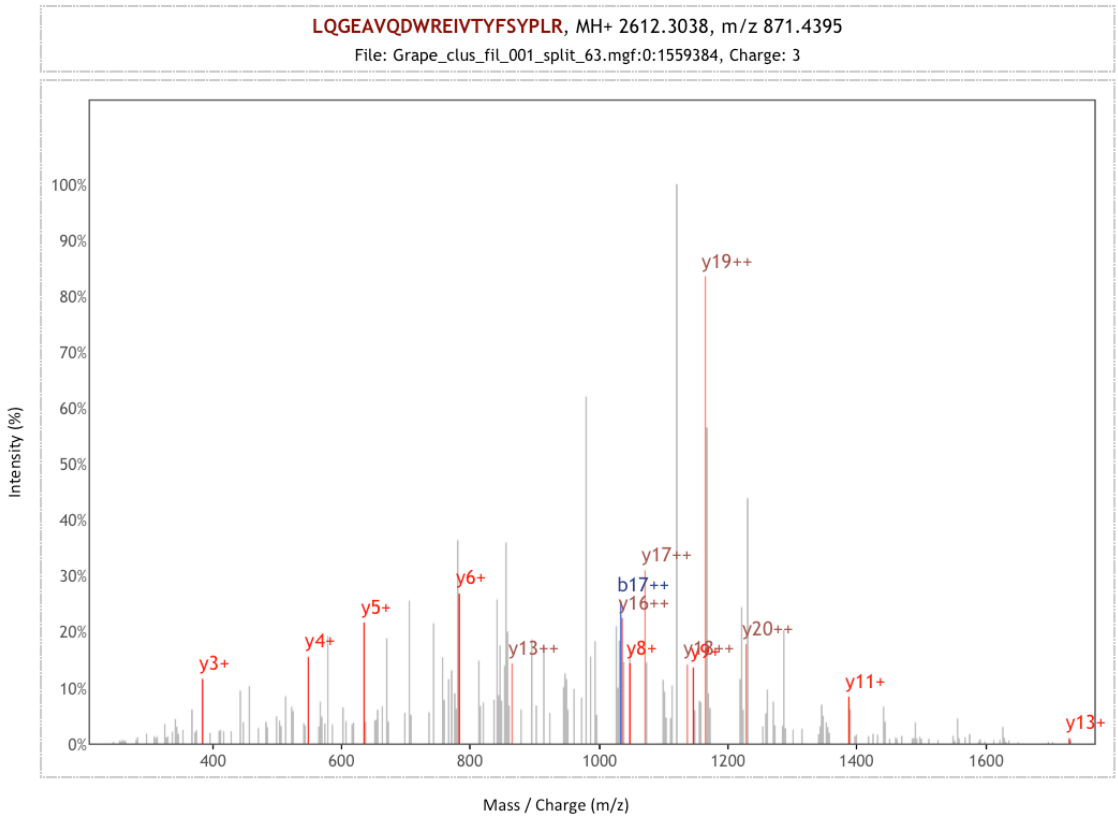
L


SGSNADAGELRDAFQLYDGDK, MH+ 2229.0054, m/z 1115.0064
File: Grape_clus_fil_001_split_9.mgf:0:294485, Charge: 2
Intensity (%)
Mass / Charge (m/z)



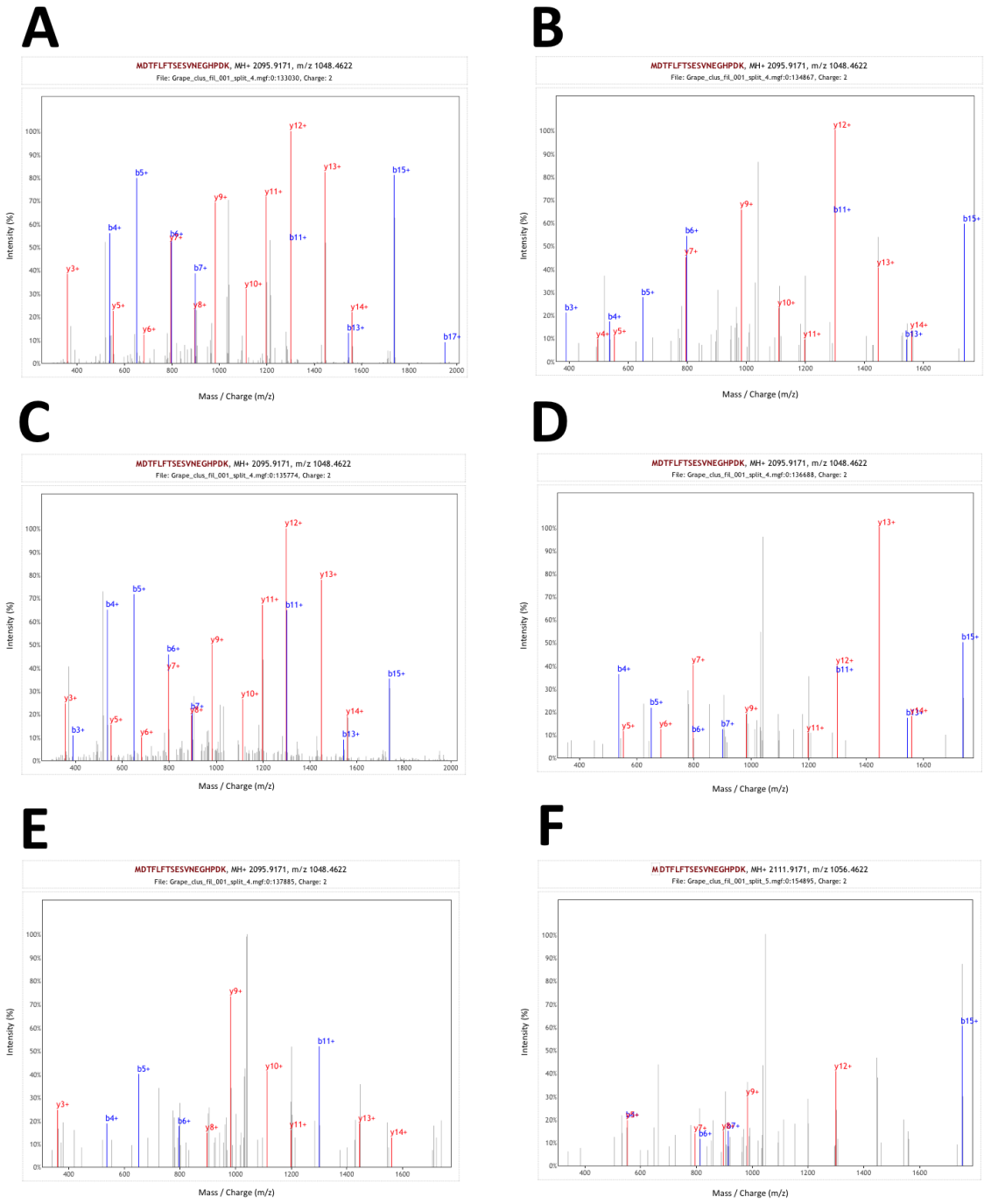
Appendix Figure 5.13 Supporting MS/MS spectra for a novel exon annotation event

Fifteen MS/MS spectra (A-O) supporting novel peptides annotating a novel exon event, illustrated in Figure 5.10.

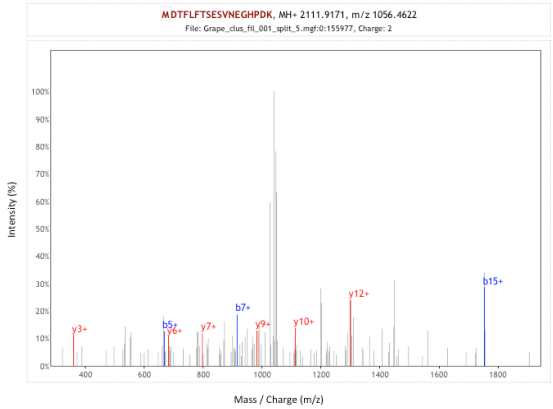


Appendix Figure 5.14 Supporting MS/MS spectrum for a N-terminal acetylated peptide suggesting a conflict with the reference annotation

A single MS/MS spectrum supporting an N-terminal acetylated peptide in a known reference protein, illustrated in Figure 5.11.

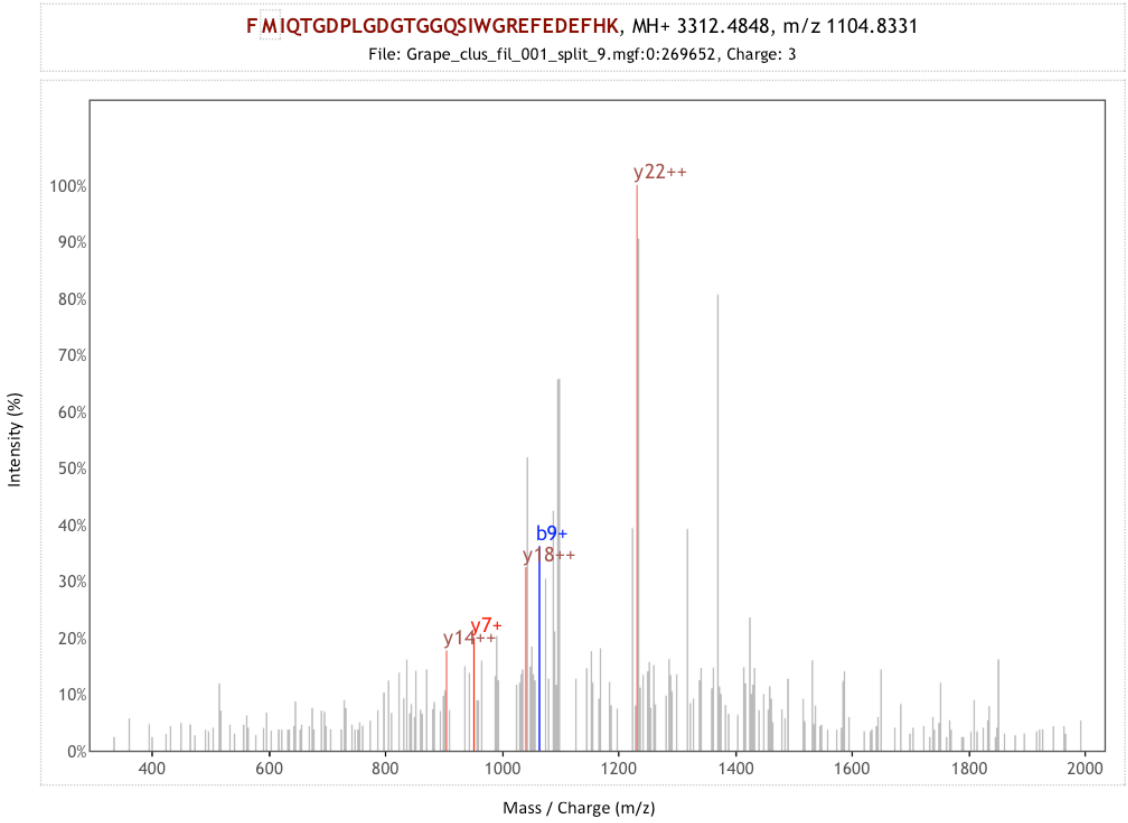

A
MDTFLFTSESVNEGHPDK, MH+ 2095.9171, m/z 1048.4622
File: Grape_clus_fil_001_split_4.mgf:0:133030, Charge: 2
B
MDTFLFTSESVNEGHPDK, MH+ 2095.9171, m/z 1048.4622
File: Grape_clus_fil_001_split_4.mgf:0:134867, Charge: 2
C
MDTFLFTSESVNEGHPDK, MH+ 2095.9171, m/z 1048.4622
File: Grape_clus_fil_001_split_4.mgf:0:135774, Charge: 2
D
MDTFLFTSESVNEGHPDK, MH+ 2095.9171, m/z 1048.4622
File: Grape_clus_fil_001_split_4.mgf:0:136688, Charge: 2
E
MDTFLFTSESVNEGHPDK, MH+ 2095.9171, m/z 1048.4622
File: Grape_clus_fil_001_split_4.mgf:0:137885, Charge: 2
F
MDTFLFTSESVNEGHPDK, MH+ 2111.9171, m/z 1056.4622
File: Grape_clus_fil_001_split_5.mgf:0:154895, Charge: 2
Intensity (%)
Mass / Charge (m/z)

G



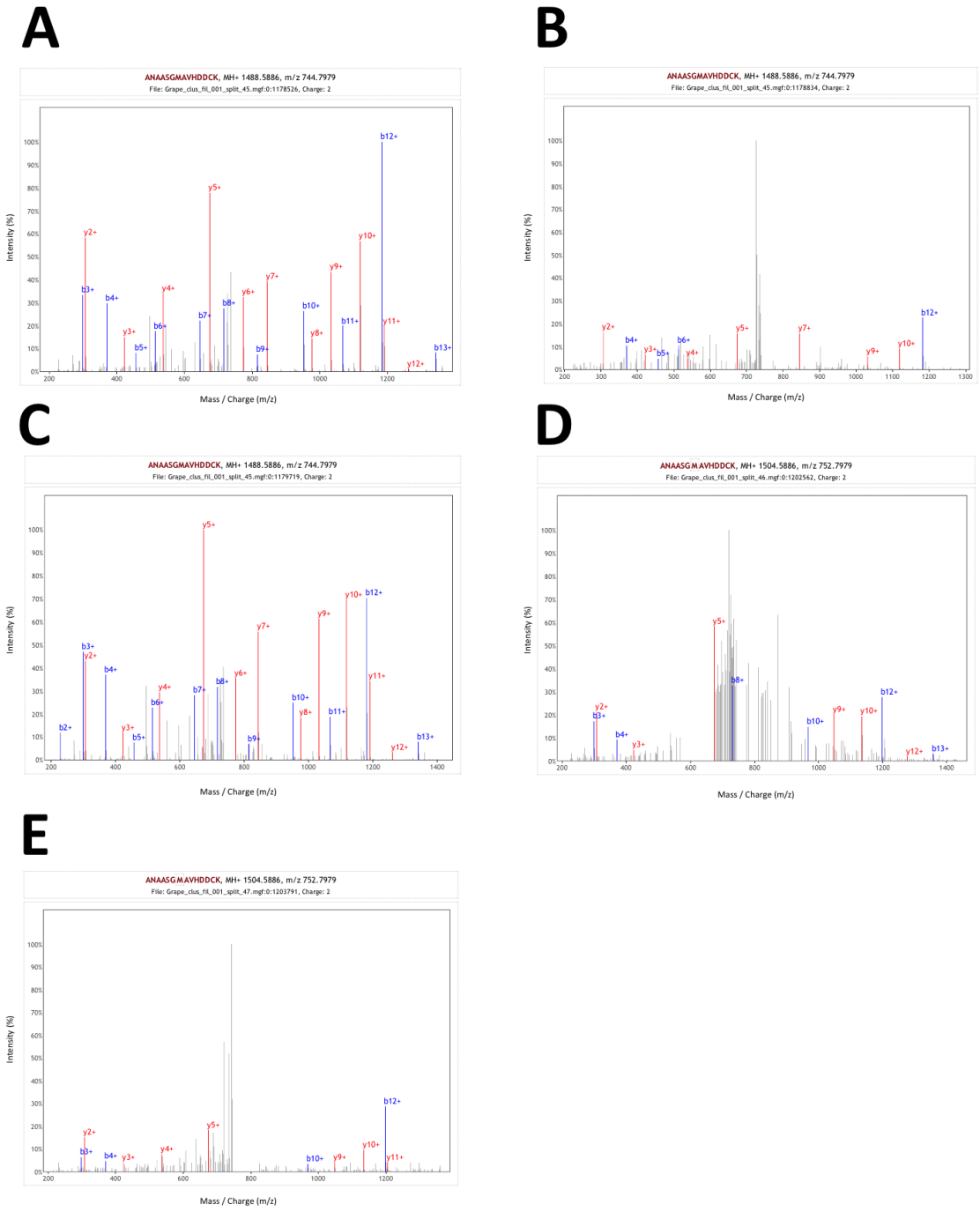
Appendix Figure 5.15 Supporting MS/MS spectra for an N-terminal acetylated peptide suggesting a conflict with the reference annotation

Seven MS/MS spectra (A-G) supporting an N-terminal acetylated peptide in a known reference protein, illustrated in Figure 5.12.



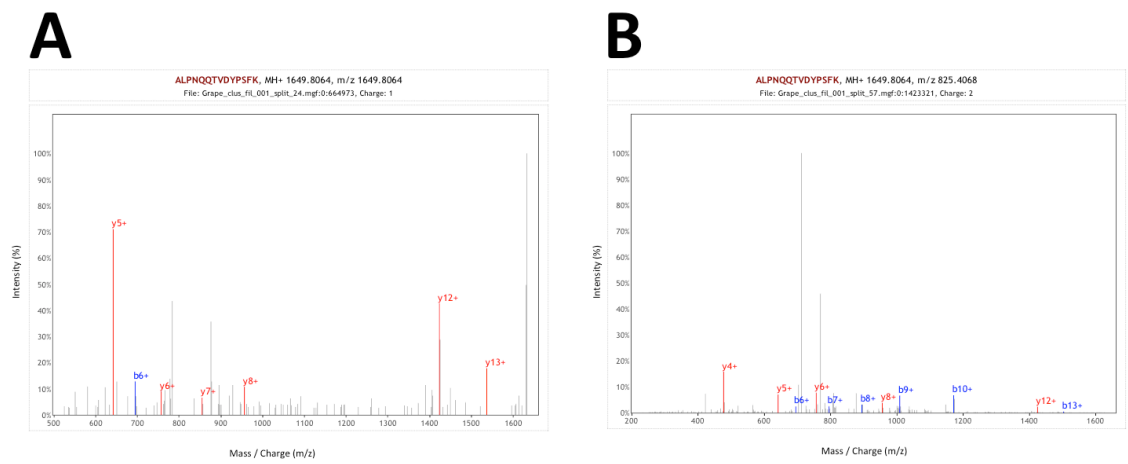
Appendix Figure 5.16 Supporting MS/MS spectrum for an N-terminal acetylated peptide suggesting a conflict with the reference annotation

A single MS/MS spectrum supporting an N-terminal acetylated peptide in a known reference protein, illustrated in Figure 5.13.



Appendix Figure 5.17 Supporting MS/MS spectra for an N-terminal acetylated peptide suggesting a conflict with the reference annotation

Five representative MS/MS spectra (A-E; from a total of 23), supporting a N-terminal acetylated peptide in a known reference protein, illustrated in Figure 5.14.



Appendix Figure 5.18 Supporting MS/MS spectra for an N-terminal acetylated peptide suggesting a conflict with the reference annotation

Two MS/MS spectra (A-B) supporting an N-terminal acetylated peptide in a known reference protein, illustrated in Figure 5.15.

Appendix File 6.1 Reference predictions are in zip file 'AppendixFile6.1.zip' on the DVD provided.

Appendix File 6.2 Clustering, quality filtering and precursor mass tolerance optimization results are in excel file 'AppendixFile6.2.xlsx' on the DVD provided.

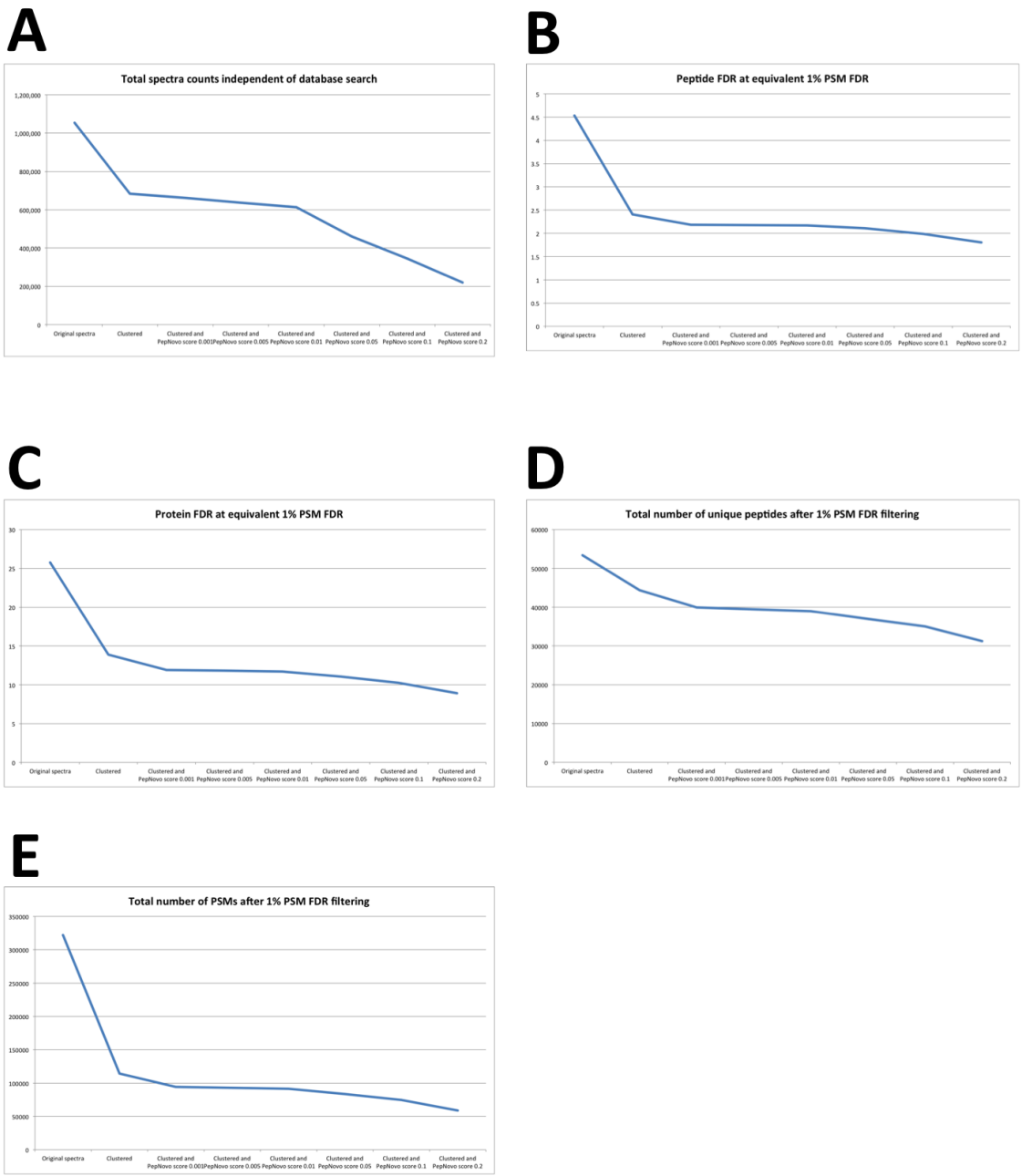
Appendix File 6.3 Processed proteogenomics results are in excel file 'AppendixFile6.3.xlsx' on the DVD provided.

Appendix File 6.4 Raw proteogenomics results are in zip file 'AppendixFile6.4.zip' on the DVD provided.

Appendix File 6.5 Augustus gene predictions are in zip file 'AppendixFile6.5.zip' on the DVD provided.

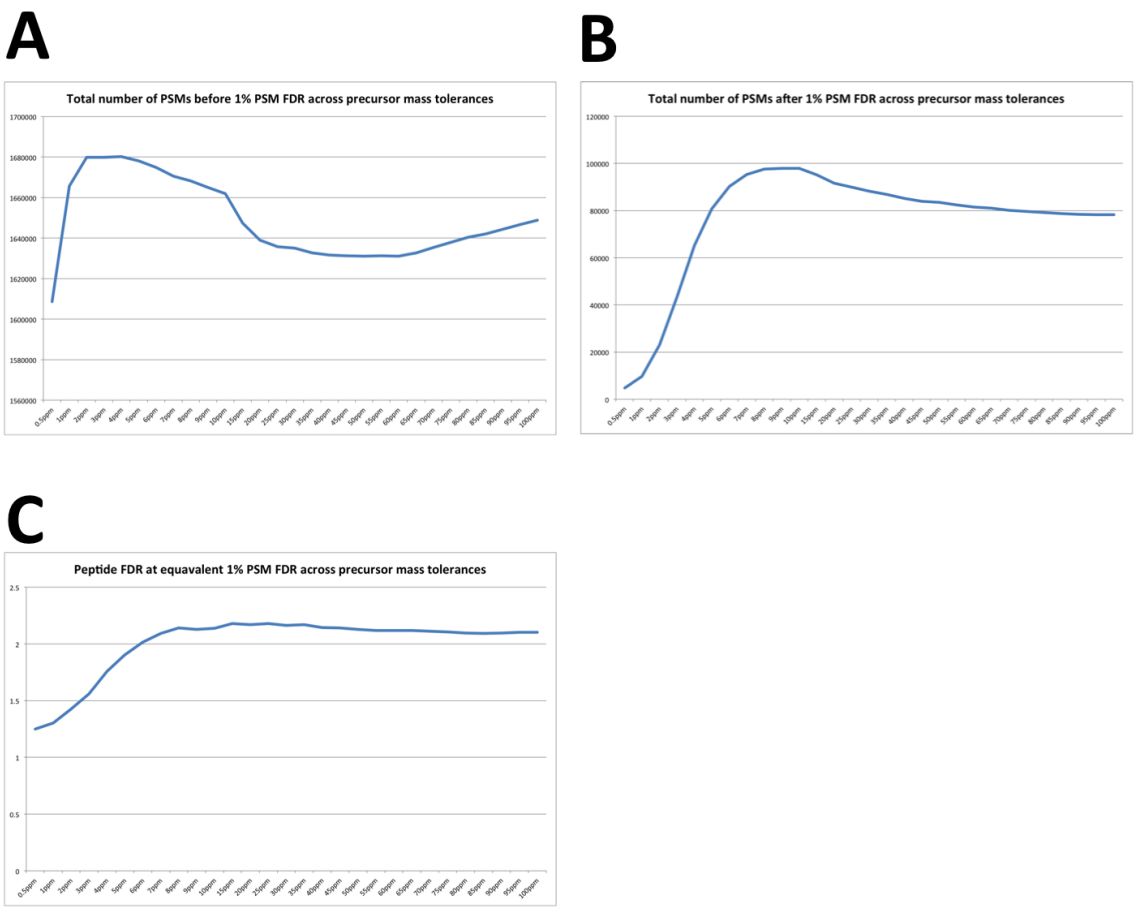
Appendix File 6.6 Augustus gene predictions with incorporated novel peptides are in excel file 'AppendixFile6.6.xlsx' on the DVD provided.

Appendix File 6.7 Novel N-terminal acetylated peptides are in excel file 'AppendixFile6.7.xlsx' on the DVD.



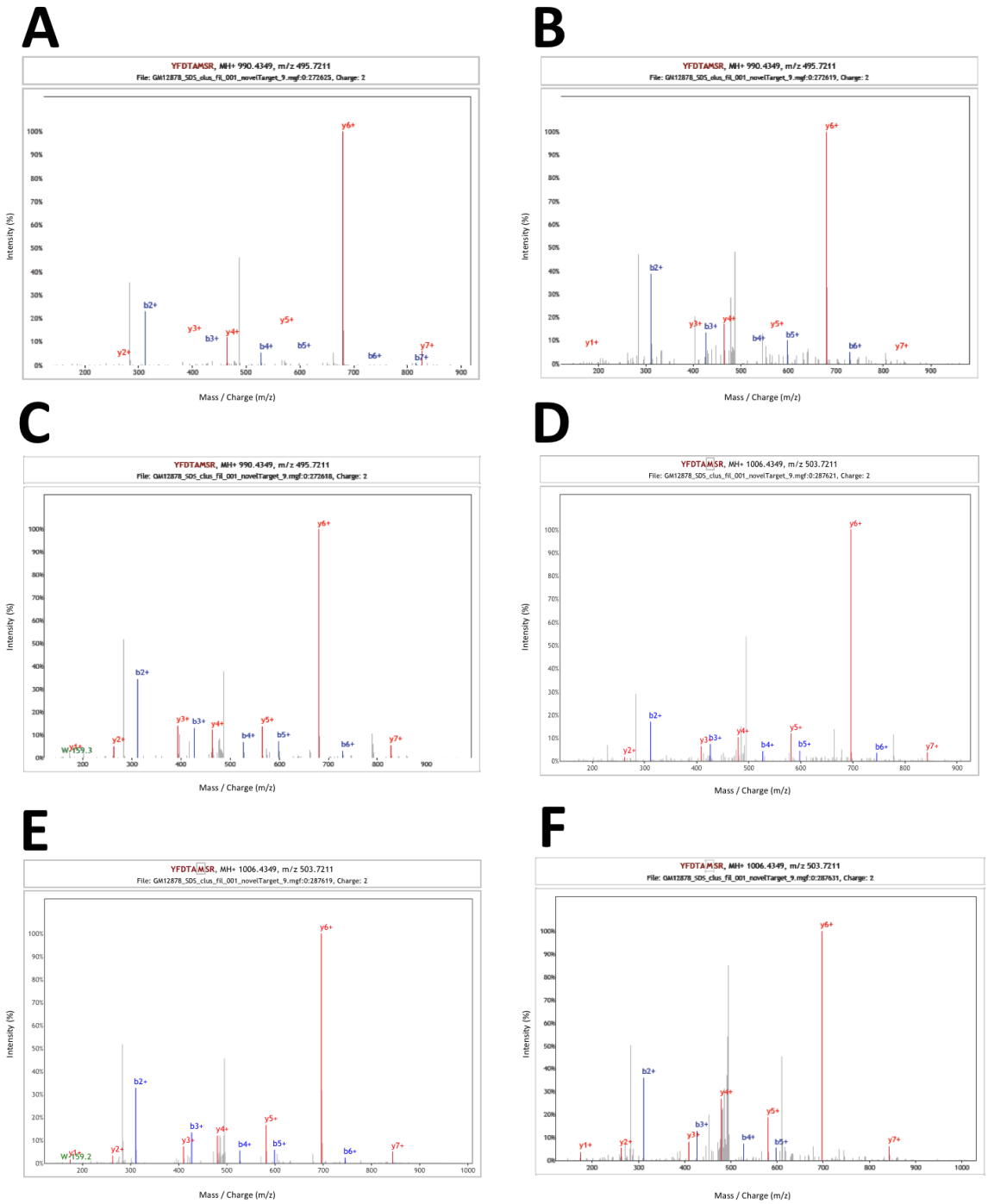
Appendix Figure 6.1 Pre and post clustering with and without PepNovo quality filtering

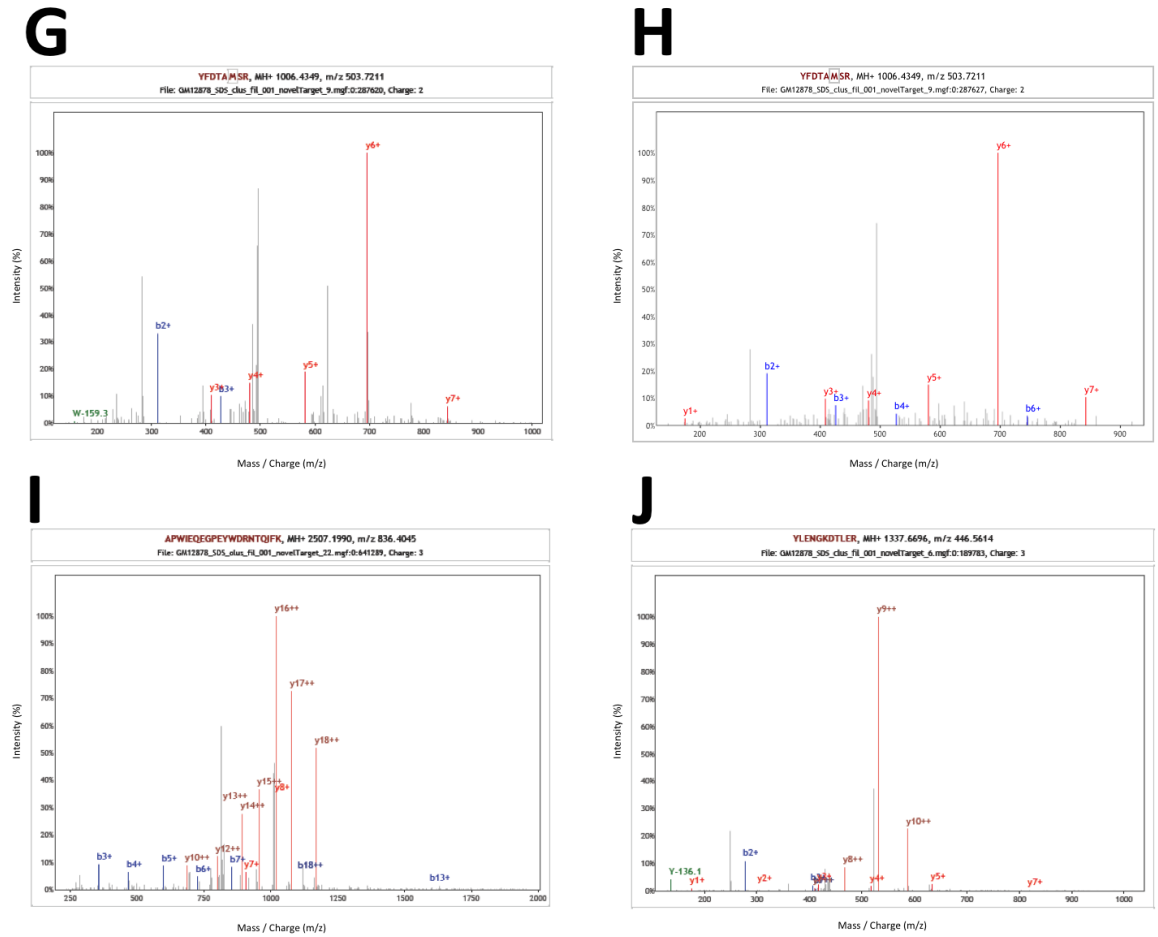
(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



Appendix Figure 6.2 Precursor mass tolerance optimisation

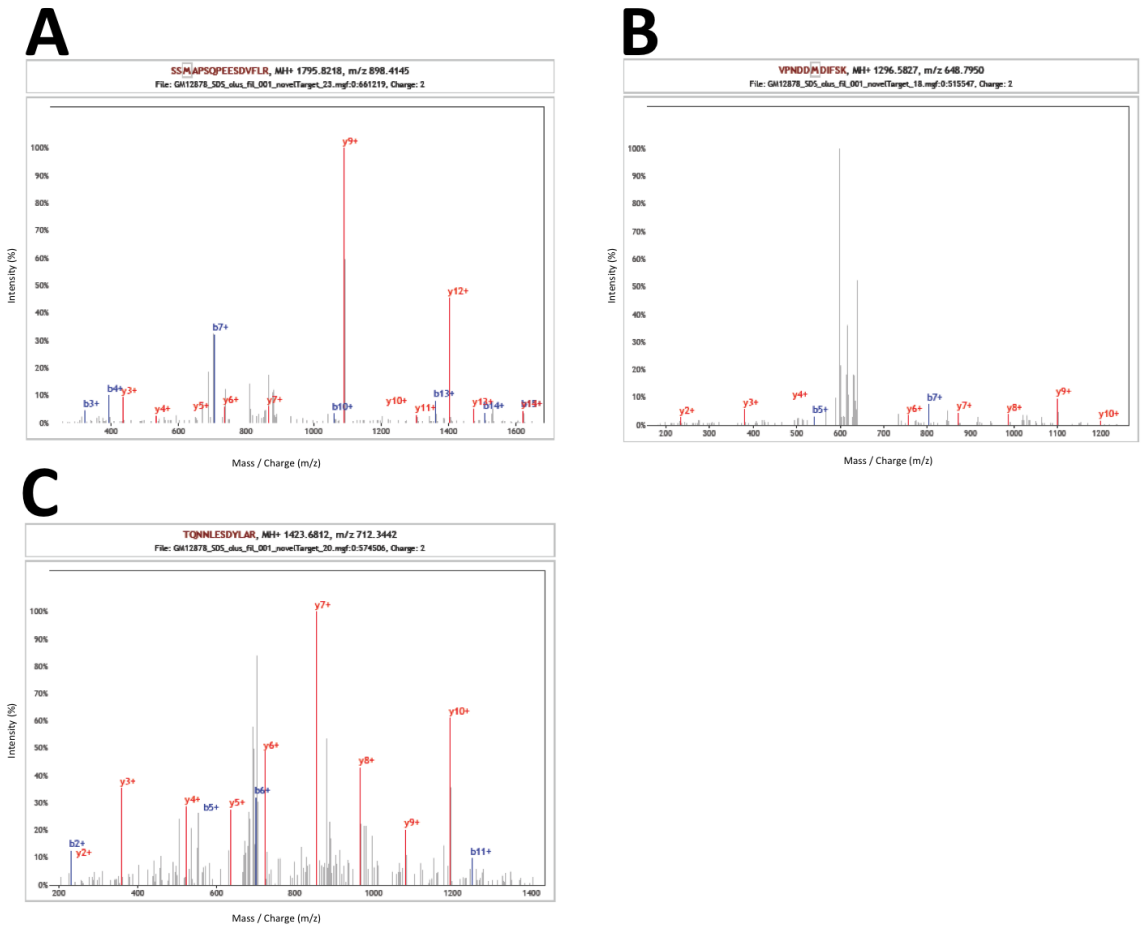
(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.


A
YFDTAMSR, MH+ 990.4349, m/z 495.7211
File: GM12878_SDS_clus_fil_001_novelTarget_9.mgf:0:272625, Charge: 2
B
YFDTAMSR, MH+ 990.4349, m/z 495.7211
File: GM12878_SDS_clus_fil_001_novelTarget_9.mgf:0:272619, Charge: 2
C
YFDTAMSR, MH+ 990.4349, m/z 495.7211
File: GM12878_SDS_clus_fil_001_novelTarget_9.mgf:0:272618, Charge: 2
D
YFDTAMSR, MH+ 1006.4349, m/z 503.7211
File: GM12878_SDS_clus_fil_001_novelTarget_9.mgf:0:287621, Charge: 2
E
YFDTAMSR, MH+ 1006.4349, m/z 503.7211
File: GM12878_SDS_clus_fil_001_novelTarget_9.mgf:0:287619, Charge: 2
F
YFDTAMSR, MH+ 1006.4349, m/z 503.7211
File: GM12878_SDS_clus_fil_001_novelTarget_9.mgf:0:287631, Charge: 2
Intensity (%)
Mass / Charge (m/z)



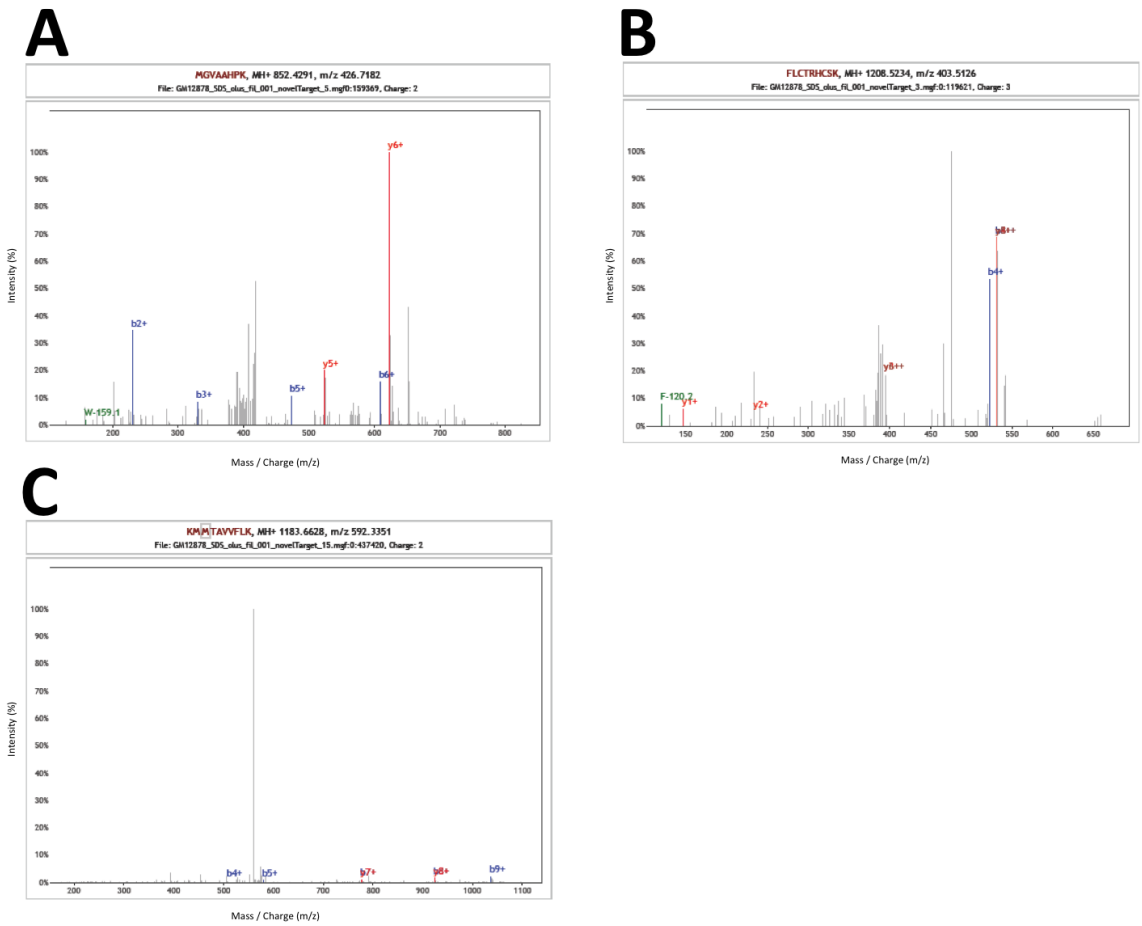
Appendix Figure 6.3 Supporting MS/MS spectra for a novel gene event

Ten MS/MS spectra (A-J) supporting novel peptides annotating a novel gene event, illustrated in Figure 6.2.



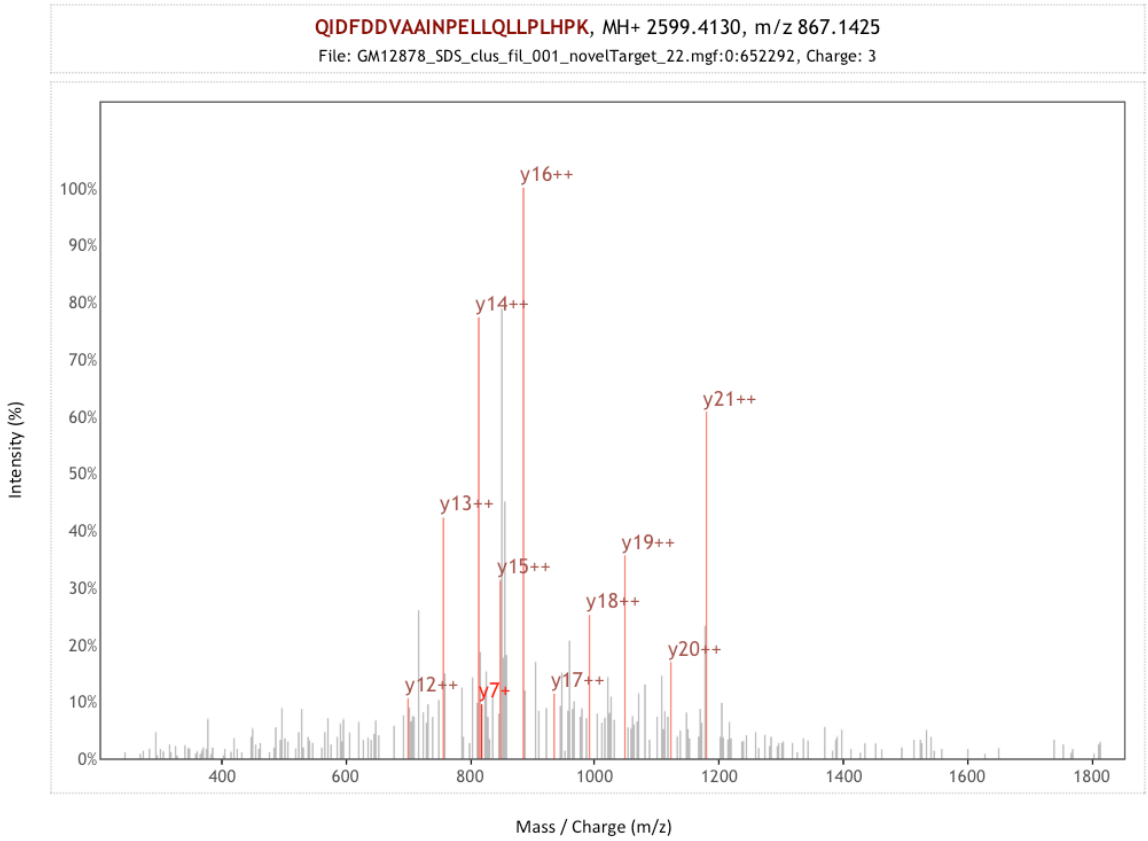
Appendix Figure 6.4 Supporting MS/MS spectra for a gene boundary and novel exon annotation events

Three MS/MS spectra (A-C) supporting novel peptides annotating a gene boundary and novel exon events, illustrated in Figure 6.3.



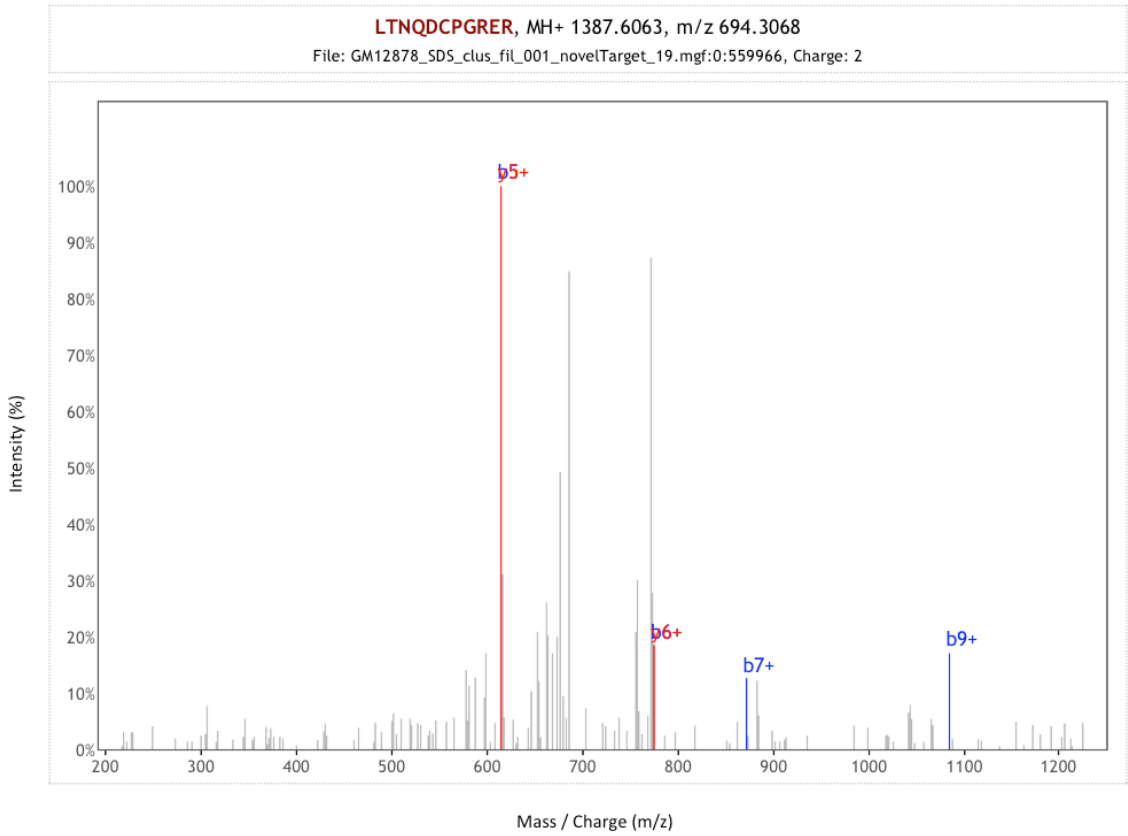
Appendix Figure 6.5 Supporting MS/MS spectra for a reverse strand annotation event and erroneously included frame-shift annotation event

Three MS/MS spectra (A-C) supporting novel peptides annotating a reverse strand event, and a likely erroneous frame-shift event from peptide “KMMTAVVFLK”, with its lower intensity MS/MS spectrum, illustrated in Figure 6.4.



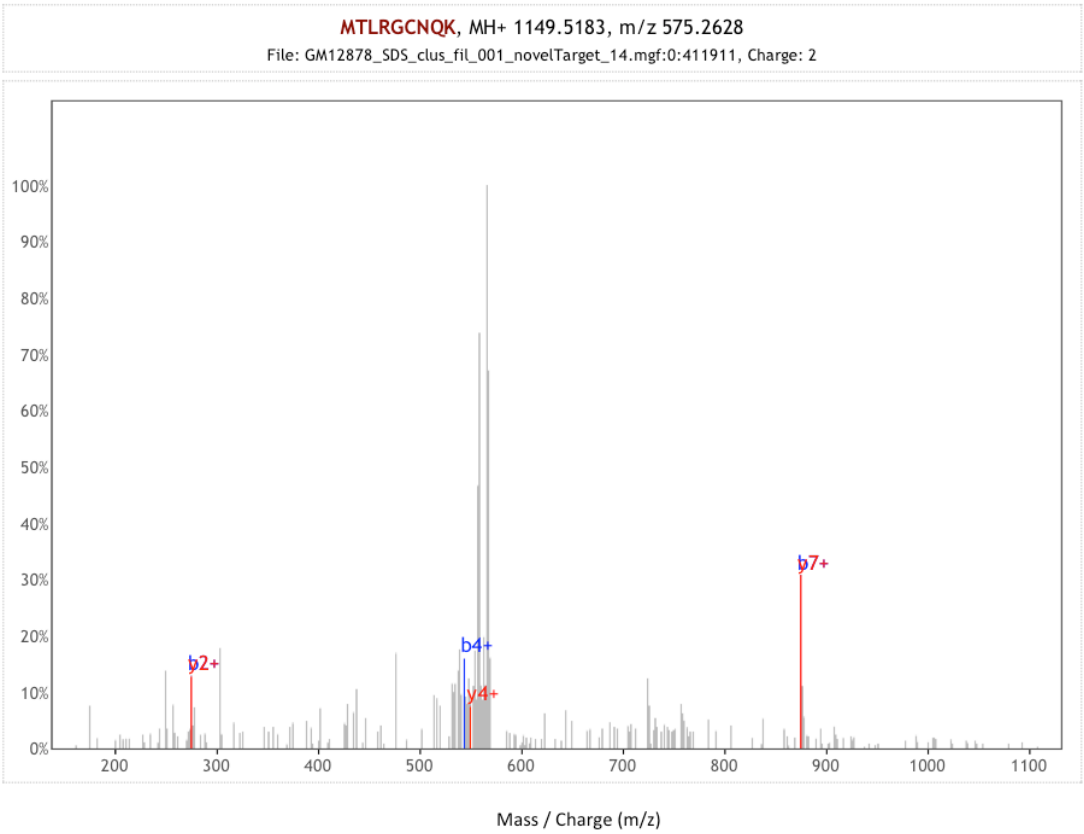
Appendix Figure 6.6 Supporting MS/MS spectrum for an exon boundary and translated UTR annotation event

A single MS/MS spectrum supporting a novel peptide annotating an exon boundary and translated UTR event, illustrated in Figure 6.5.



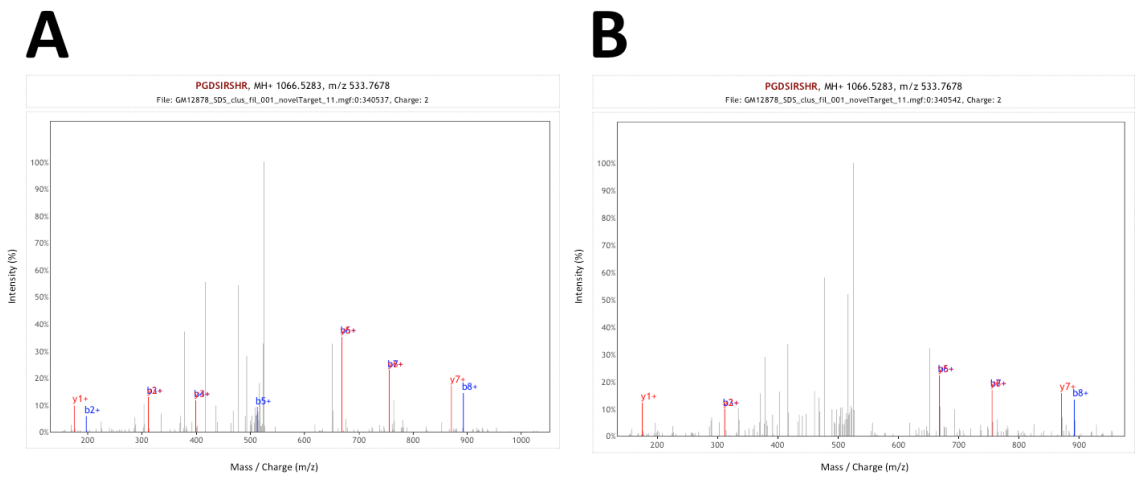
Appendix Figure 6.7 Supporting MS/MS spectrum for a novel and unique N-terminal acetylated peptide

A single MS/MS spectrum supporting a novel and unique N-terminal acetylated peptide, identified in a gene boundary and reverse strand event, and which was not incorporated into any Augustus gene predictions.



Appendix Figure 6.8 Supporting MS/MS spectrum for a novel and unique N-terminal acetylated peptide

A single MS/MS spectrum supporting a novel and unique N-terminal acetylated peptide, identified in a gene boundary and reverse strand event, and which was incorporated into an Augustus gene prediction.



Appendix Figure 6.9 Supporting MS/MS spectra for a novel and shared N-terminal acetylated peptide

Two MS/MS spectra supporting a novel and shared N-terminal acetylated peptide, identified in a gene boundary and reverse strand event, which was not incorporated into the Augustus gene prediction at that location. However, the peptide at one of 76 other locations was incorporated into a prediction at that location.

Appendix File 7.1 Reference predictions are in zip file 'AppendixFile7.1.zip' on the DVD provided.

Appendix File 7.2 Clustering, quality filtering and precursor mass tolerance optimization results are in excel file 'AppendixFile7.2.xlsx' on the DVD provided.

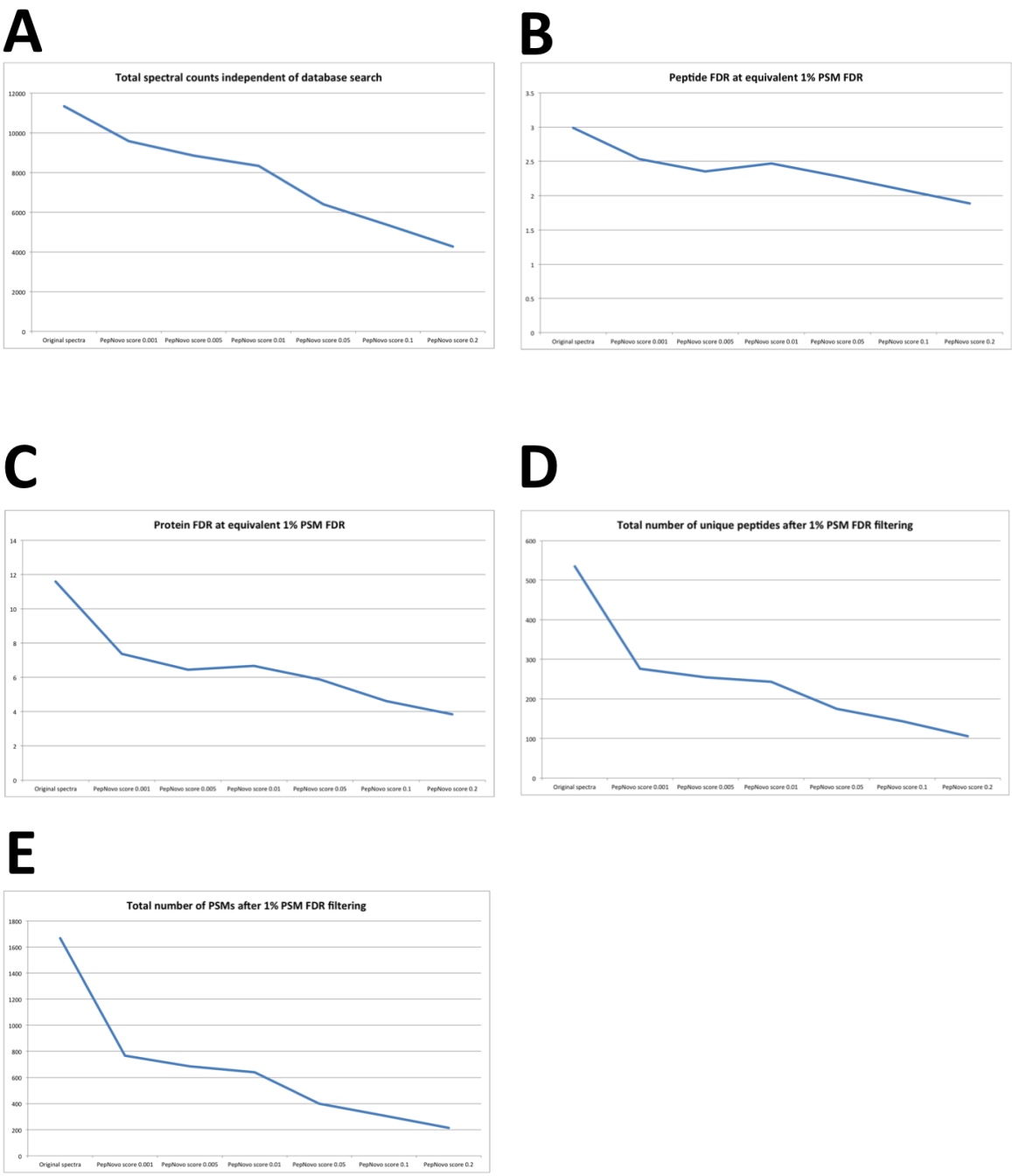
Appendix File 7.3 Processed proteogenomics results are in excel file 'AppendixFile7.3.xlsx' on the DVD provided.

Appendix File 7.4 Raw proteogenomics results are in zip file 'AppendixFile7.4.zip' on the DVD provided.

Appendix File 7.5 Augustus gene predictions are in zip file 'AppendixFile7.5.zip' on the DVD provided.

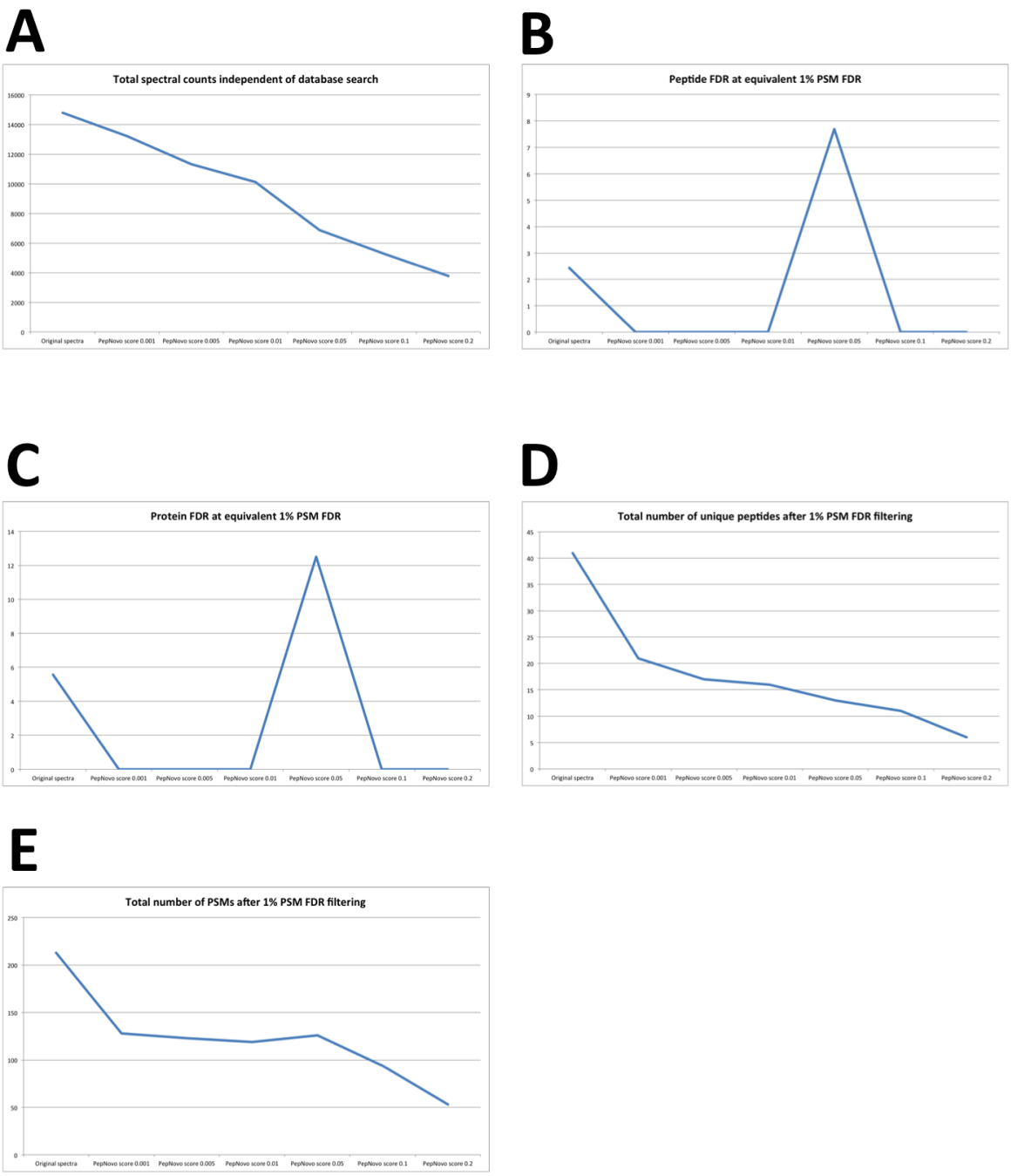
Appendix File 7.6 Augustus gene predictions with incorporated novel peptides are in excel file 'AppendixFile7.6.xlsx' on the DVD provided.

Appendix File 7.7 Known and novel N-terminal acetylated peptides are in excel file 'AppendixFile7.7.xlsx' on the DVD.



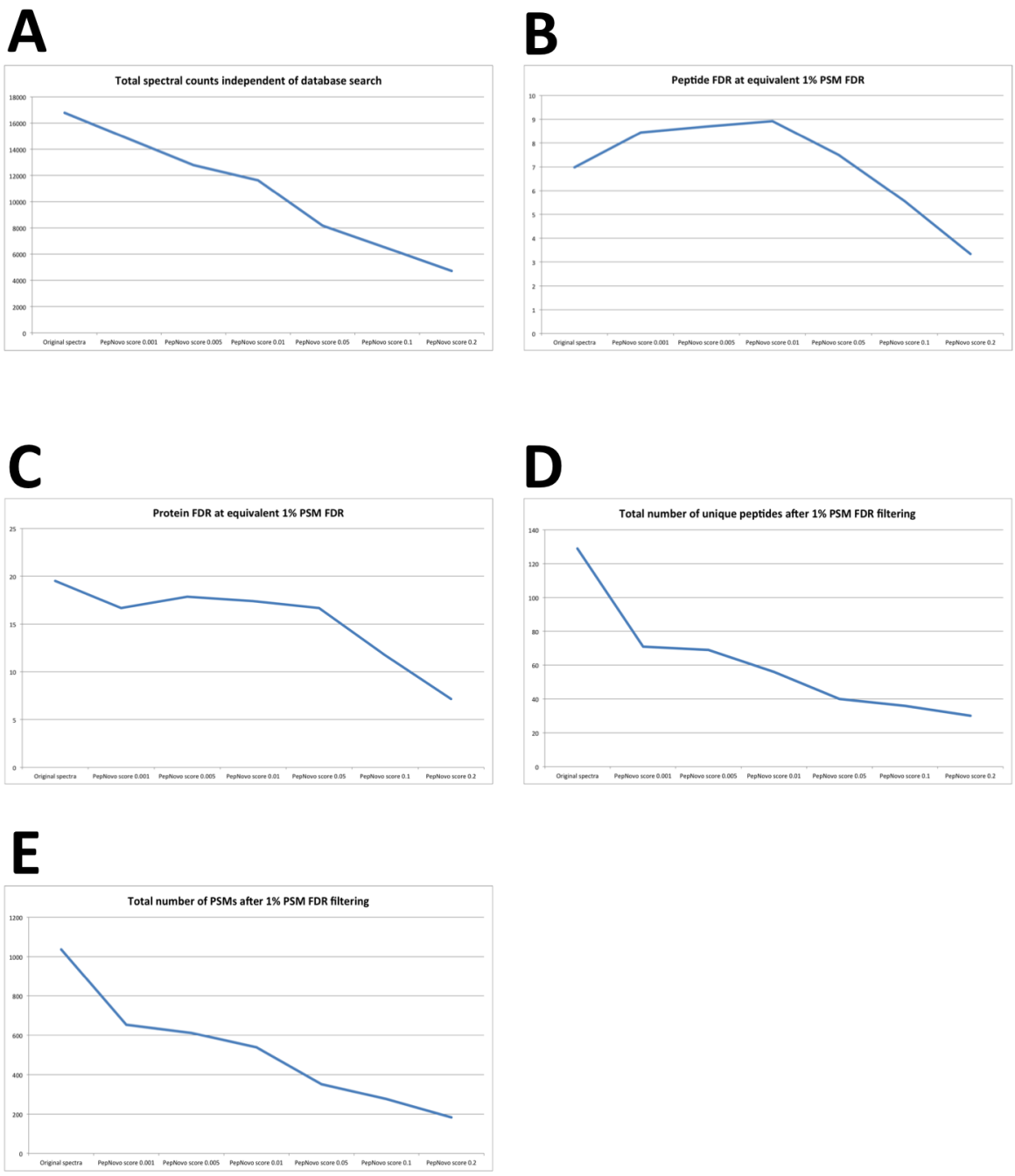
Appendix Figure 7.1 Pre and post PepNovo quality filtering for trypsin-digested wheat flour

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



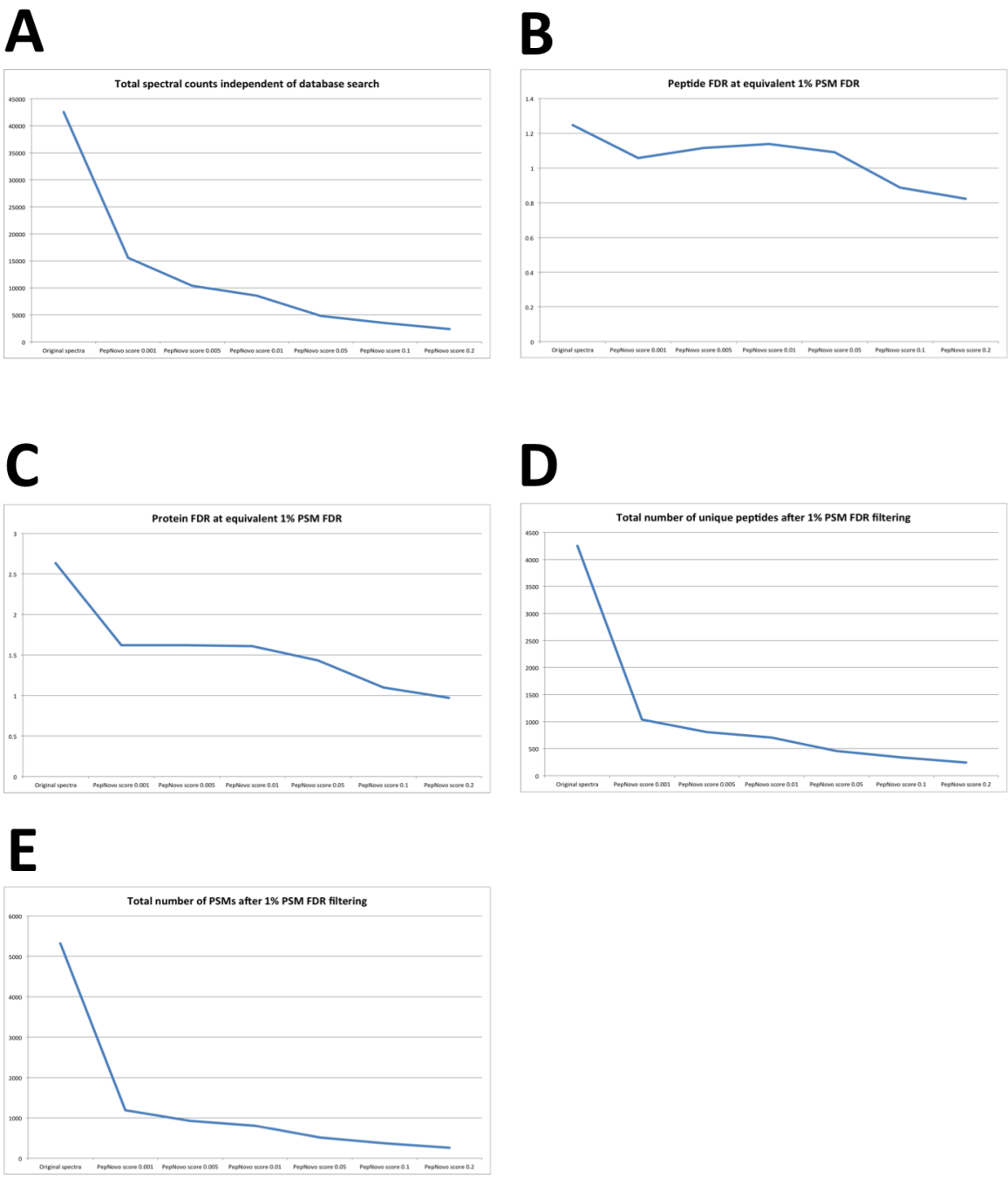
Appendix Figure 7.2 Pre and post PepNovo quality filtering for chymotrypsin-digested wheat flour

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



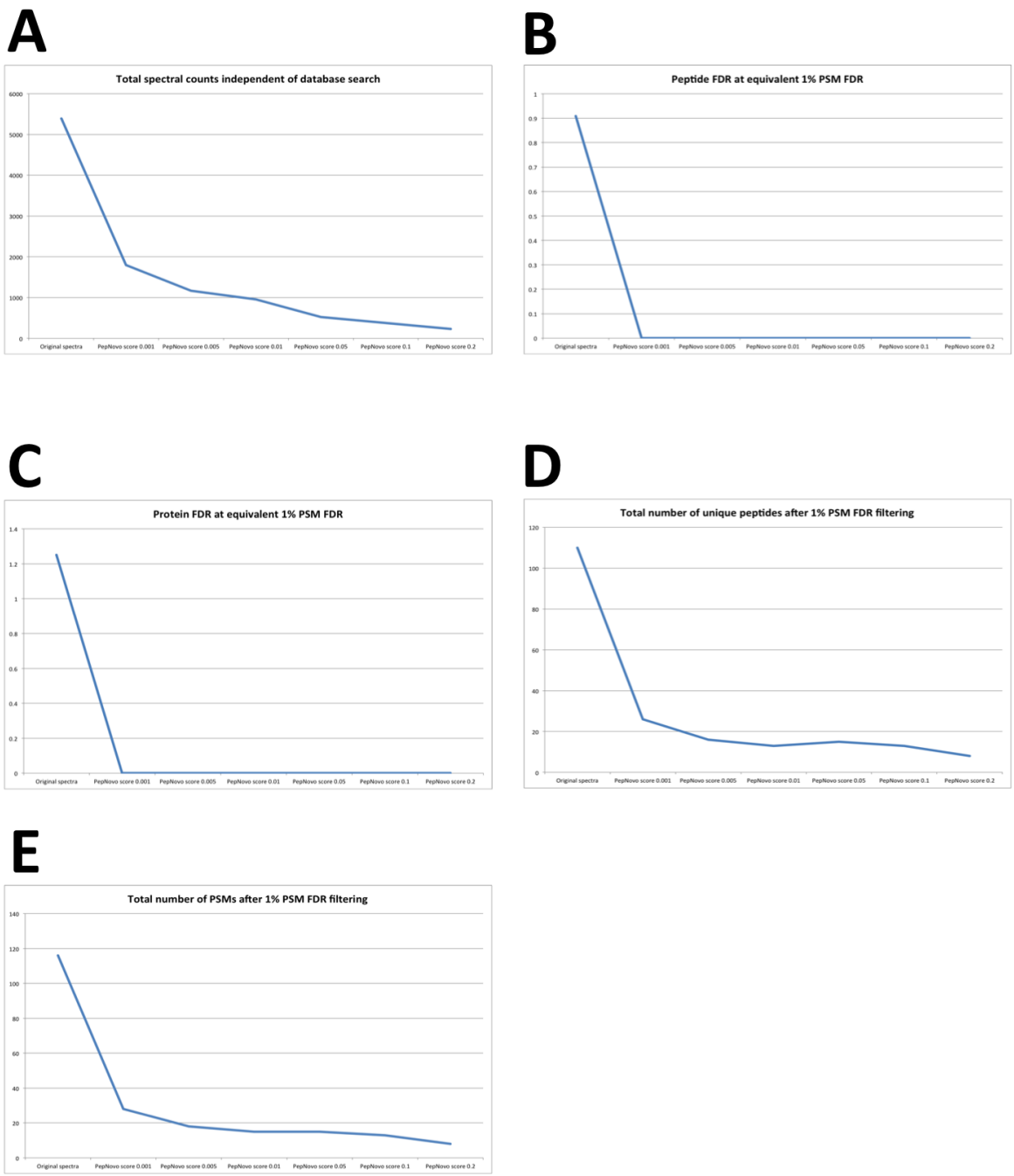
Appendix Figure 7.3 Pre and post PepNovo quality filtering for thermolysin-digested wheat flour

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



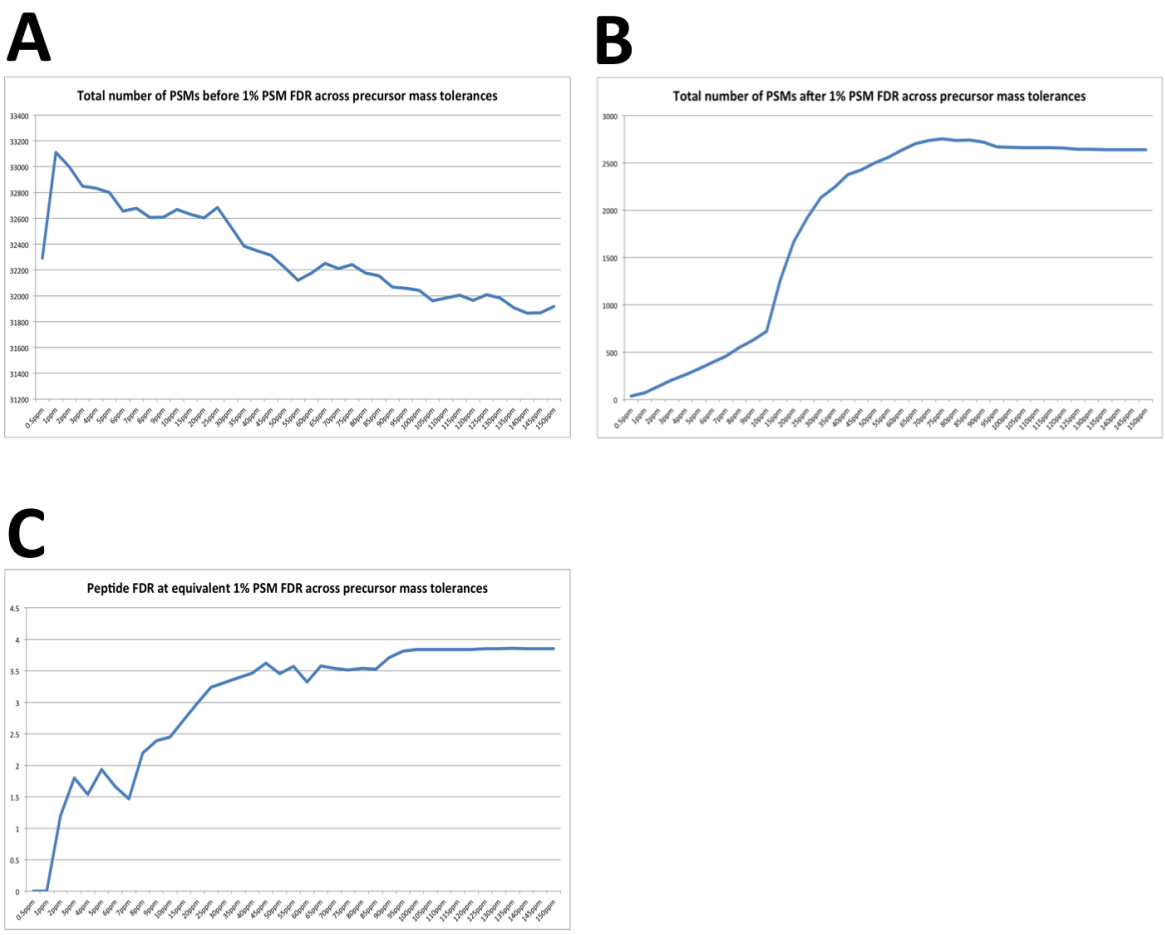
Appendix Figure 7.4 Pre and post PepNovo quality filtering for trypsin-digested meiotic tissue

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



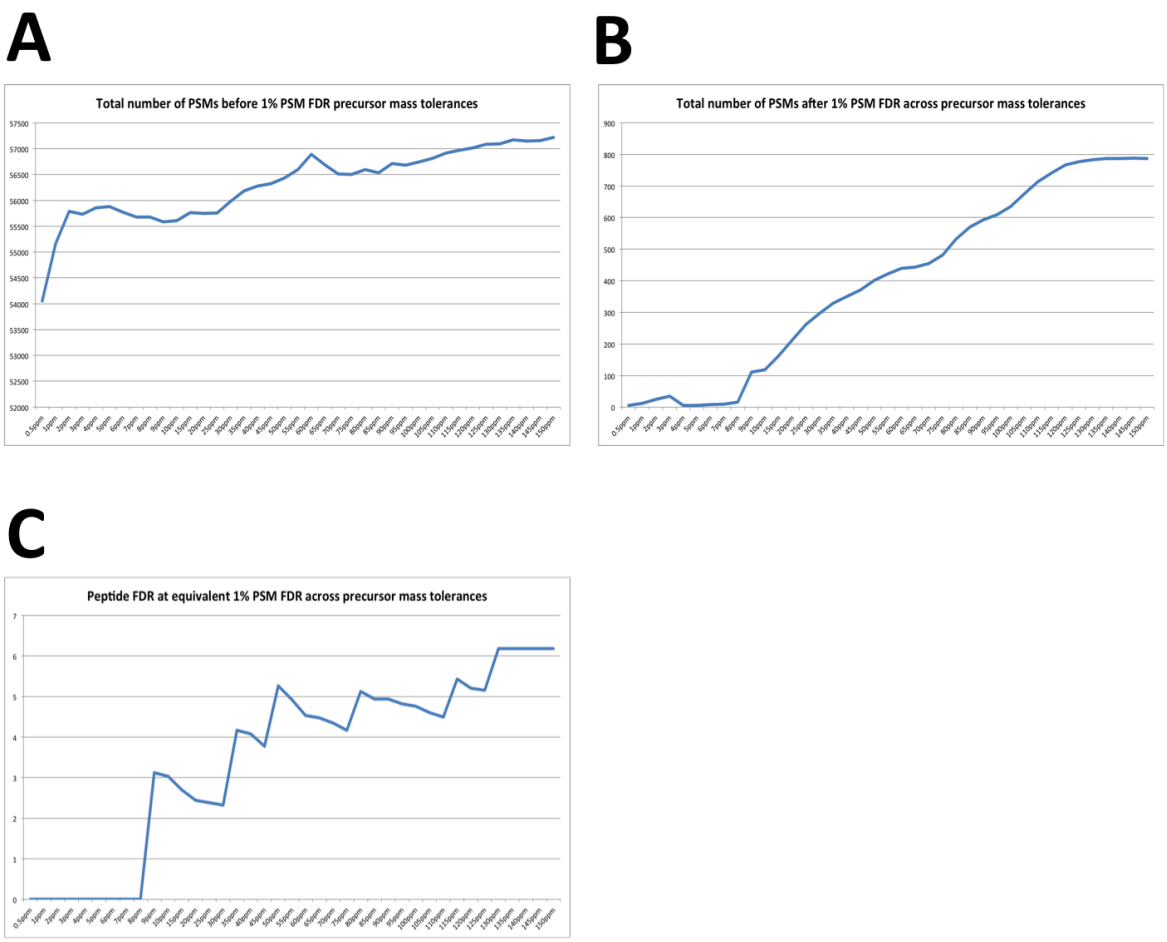
Appendix Figure 7.5 Pre and post PepNovo quality filtering for AspN-digested meiotic tissue

(A) Total spectral counts independent of database search. (B) Peptide FDR at equivalent 1% PSM FDR. (C) Protein FDR at equivalent 1% PSM FDR. (D) Total number of unique peptides after 1% PSM FDR filtering. (E) Total number of PSMs after 1% PSM FDR filtering.



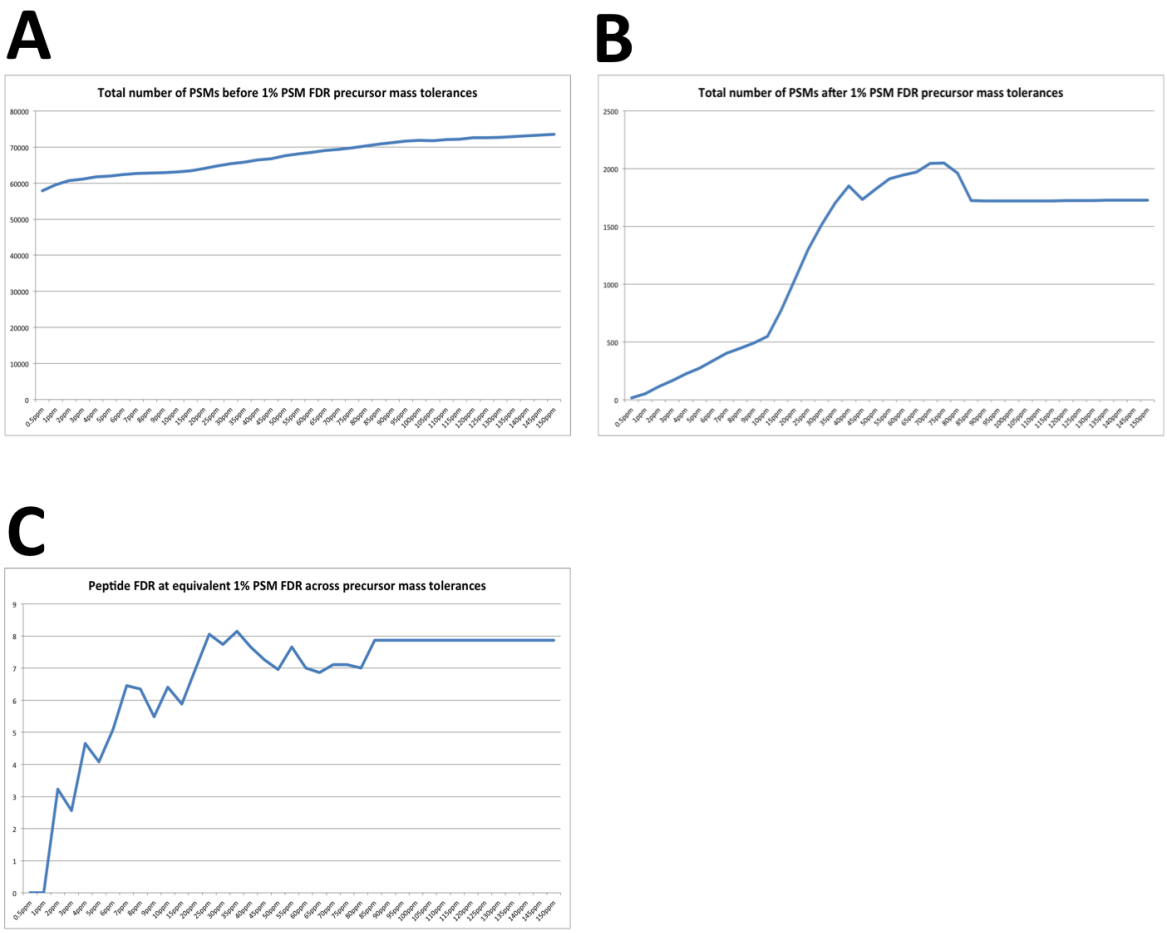
Appendix Figure 7.6 Precursor mass tolerance optimisation for trypsin-digested wheat flour

(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.



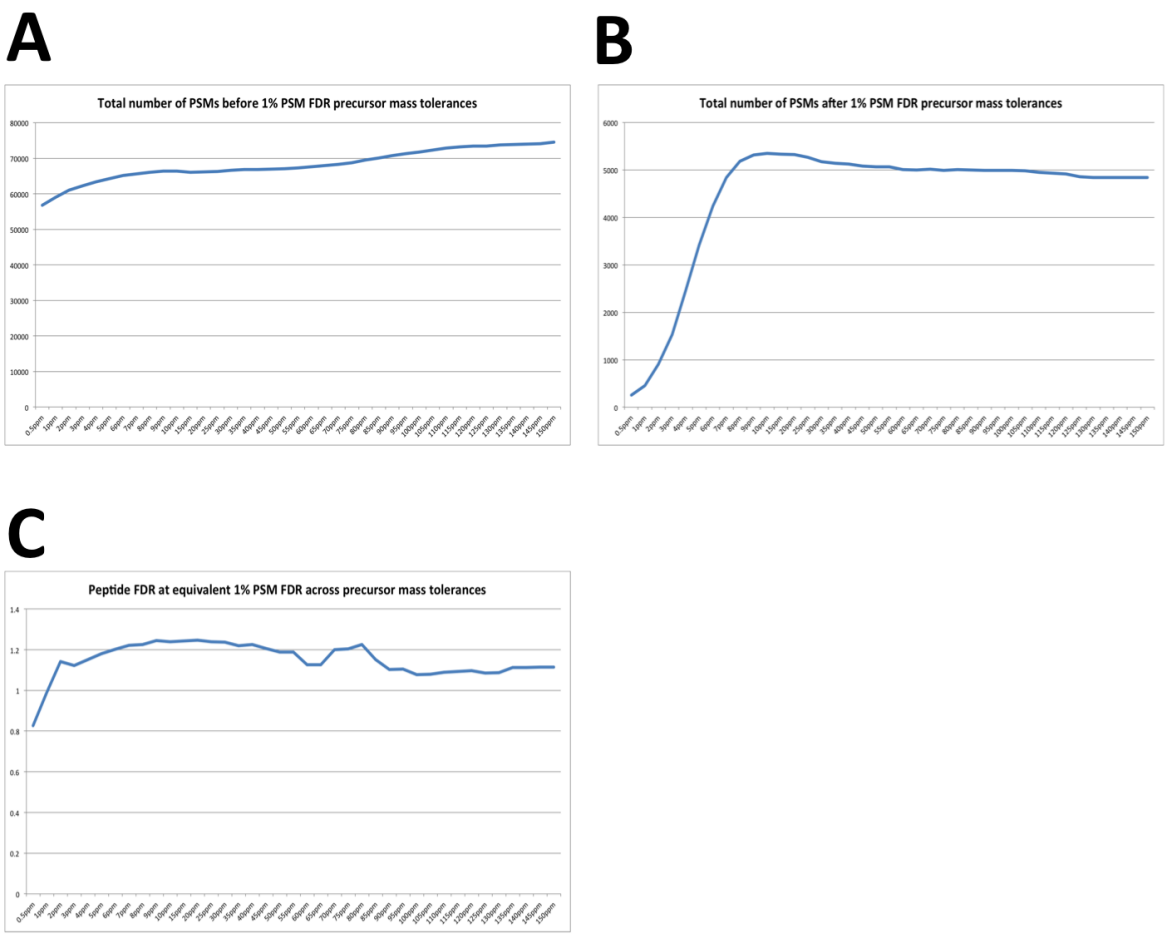
Appendix Figure 7.7 Precursor mass tolerance optimisation for chymotrypsin-digested wheat flour

(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.



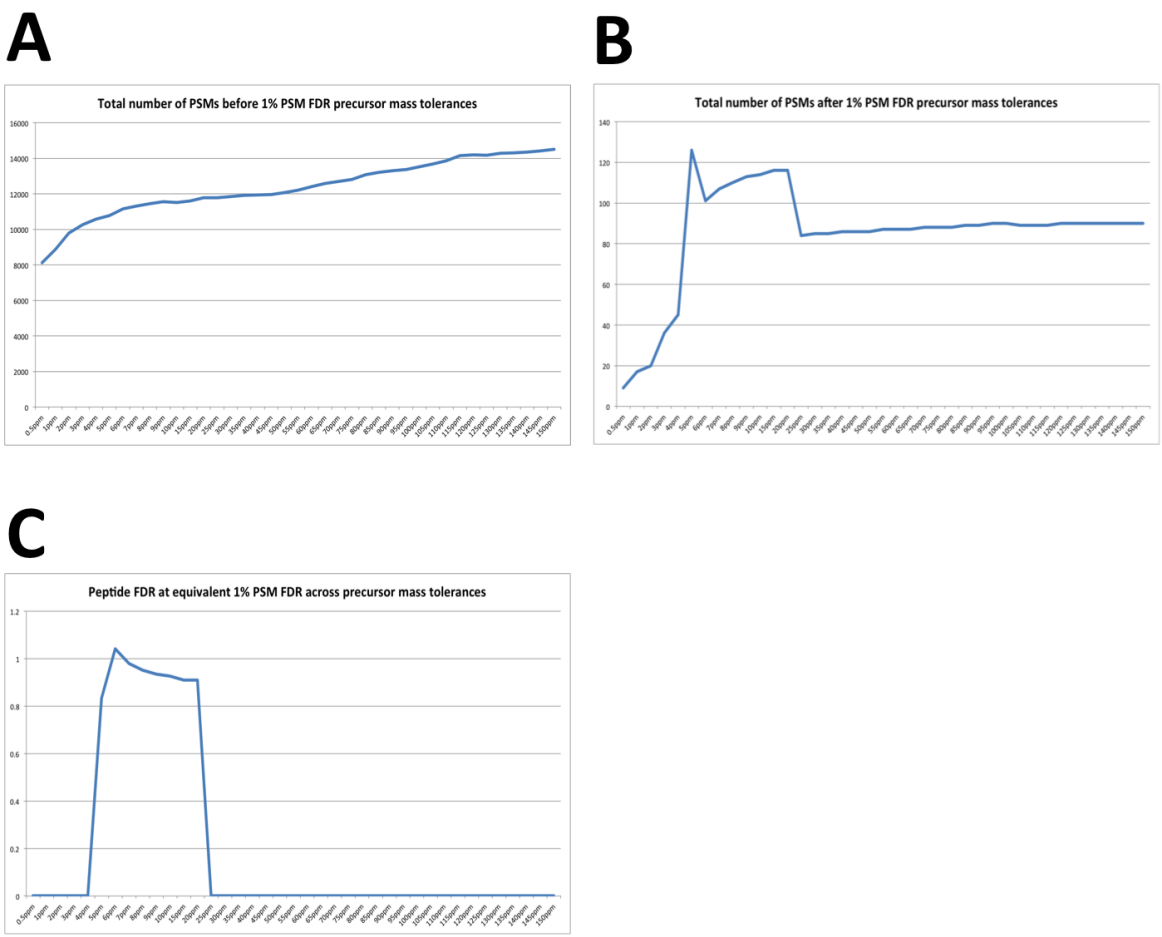
Appendix Figure 7.8 Precursor mass tolerance optimisation for thermolysin-digested wheat flour

(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.



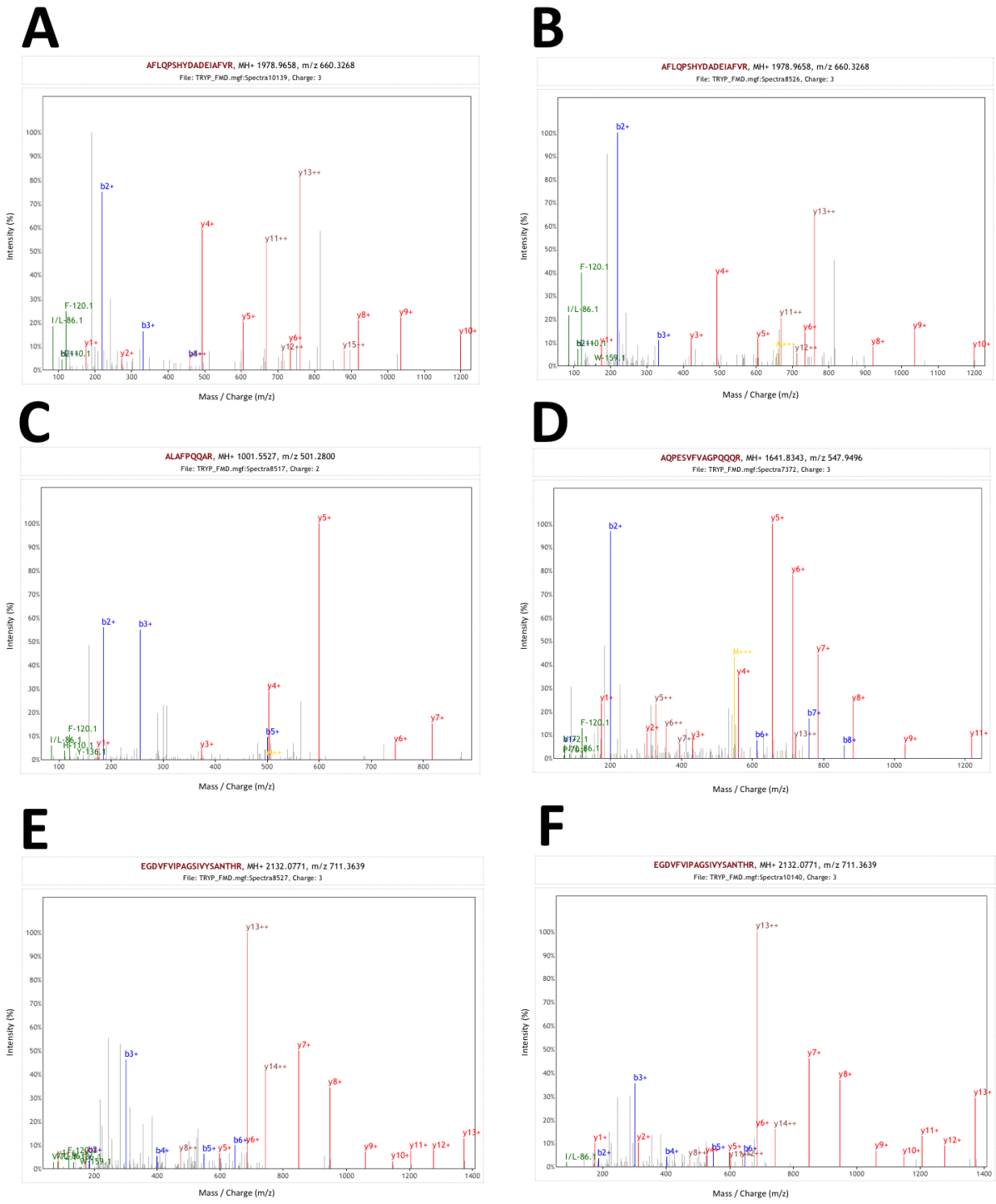
Appendix Figure 7.9 Precursor mass tolerance optimisation for trypsin-digested meiotic tissue

(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.

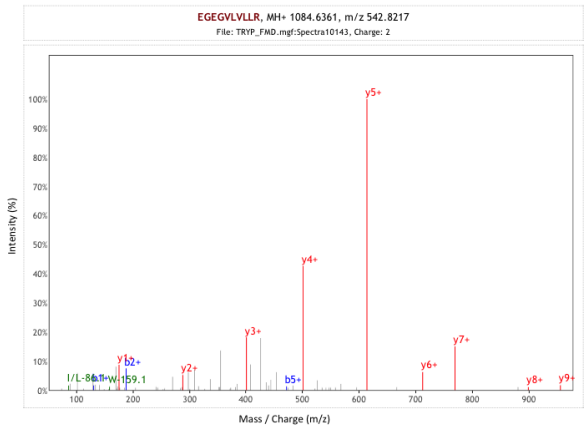


Appendix Figure 7.10 Precursor mass tolerance optimisation for AspN-digested meiotic tissue

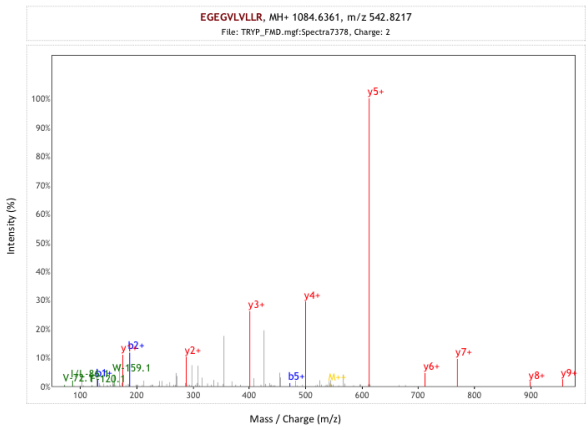
(A) Total number of PSMs before 1% PSM FDR filtering. (B) Total number of PSMs after 1% PSM FDR filtering. (C) Peptide FDR at equivalent 1% PSM FDR across precursor mass tolerances.


A
AFLQPSHYDADEIAFVR, MH+ 1978.9658, m/z 660.3268
File: TRYP_FMD.mgf:Spectra10139, Charge: 3
B
AFLQPSHYDADEIAFVR, MH+ 1978.9658, m/z 660.3268
File: TRYP_FMD.mgf:Spectra8526, Charge: 3
C
ALAFPQQAR, MH+ 1001.5527, m/z 501.2800
File: TRYP_FMD.mgf:Spectra8517, Charge: 2
D
AQPESVFVAGPQQQR, MH+ 1641.8343, m/z 547.9496
File: TRYP_FMD.mgf:Spectra7372, Charge: 3
E
EGDVFVIPAGSIVYSANTHR, MH+ 2132.0771, m/z 711.3639
File: TRYP_FMD.mgf:Spectra8527, Charge: 3
F
EGDVFVIPAGSIVYSANTHR, MH+ 2132.0771, m/z 711.3639
File: TRYP_FMD.mgf:Spectra10140, Charge: 3
Intensity (%)
Mass / Charge (m/z)

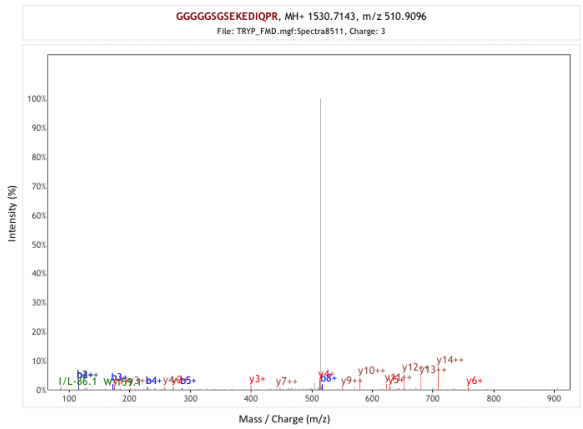
G


EGEGVLVLLR, MH+ 1084.6361, m/z 542.8217
File: TRYP_FMD.mgf:Spectra10143, Charge: 2
Intensity (%)
Mass / Charge (m/z)

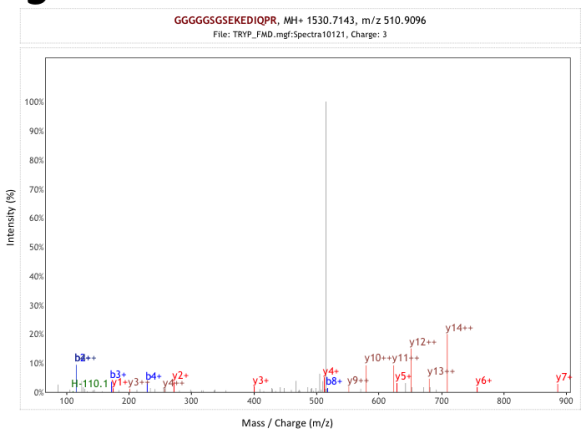
H


EGEGVLVLLR, MH+ 1084.6361, m/z 542.8217
File: TRYP_FMD.mgf:Spectra7378, Charge: 2
Intensity (%)
Mass / Charge (m/z)

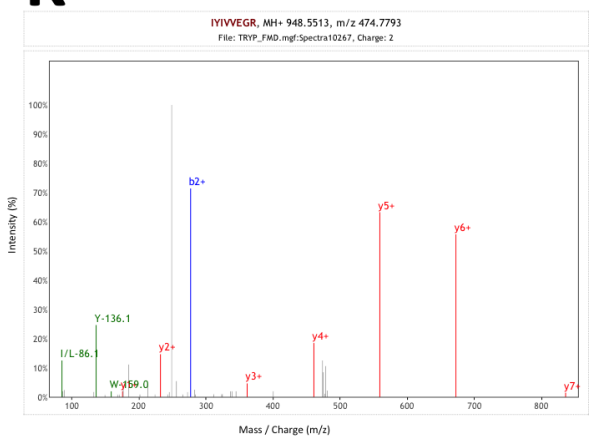
I


GGGGGSGSEKEDIQPR, MH+ 1530.7143, m/z 510.9096
File: TRYP_FMD.mgf:Spectra8511, Charge: 3
Intensity (%)
Mass / Charge (m/z)

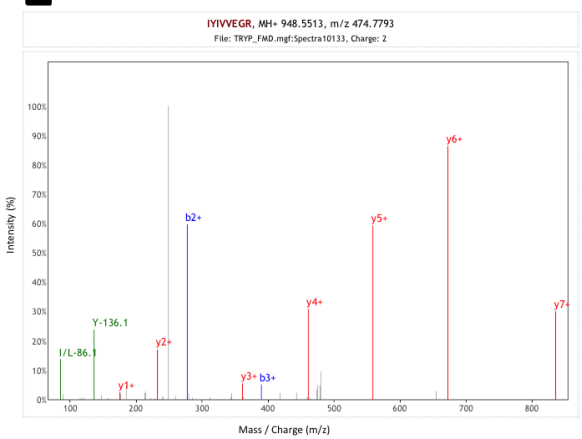
J

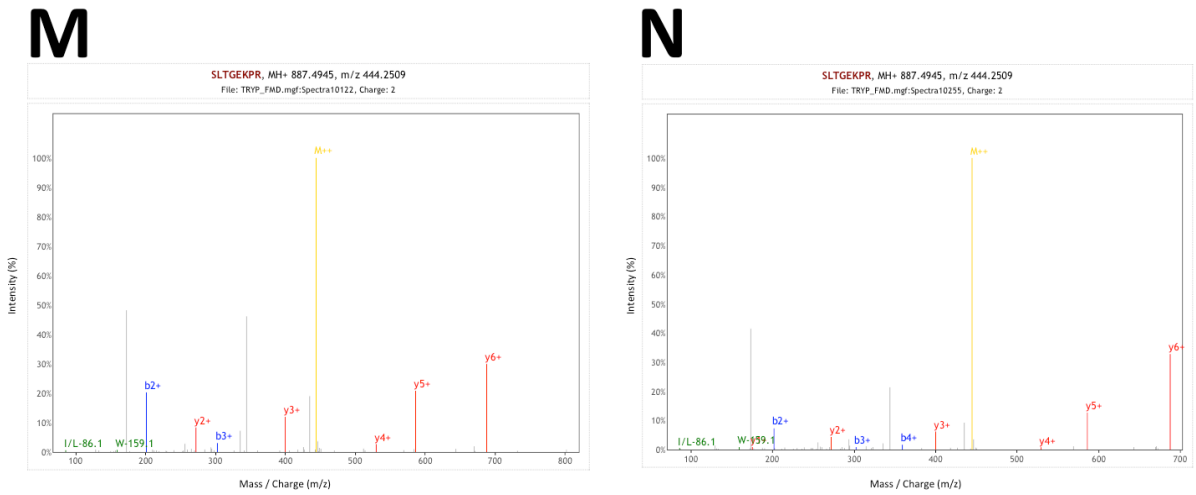

GGGGGSGSEKEDIQPR, MH+ 1530.7143, m/z 510.9096
File: TRYP_FMD.mgf:Spectra10121, Charge: 3
Intensity (%)
Mass / Charge (m/z)

K


IYIVVEGR, MH+ 948.5513, m/z 474.7793
File: TRYP_FMD.mgf:Spectra10267, Charge: 2
Intensity (%)
Mass / Charge (m/z)

L

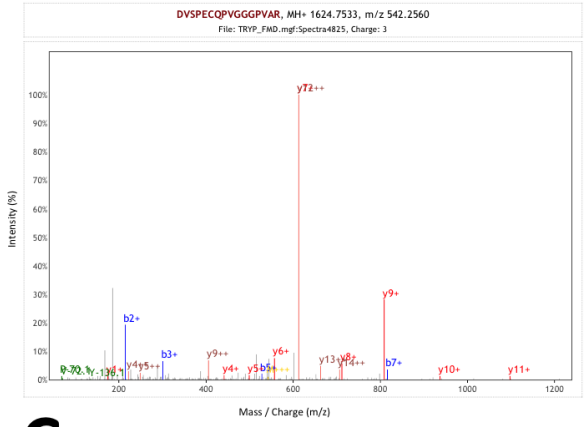

IYIVVEGR, MH+ 948.5513, m/z 474.7793
File: TRYP_FMD.mgf:Spectra10133, Charge: 2
Intensity (%)
Mass / Charge (m/z)



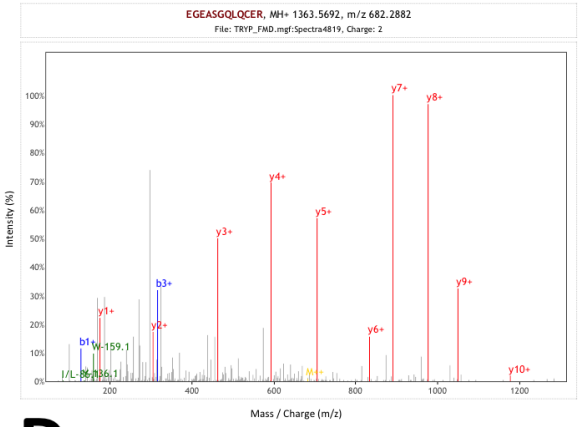
Appendix Figure 7.11 Supporting MS/MS spectra for a novel gene annotation event

Fourteen representative MS/MS spectra (A-N; from a total of 24), supporting novel peptides annotating a novel gene event, illustrated in Figure 7.1.

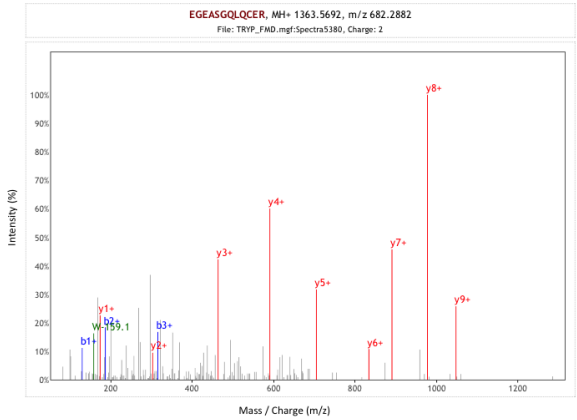
A


DVSPECQPVGGGPVAR, MH+ 1624.7533, m/z 542.2560
File: TRYP_FMD.mgf:Spectra4825, Charge: 3
Intensity (%)
Mass / Charge (m/z)

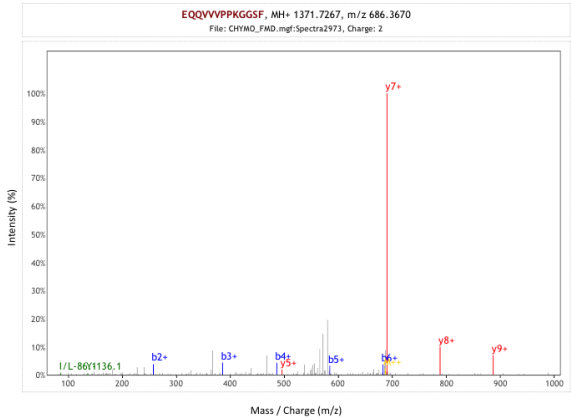
B


EGEASGQLQCER, MH+ 1363.5692, m/z 682.2882
File: TRYP_FMD.mgf:Spectra4819, Charge: 2
Intensity (%)
Mass / Charge (m/z)

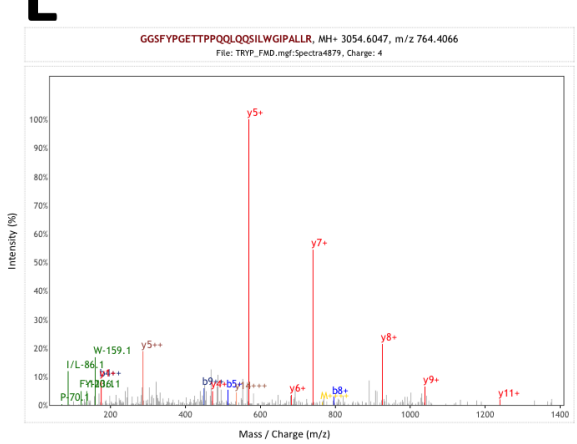
C


EGEASGQLQCER, MH+ 1363.5692, m/z 682.2882
File: TRYP_FMD.mgf:Spectra5380, Charge: 2
Intensity (%)
Mass / Charge (m/z)

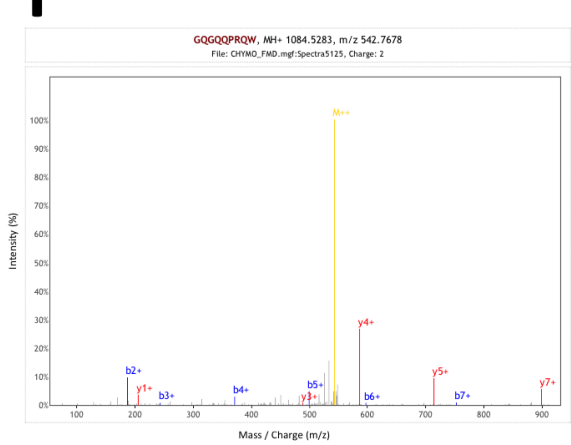
D


EQQVVVPPKGGSF, MH+ 1371.7267, m/z 686.3670
File: CHYMO_FMD.mgf:Spectra2973, Charge: 2
Intensity (%)
Mass / Charge (m/z)

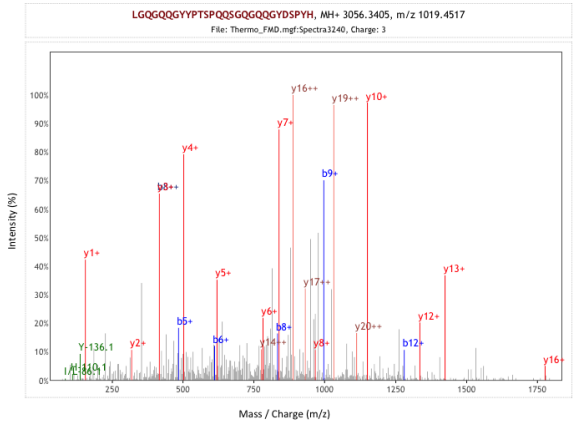
E


GGSFYPGETTPPQQLQQSILWGIPALLR, MH+ 3054.6047, m/z 764.4066
File: TRYP_FMD.mgf:Spectra4879, Charge: 4
Intensity (%)
Mass / Charge (m/z)

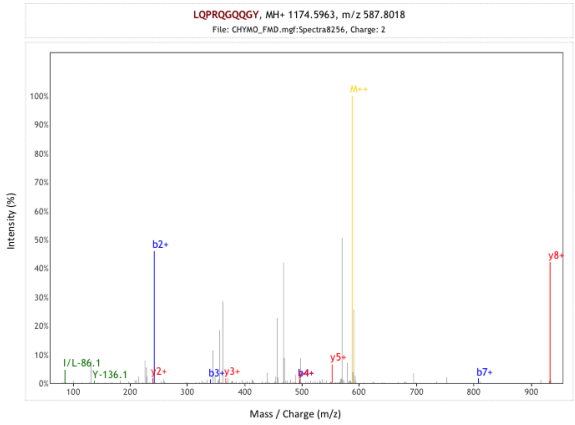
F


GQGQQPRQW, MH+ 1084.5283, m/z 542.7678
File: CHYMO_FMD.mgf:Spectra5125, Charge: 2
Intensity (%)
Mass / Charge (m/z)

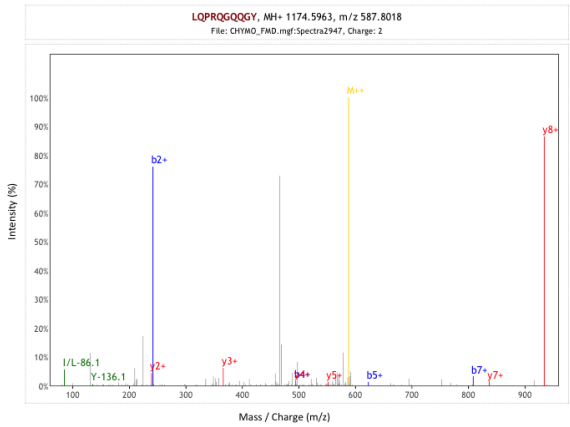
G


LGQGQQGYYPTSPQQSGQGQQGYDSPYH, MH+ 3056.3405, m/z 1019.4517
File: Thermo_FMD.mgf:Spectra3240, Charge: 3
Intensity (%)
Mass / Charge (m/z)

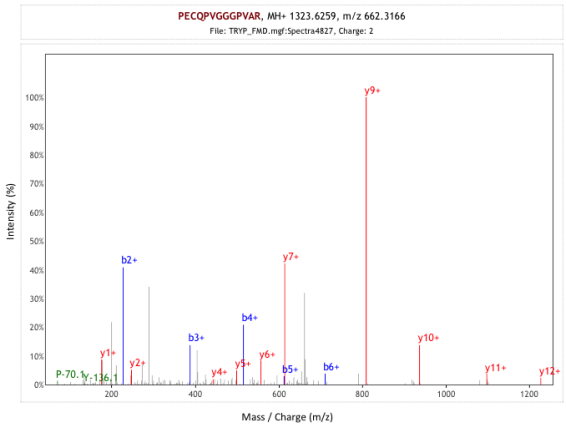
H


LQPRQGQQGY, MH+ 1174.5963, m/z 587.8018
File: CHYMO_FMD.mgf:Spectra8256, Charge: 2
Intensity (%)
Mass / Charge (m/z)

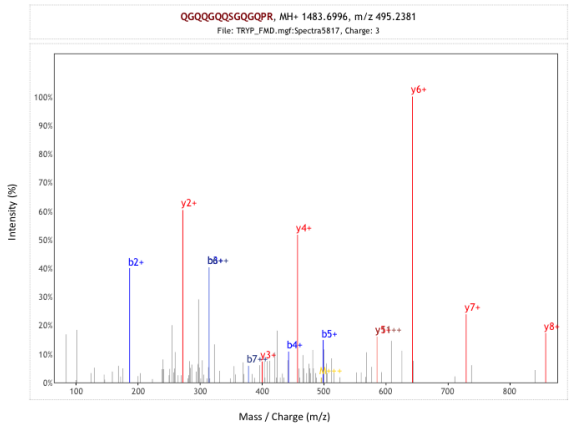
I


LQPRQGQQGY, MH+ 1174.5963, m/z 587.8018
File: CHYMO_FMD.mgf:Spectra2947, Charge: 2
Intensity (%)
Mass / Charge (m/z)

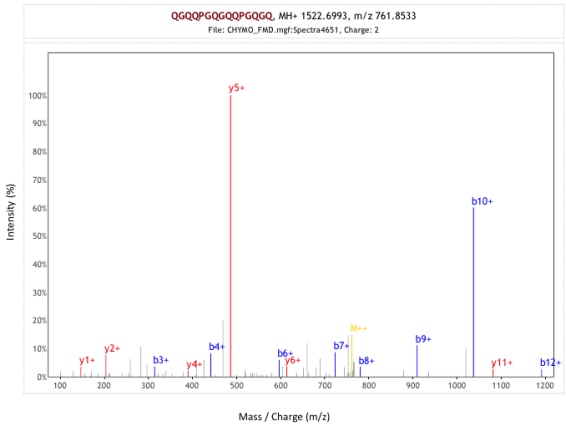
J


PECQPVGGGPVAR, MH+ 1323.6259, m/z 662.3166
File: TRYP_FMD.mgf:Spectra4827, Charge: 2
Intensity (%)
Mass / Charge (m/z)

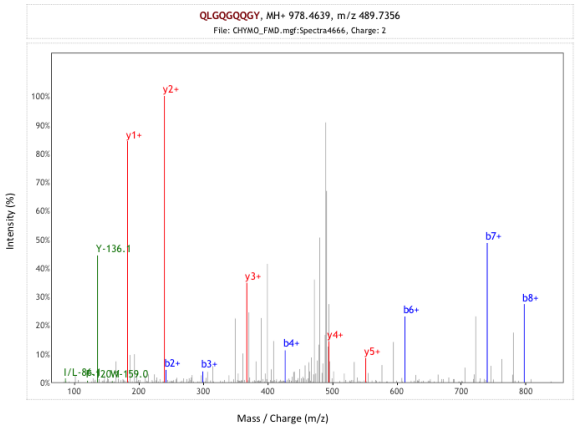
K


QGQQGQQSGQGQPR, MH+ 1483.6996, m/z 495.2381
File: TRYP_FMD.mgf:Spectra5817, Charge: 3
Intensity (%)
Mass / Charge (m/z)

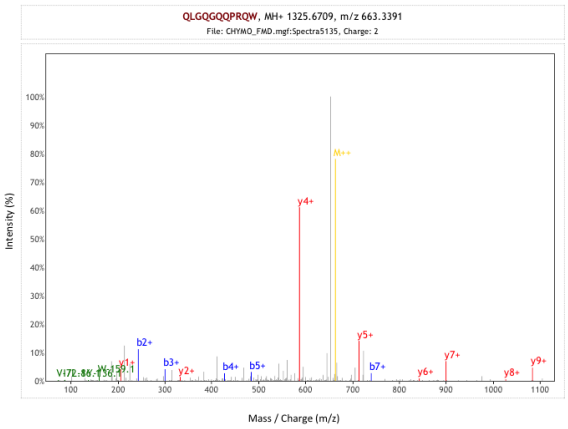
L


QGQQPGQGQQPGQGQ, MH+ 1522.6993, m/z 761.8533
File: CHYMO_FMD.mgf:Spectra4651, Charge: 2
Intensity (%)
Mass / Charge (m/z)

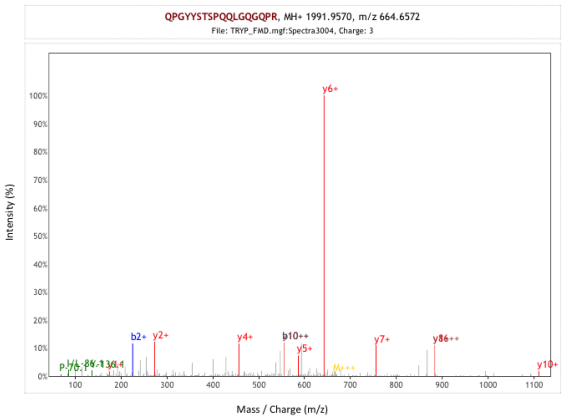
M


QLGQGQQGY, MH+ 978.4639, m/z 489.7356
File: CHYMO_FMD.mgf:Spectra4666, Charge: 2
Intensity (%)
Mass / Charge (m/z)

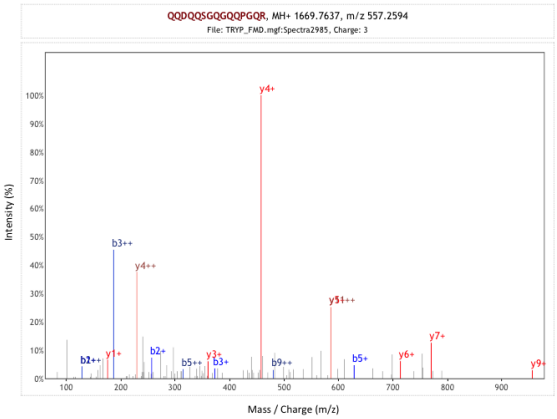
N


QLGQGQQPRQW, MH+ 1325.6709, m/z 663.3391
File: CHYMO_FMD.mgf:Spectra5135, Charge: 2
Intensity (%)
Mass / Charge (m/z)

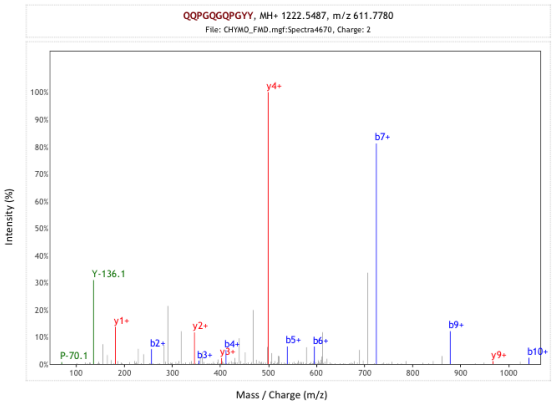
O


QPGYYSTSPQQLGQGQPR, MH+ 1991.9570, m/z 664.6572
File: TRYP_FMD.mgf:Spectra3004, Charge: 3
Intensity (%)
Mass / Charge (m/z)

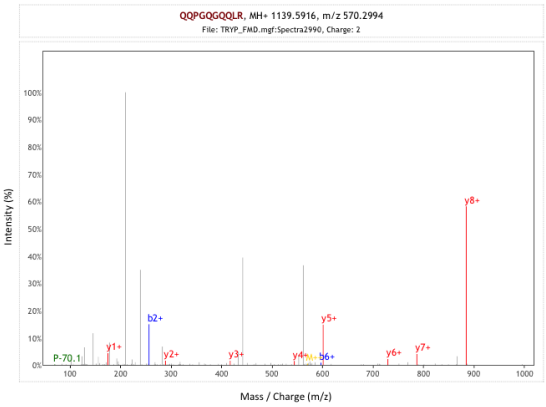
P


QQDQQSGQGQQPGQR, MH+ 1669.7637, m/z 557.2594
File: TRYP_FMD.mgf:Spectra2985, Charge: 3
Intensity (%)
Mass / Charge (m/z)

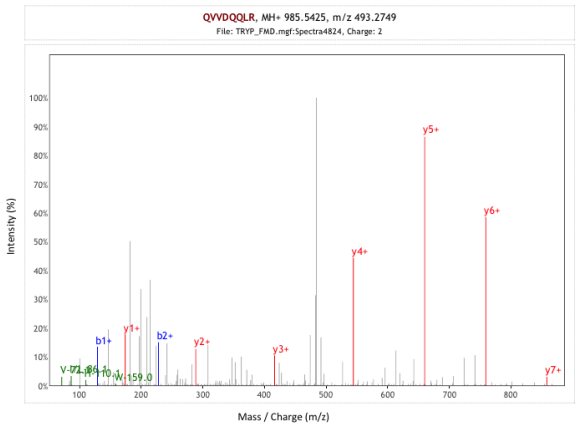
Q


QQPGQGQPGYY, MH+ 1222.5487, m/z 611.7780
File: CHYMO_FMD.mgf:Spectra4670, Charge: 2
Intensity (%)
Mass / Charge (m/z)

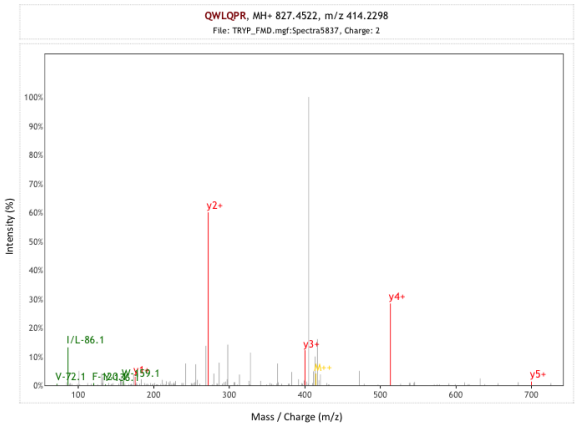
R


QQPGQGQQLR, MH+ 1139.5916, m/z 570.2994
File: TRYP_FMD.mgf:Spectra2990, Charge: 2
Intensity (%)
Mass / Charge (m/z)

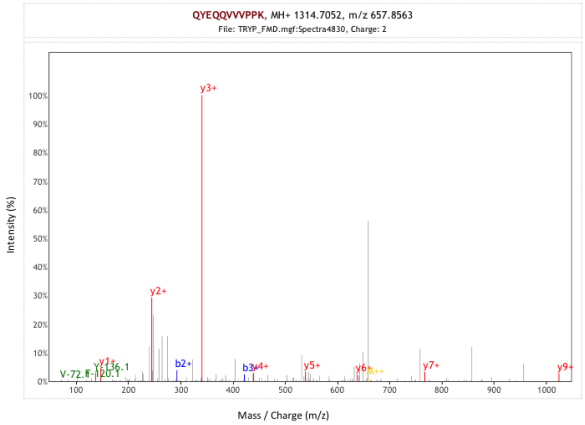
S


QVVDQQLR, MH+ 985.5425, m/z 493.2749
File: TRYP_FMD.mgf:Spectra4824, Charge: 2
Intensity (%)
Mass / Charge (m/z)

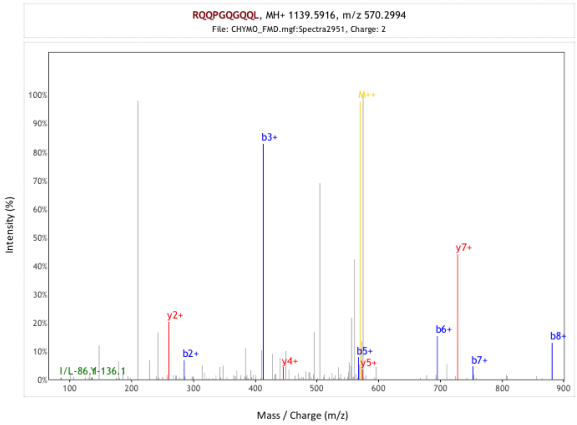
T


QWLQPR, MH+ 827.4522, m/z 414.2298
File: TRYP_FMD.mgf:Spectra5837, Charge: 2
Intensity (%)
Mass / Charge (m/z)

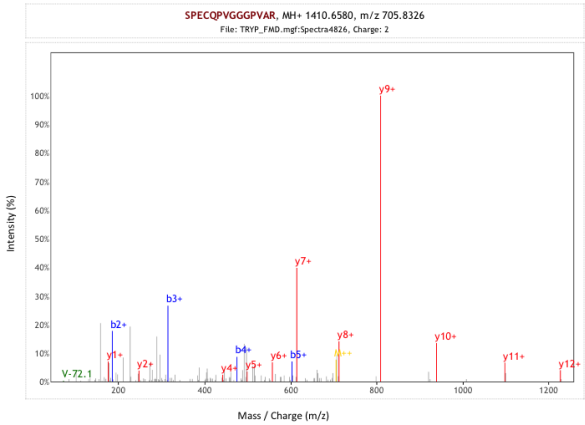
U


QYEQQVVVPPK, MH+ 1314.7052, m/z 657.8563
File: TRYP_FMD.mgf:Spectra4830, Charge: 2
Intensity (%)
Mass / Charge (m/z)

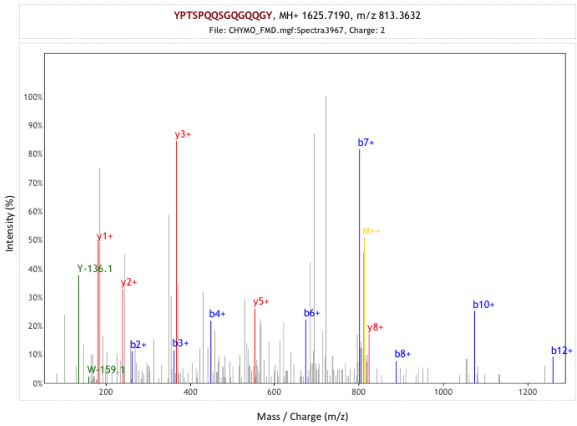
V

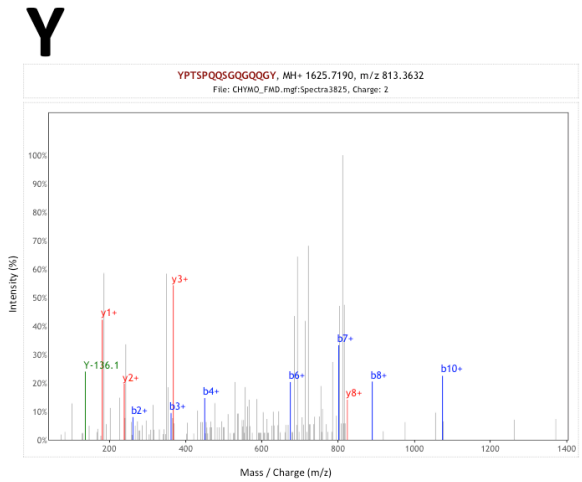

RQQPGQGQQL, MH+ 1139.5916, m/z 570.2994
File: CHYMO_FMD.mgf:Spectra2951, Charge: 2
Intensity (%)
Mass / Charge (m/z)

W


SPECQPVGGGPVAR, MH+ 1410.6580, m/z 705.8326
File: TRYP_FMD.mgf:Spectra4826, Charge: 2
Intensity (%)
Mass / Charge (m/z)

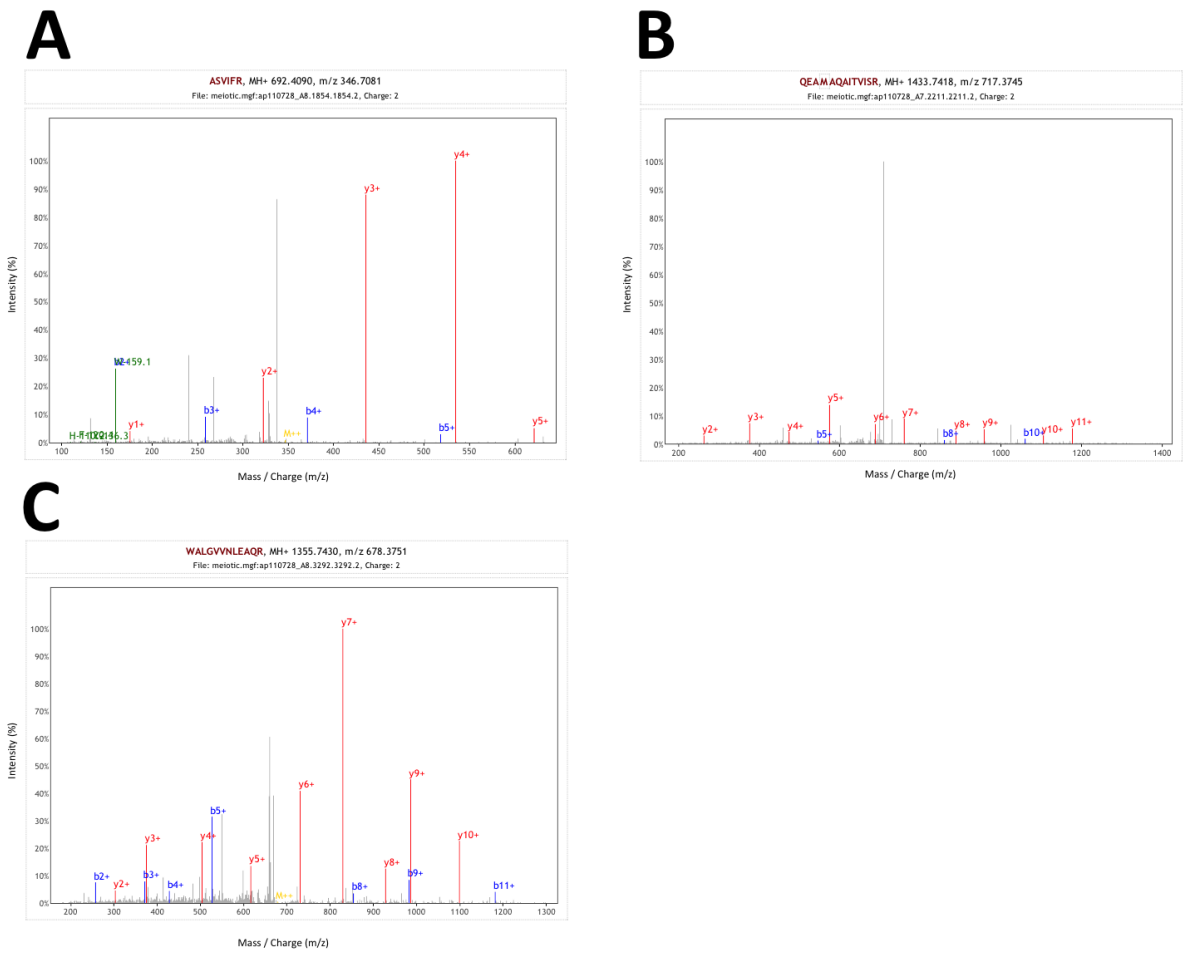
X


YPTSPQQSGQGQQGY, MH+ 1625.7190, m/z 813.3632
File: CHYMO_FMD.mgf:Spectra3967, Charge: 2
Intensity (%)
Mass / Charge (m/z)



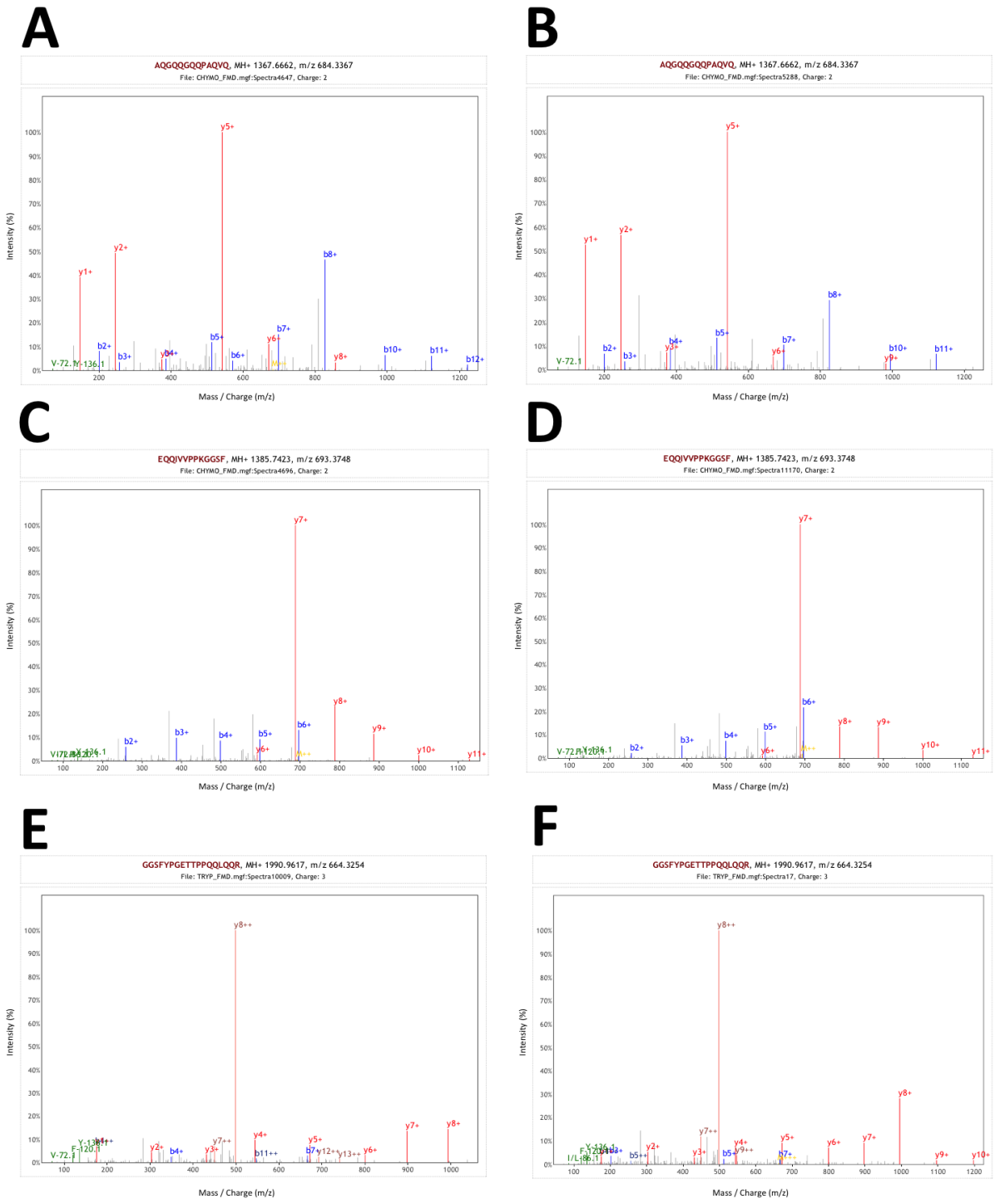
Appendix Figure 7.12 Supporting MS/MS spectra for a novel gene annotation misidentified as a gene boundary annotation event

Twenty-five representative MS/MS spectra (A-Y; from a total of 50) supporting novel peptides annotating a gene boundary event, which was actually a novel gene annotation, illustrated in Figure 7.2.

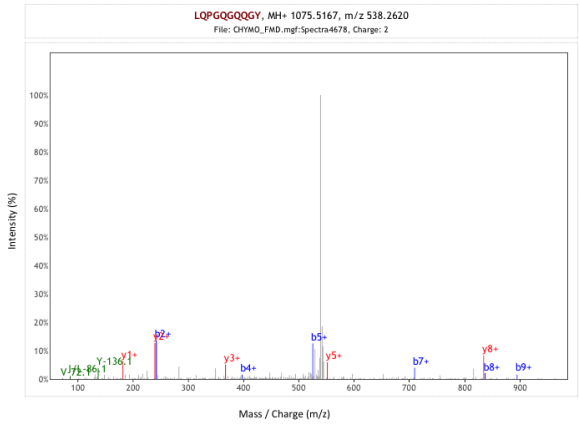


Appendix Figure 7.13 Supporting MS/MS spectra for a reverse strand annotation event

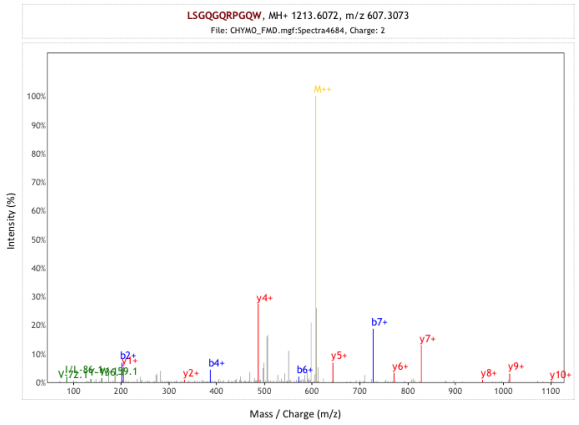
Three MS/MS spectra (A-C) supporting novel peptides annotating a reverse strand event, illustrated in Figure 7.3.


A
AQGQQGQQPAQVQ, MH+ 1367.6662, m/z 684.3367
File: CHYMO_FMD.mgf:Spectra4647, Charge: 2
B
AQGQQGQQPAQVQ, MH+ 1367.6662, m/z 684.3367
File: CHYMO_FMD.mgf:Spectra5288, Charge: 2
C
EQQIVVPPKGGSF, MH+ 1385.7423, m/z 693.3748
File: CHYMO_FMD.mgf:Spectra4696, Charge: 2
D
EQQIVVPPKGGSF, MH+ 1385.7423, m/z 693.3748
File: CHYMO_FMD.mgf:Spectra11170, Charge: 2
E
GGSFYPGETTPPQQLQQR, MH+ 1990.9617, m/z 664.3254
File: TRYP_FMD.mgf:Spectra10009, Charge: 3
F
GGSFYPGETTPPQQLQQR, MH+ 1990.9617, m/z 664.3254
File: TRYP_FMD.mgf:Spectra17, Charge: 3
Intensity (%)
Mass / Charge (m/z)

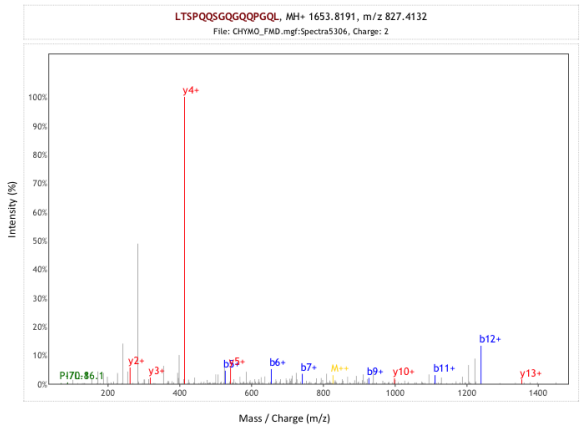
G


LQPGQGQQGY, MH+ 1075.5167, m/z 538.2620
File: CHYMO_FMD.mgf:Spectra4678, Charge: 2
Intensity (%)
Mass / Charge (m/z)

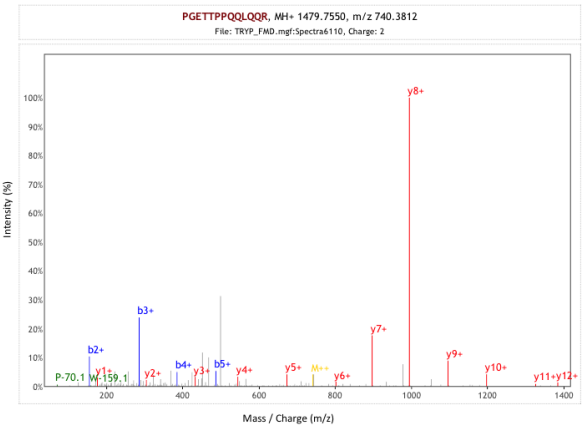
H


LSGQGQRPGQW, MH+ 1213.6072, m/z 607.3073
File: CHYMO_FMD.mgf:Spectra4684, Charge: 2
Intensity (%)
Mass / Charge (m/z)

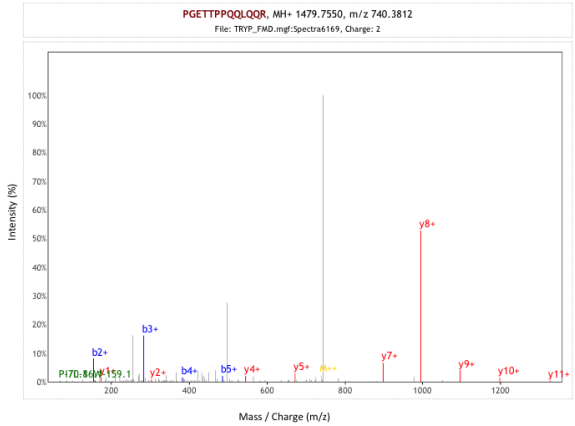
I


LTSPQQSGQGQQPGQL, MH+ 1653.8191, m/z 827.4132
File: CHYMO_FMD.mgf:Spectra5306, Charge: 2
Intensity (%)
Mass / Charge (m/z)

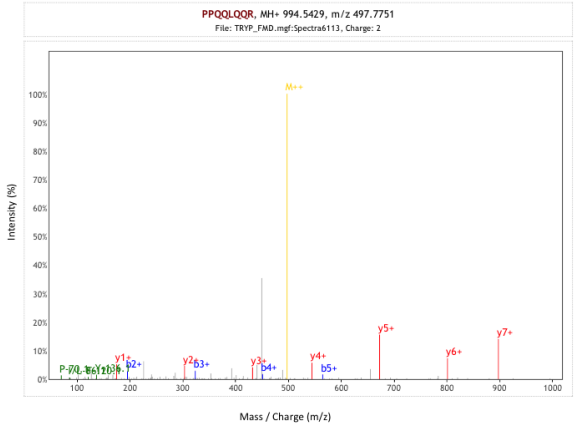
J


PGETTPPQQLQQR, MH+ 1479.7550, m/z 740.3812
File: TRYP_FMD.mgf:Spectra6110, Charge: 2
Intensity (%)
Mass / Charge (m/z)

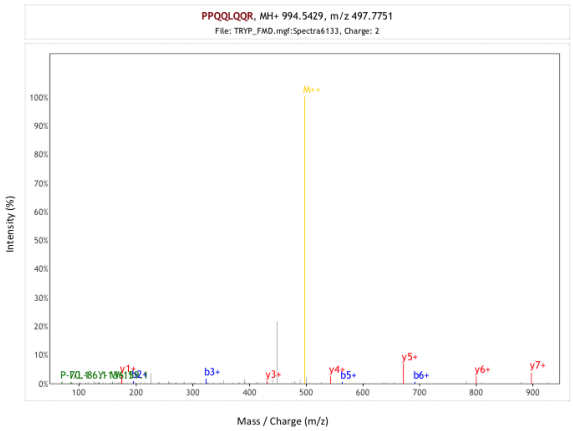
K


PGETTPPQQLQQR, MH+ 1479.7550, m/z 740.3812
File: TRYP_FMD.mgf:Spectra6169, Charge: 2
Intensity (%)
Mass / Charge (m/z)

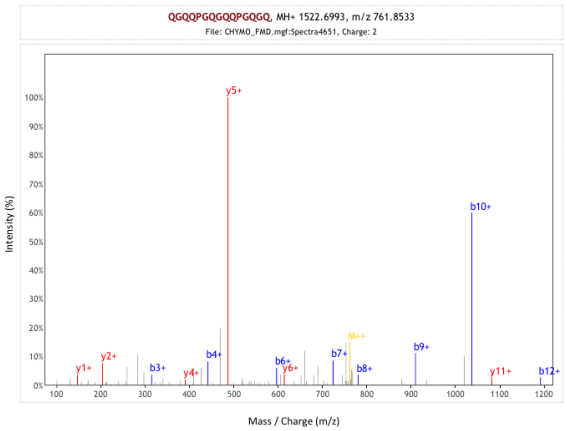
L


PPQQLQQR, MH+ 994.5429, m/z 497.7751
File: TRYP_FMD.mgf:Spectra6113, Charge: 2
Intensity (%)
Mass / Charge (m/z)

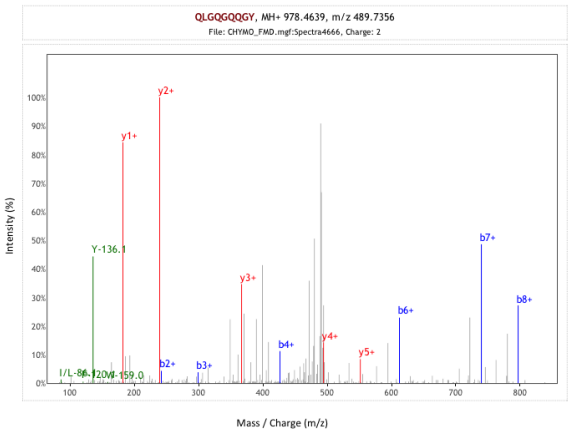
M


PPQQLQQR, MH+ 994.5429, m/z 497.7751
File: TRYP_FMD.mgf:Spectra6133, Charge: 2
Intensity (%)
Mass / Charge (m/z)

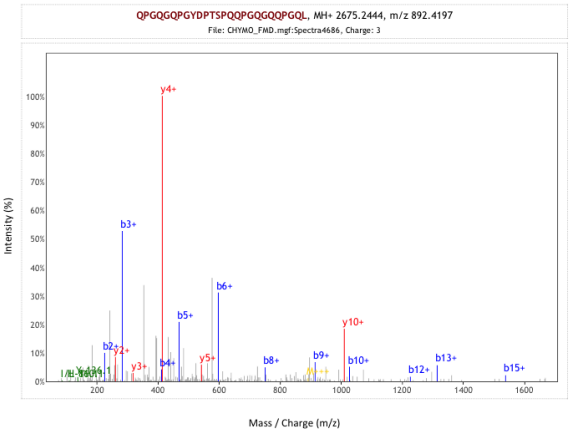
N


QGQQPGQGQQPGQGQ, MH+ 1522.6993, m/z 761.8533
File: CHYMO_FMD.mgf:Spectra4651, Charge: 2
Intensity (%)
Mass / Charge (m/z)

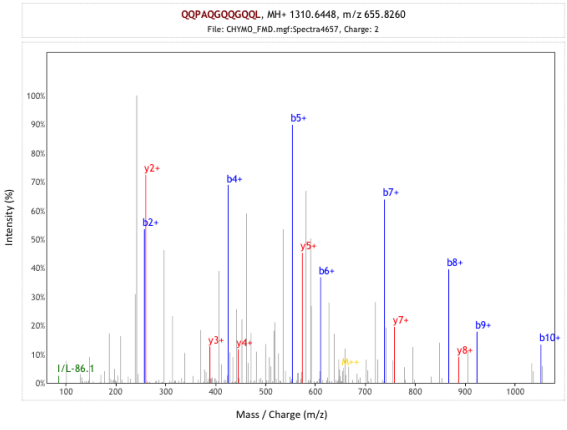
O


QLGQGQQGY, MH+ 978.4639, m/z 489.7356
File: CHYMO_FMD.mgf:Spectra4666, Charge: 2
Intensity (%)
Mass / Charge (m/z)

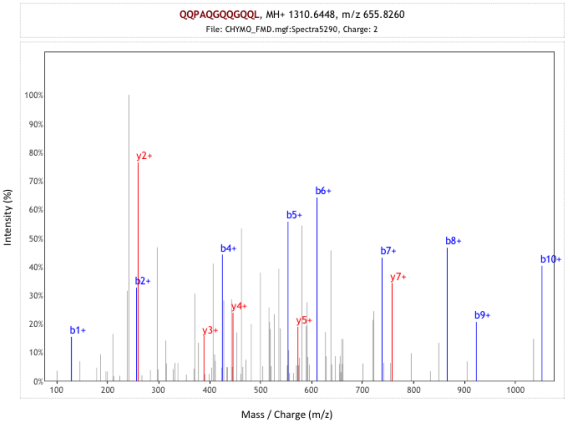
P


QPGQGQPGYDPTSPQQPGQGQQPGQL, MH+ 2675.2444, m/z 892.4197
File: CHYMO_FMD.mgf:Spectra4686, Charge: 3
Intensity (%)
Mass / Charge (m/z)

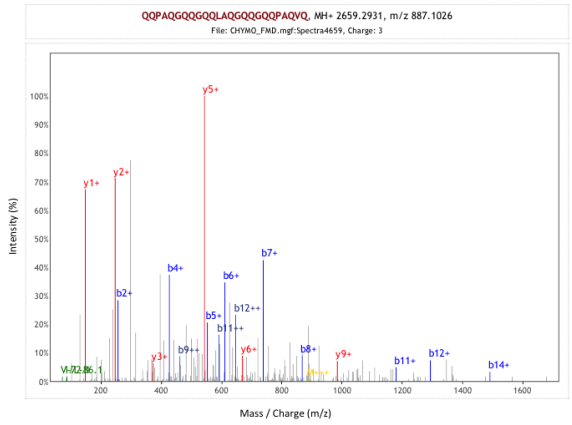
Q


QQPAQGQQGQQL, MH+ 1310.6448, m/z 655.8260
File: CHYMO_FMD.mgf:Spectra4657, Charge: 2
Intensity (%)
Mass / Charge (m/z)

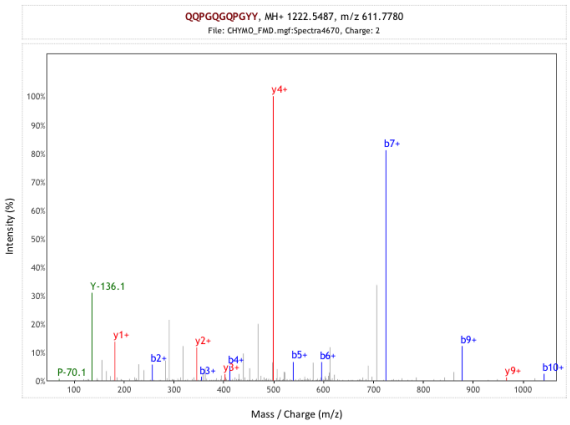
R


QQPAQGQQGQQL, MH+ 1310.6448, m/z 655.8260
File: CHYMO_FMD.mgf:Spectra5290, Charge: 2
Intensity (%)
Mass / Charge (m/z)

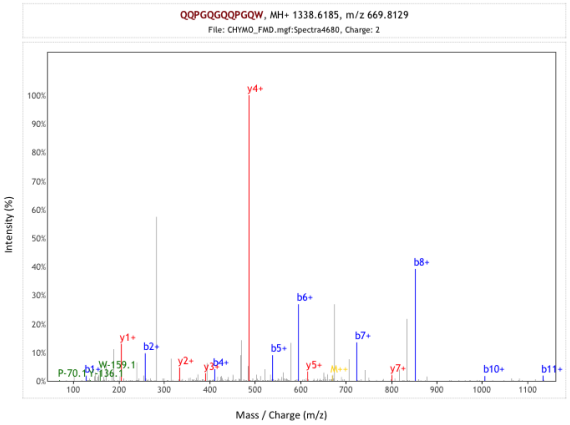
S


QQPAQGQQGQQLAQGQQGQQPAQVQ, MH+ 2659.2931, m/z 887.1026
File: CHYMO_FMD.mgf:Spectra4659, Charge: 3
Intensity (%)
Mass / Charge (m/z)

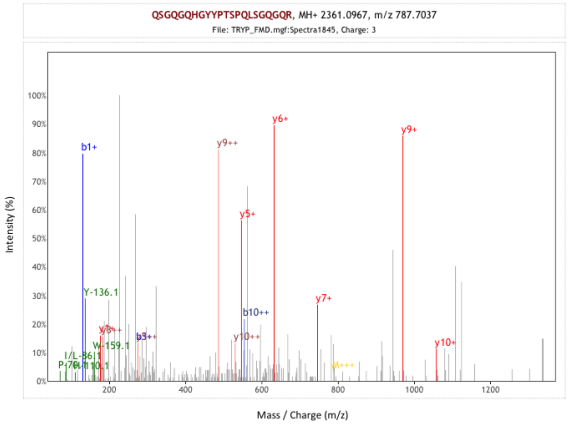
T


QQPGQGQPGYY, MH+ 1222.5487, m/z 611.7780
File: CHYMO_FMD.mgf:Spectra4670, Charge: 2
Intensity (%)
Mass / Charge (m/z)

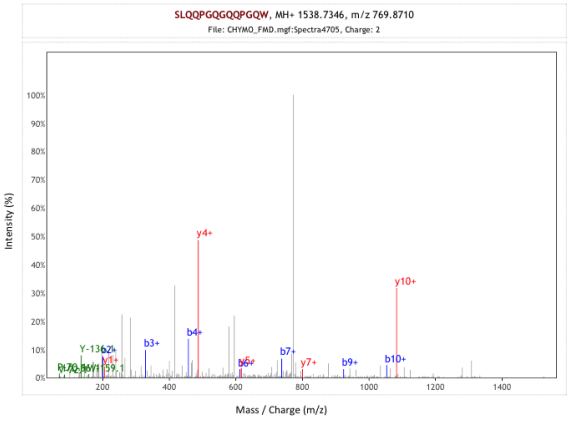
U


QQPGQGQQPGQW, MH+ 1338.6185, m/z 669.8129
File: CHYMO_FMD.mgf:Spectra4680, Charge: 2
Intensity (%)
Mass / Charge (m/z)

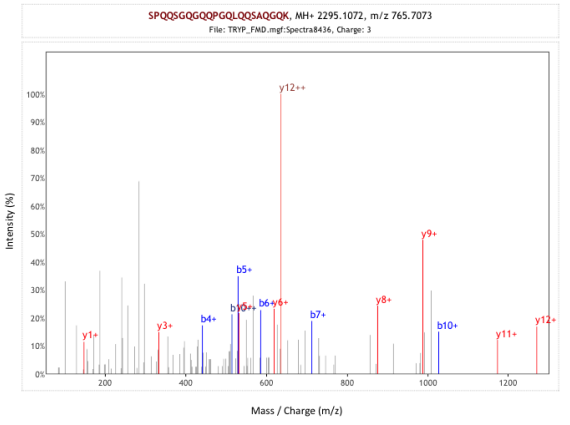
V

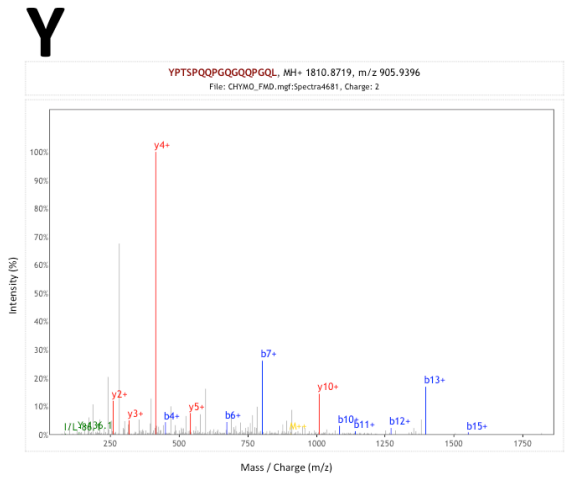

QSGQGQHGYYPTSPQLSGQGQR, MH+ 2361.0967, m/z 787.7037
File: TRYP_FMD.mgf:Spectra1845, Charge: 3
Intensity (%)
Mass / Charge (m/z)

W


SLQQPGQGQQPGQW, MH+ 1538.7346, m/z 769.8710
File: CHYMO_FMD.mgf:Spectra4705, Charge: 2
Intensity (%)
Mass / Charge (m/z)

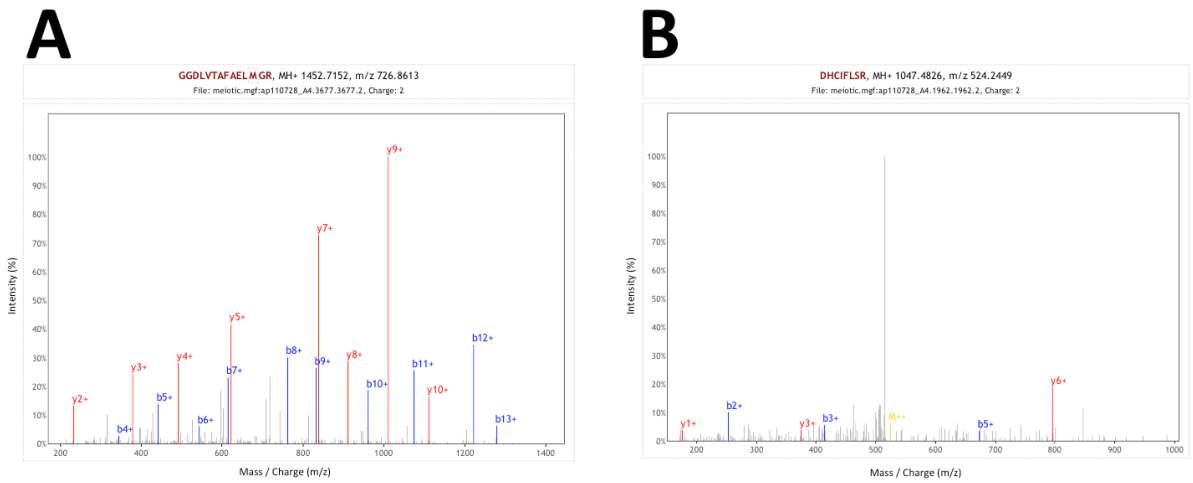
X


SPQQSGQGQQPGQLQQSAQGQK, MH+ 2295.1072, m/z 765.7073
File: TRYP_FMD.mgf:Spectra8436, Charge: 3
Intensity (%)
Mass / Charge (m/z)



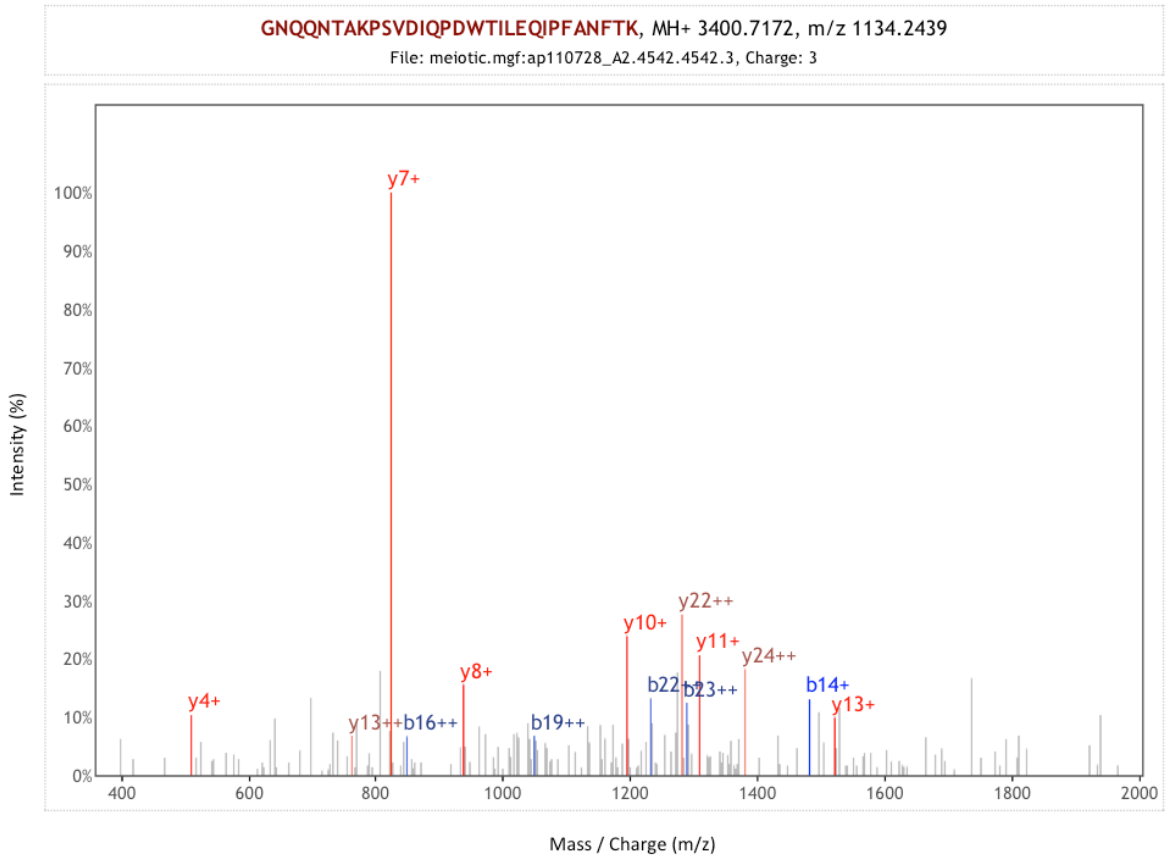
Appendix Figure 7.14 Supporting MS/MS spectra for an exon boundary annotation via a translated UTR annotation event

Twenty-five representative MS/MS spectra (A-Y; from a total of 34) supporting novel peptides annotating a translated UTR event, which was more likely an exon boundary annotation, illustrated in Figure 7.4.



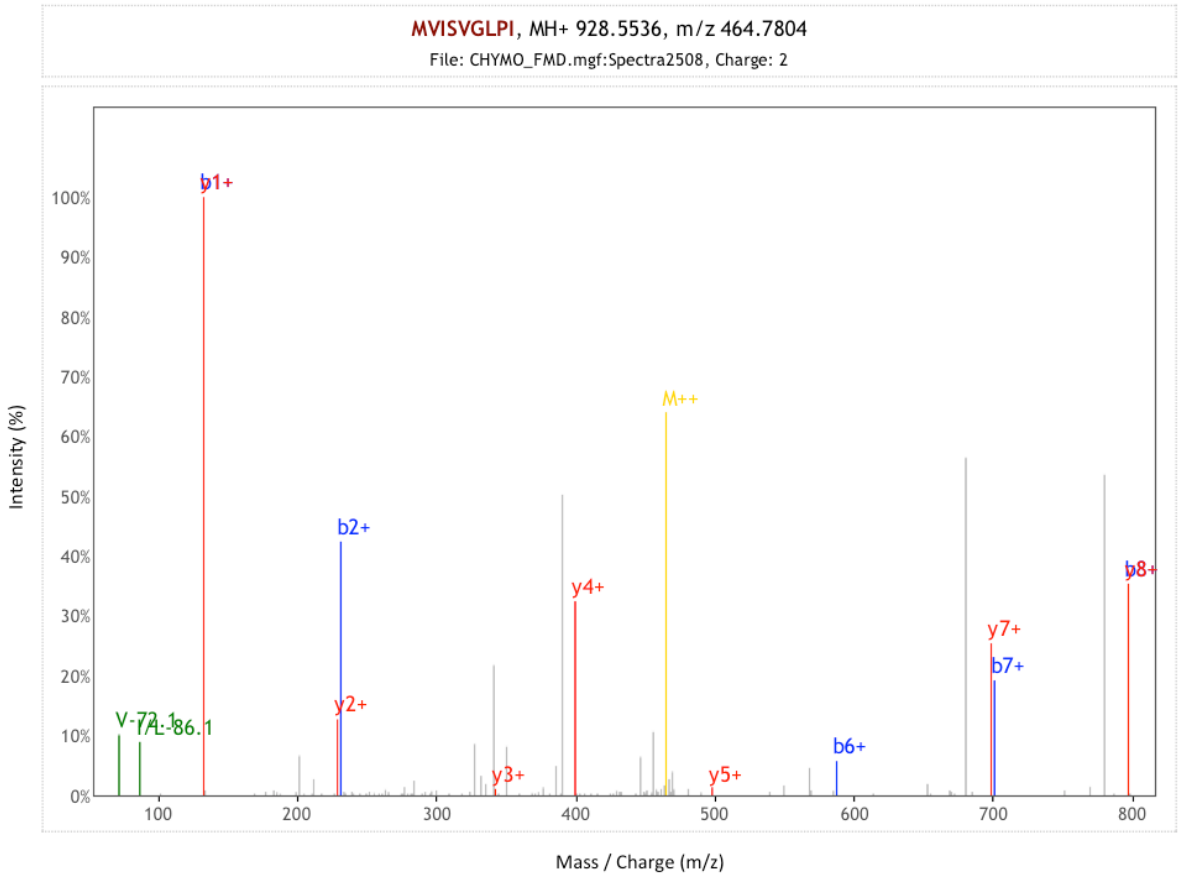
Appendix Figure 7.15 Supporting MS/MS spectra for an exon boundary annotation event

Two MS/MS spectra (A-B) supporting novel peptides annotating an exon boundary event, illustrated in Figure 7.5.



Appendix Figure 7.16 Supporting MS/MS spectrum for a frame-shift annotation event

A single MS/MS spectrum supporting a novel peptide annotating a frame-shift event, illustrated in Figure 7.6.



Appendix Figure 7.17 Supporting MS/MS spectrum for a novel exon annotation event

A single MS/MS spectrum supporting a novel peptide annotating a novel exon event, illustrated in Figure 7.7.

REFERENCES

1. Pennisi E: **DNA Sequencing No Genome Left Behind**. *Science* 2009, **326**(5954):794-795.

2. Azvolinsky A: **Sequencing the Tree of Life**. In: *The Scientist.* 2014.

3. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM: **Non-model organisms, a species endangered by proteogenomics**. *J Proteomics* 2014, **105**:5-18.

4. Delmotte N, Ahrens CH, Knief C, Qeli E, Koch M, Fischer HM, Vorholt JA, Hennecke H, Pessi G: **An integrated proteomics and transcriptomics reference data set provides new insights into the Bradyrhizobium japonicum bacteroid metabolism in soybean root nodules**. *Proteomics* 2010, **10**(7):1391-1400.

5. Koch M, Delmotte N, Rehrauer H, Vorholt JA, Pessi G, Hennecke H: **Rhizobial adaptation to hosts, a new facet in the legume root-nodule symbiosis**. *Molecular plant-microbe interactions : MPMI* 2010, **23**(6):784-790.

6. Chapman B, Bellgard M: **High-throughput parallel proteogenomics: A bacterial case study**. *Proteomics* 2014, **14**(23-24):2780-2789.

7. Vitulo N, Forcato C, Carpinelli EC, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C *et al*: **A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype**. *BMC Plant Biol* 2014, **14**:99.

8. Chapman B, Castellana N, Apffel A, Ghan R, Cramer GR, Bellgard M, Haynes PA, Van Sluyter SC: **Plant proteogenomics: from protein extraction to improved gene predictions**. *Methods in molecular biology* 2013, **1002**:267-294.

9. Cramer GR, Van Sluyter SC, Hopper DW, Pascovici D, Keighley T, Haynes PA: **Proteomic analysis indicates massive changes in metabolism prior to the inhibition of growth and photosynthesis of grapevine (Vitis vinifera L.) in response to water deficit**. *BMC Plant Biol* 2013, **13**:49.

10. Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Dal Santo S, Minio A, Buson G, Tononi P, Zago ED *et al*: **De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity**. *BMC Genomics* 2013, **14**:41.

11. Khatun J, Yu Y, Wrobel JA, Risk BA, Gunawardena HP, Secrest A, Spitzer WJ, Xie L, Wang L, Chen X *et al*: **Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions**. *BMC Genomics* 2013, **14**:141.

12. Mayer K, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski A, Sourdille P *et al*: **A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome**. *Science* 2014, **345**(6194):1251788.

13. Dupont FM, Vensel WH, Tanaka CK, Hurkman WJ, Altenbach SB: **Deciphering the complexities of the wheat flour proteome using quantitative two-dimensional electrophoresis, three proteases and tandem mass spectrometry**. *Proteome Sci* 2011, **9**:10.

14. Greer E, Martin AC, Pendle A, Colas I, Jones AM, Moore G, Shaw P: **The Ph1 locus suppresses Cdk2-type activity during premeiosis and meiosis in wheat**. *The Plant Cell* 2012, **24**(1):152-162.

15. Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KF, Olsen OA: **Genome interplay in the grain transcriptome of hexaploid bread wheat**. *Science* 2014, **345**(6194):1250091.

16. Crick FH: **On protein synthesis**. *Symp Soc Exp Biol* 1958, **12**:138-163.

17. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**(5258):561-563.

18. Konarska MM, Padgett RA, Sharp PA: **Recognition of cap structure in splicing in vitro of mRNA precursors**. *Cell* 1984, **38**(3):731-736.

19. Ohno M, Sakamoto H, Shimura Y: **Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure**. *P Natl Acad Sci USA* 1987, **84**(15):5187-5191.

20. Huang Y, Carmichael GG: **Role of polyadenylation in nucleocytoplasmic transport of mRNA**. *Molecular and cellular biology* 1996, **16**(4):1534-1542.

21. Sachs A: **The role of poly(A) in the translation and stability of mRNA**. *Curr Opin Cell Biol* 1990, **2**(6):1092-1098.

22. Coller JM, Gray NK, Wickens MP: **mRNA stabilization by poly(A) binding protein is independent of poly(A) and requires translation**. *Genes Dev* 1998, **12**(20):3226-3235.

23. Guhaniyogi J, Brewer G: **Regulation of mRNA stability in mammalian cells**. *Gene* 2001, **265**(1-2):11-23.

24. Slomovic S, Fremder E, Staals RH, Pruijn GJ, Schuster G: **Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells**. *P Natl Acad Sci USA* 2010, **107**(16):7407-7412.

25. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R: **Comprehensive splice-site analysis using comparative genomics**. *Nucleic acids research* 2006, **34**(14):3955-3967.

26. Graveley BR: **Alternative splicing: increasing diversity in the proteomic world**. *Trends Genet* 2001, **17**(2):100-107.

27. Artamonova, II, Gelfand MS: **Comparative genomics and evolution of alternative splicing: the pessimists' science**. *Chem Rev* 2007, **107**(8):3407-3430.

28. Claverie JM: **Fewer genes, more noncoding RNA**. *Science* 2005, **309**(5740):1529-1530.

29. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

30. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.

31. Pennisi E: **Genomics. ENCODE project writes eulogy for junk DNA**. *Science* 2012, **337**(6099):1159, 1161.

32. Kim N, Alekseyenko AV, Roy M, Lee C: **The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species**. *Nucleic acids research* 2007, **35**(Database issue):D93-98.

33. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 2010, **463**(7280):457-463.

34. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

35. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

36. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D *et al*: **GENCODE: producing a reference annotation for ENCODE**. *Genome Biol* 2006, **7 Suppl 1**:S4 1-9.

37. Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH: **Genomics in C. elegans: so many genes, such a little worm**. *Genome research* 2005, **15**(12):1651-1660.

38. Russel PJ: **iGenetics: A Molecular Approach. 3rd edition**. In.: Pearson Education.(828 s). ISBN; 2010.

39. Wood EJ: **Cellular and molecular immunology (5th ed.): Abbas A. K., and Lichtman, A. H**. *Biochemistry and Molecular Biology Education* 2004, **32**(1):65-66.

40. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science* 1995, **269**(5223):496-512.

41. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al*: **The minimal gene complement of Mycoplasma genitalium**. *Science* 1995, **270**(5235):397-403.

42. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, 3rd *et al*: **Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence**. *Nature* 1998, **393**(6685):537-544.

43. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**(5287):546, 563-547.

44. Consortium TCeS: **Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology**. *Science* 1998, **282**(5396):2012-2018.

45. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA *et al*: **A whole-genome assembly of Drosophila**. *Science* 2000, **287**(5461):2196-2204.

46. **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.

47. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.

48. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE *et al*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. *Nature* 2004, **428**(6982):493-521.

49. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M *et al*: **Quality assessment of the human genome sequence**. *Nature* 2004, **429**(6990):365-368.

50. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems**. *J Biomed Biotechnol* 2012, **2012**:251364.

51. Collins FS, Morgan M, Patrinos A: **The Human Genome Project: lessons from large-scale biology**. *Science* 2003, **300**(5617):286-290.

52. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers**. *BMC Genomics* 2012, **13**:341.

53. Thompson JF, Steinmann KE: **Single molecule sequencing with a HeliScope genetic analysis system**. *Curr Protoc Mol Biol* 2010, **Chapter 7**:Unit7 10.

54. Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D *et al*: **Virtual terminator nucleotides for next-generation DNA sequencing**. *Nat Methods* 2009, **6**(8):593-595.

55. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M *et al*: **An integrated semiconductor device enabling non-optical genome sequencing**. *Nature* 2011, **475**(7356):348-352.

56. Karow J: **Ion Torrent Patent App Suggests Sequencing Tech Using Chemical-Sensitive Field-Effect Transistors [Internet]**. *In Science* 2009.

57. Davies K: **It's "Watson Meets Moore" as Ion Torrent Introduces Semiconductor Sequencing [Internet]**. *Bio-IT World* 2010.

58. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**(5910):133-138.

59. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: **Real-time DNA sequencing from single polymerase molecules**. *Methods in enzymology* 2010, **472**:431-455.

60. Church GB, MA), Deamer, David W. (Santa Cruz, CA), Branton, Daniel (Lexington, MA), Baldarelli, Richard (Natick, MA), Kasianowicz, John (Darnestown, MD): **Characterization of individual polymer molecules based on monomer-interface interactions**. In. United States: President & Fellows of Harvard College (Cambridge, MA),The Regents of the University of California (Oakland, CA); 1998.

61. Kasianowicz JJ, Brandin E, Branton D, Deamer DW: **Characterization of individual polynucleotide molecules using a membrane channel**. *P Natl Acad Sci USA* 1996, **93**(24):13770-13773.

62. Timp W, Mirsaidov UM, Wang D, Comer J, Aksimentiev A, Timp G: **Nanopore Sequencing: Electrical Measurements of the Code of Life**. *IEEE Trans Nanotechnol* 2010, **9**(3):281-294.

63. Check Hayden E: **Nanopore genome sequencer makes its debut [Internet]**. *Nature* 2012.

64. Schneider GF, Dekker C: **DNA sequencing with nanopores**. *Nature biotechnology* 2012, **30**(4):326-328.

65. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**(7422):56-65.

66. Imelfort M, Edwards D: **De novo sequencing of plant genomes using second-generation technologies**. *Brief Bioinform* 2009, **10**(6):609-618.

67. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing**. *Nucleic acids research* 2012, **40**(10):e72.

68. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC: **Effects of GC bias in next-generation-sequencing data on de novo genome assembly**. *PLoS One* 2013, **8**(4):e62856.

69. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing**. *Genome Res* 2010, **20**(9):1165-1173.

70. Chu HT, Hsiao WW, Tsao TT, Hsu DF, Chen CC, Lee SA, Kao CY: **SeqEntropy: genome-wide assessment of repeats for short read sequencing**. *PLoS One* 2013, **8**(3):e59484.

71. Wetzel J, Kingsford C, Pop M: **Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies**. *BMC Bioinformatics* 2011, **12**:95.

72. Prjibelski AD, Vasilinetc I, Bankevich A, Gurevich A, Krivosheeva T, Nurk S, Pham S, Korobeynikov A, Lapidus A, Pevzner PA: **ExSPAnder: a universal repeat resolver for DNA fragment assembly**. *Bioinformatics* 2014, **30**(12):i293-301.

73. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation**. *Nature reviews Genetics* 2012, **13**(5):329-342.

74. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.

75. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E *et al*: **Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence**. *Genome biology* 2002, **3**(12):RESEARCH0079.

76. **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

77. Windsor AJ, Mitchell-Olds T: **Comparative genomics as a tool for gene discovery**. *Curr Opin Biotechnol* 2006, **17**(2):161-167.

78. Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, Klein C, Martens L, Staes A, Timmerman E *et al*: **Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis**. *J Proteome Res* 2007, **6**(6):2195-2204.

79. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C, Lecompte O: **Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol**. *Genome Res* 2009, **19**(1):128-135.

80. Nielsen P, Krogh A: **Large-scale prokaryotic gene prediction and comparison to genome annotation**. *Bioinformatics* 2005, **21**(24):4322-4329.

81. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics**. *Proc Natl Acad Sci U S A* 2008, **105**(52):21034-21038.

82. Castellana NE, Shen Z, He Y, Walley JW, Cassidy CJ, Briggs SP, Bafna V: **An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays**. *Mol Cell Proteomics* 2014, **13**(1):157-167.

83. Chaerkady R, Kelkar DS, Muthusamy B, Kandasamy K, Dwivedi SB, Sahasrabuddhe NA, Kim MS, Renuse S, Pinto SM, Sharma R *et al*: **A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry**. *Genome research* 2011, **21**(11):1872-1881.

84. Borchert N, Dieterich C, Krug K, Schutz W, Jung S, Nordheim A, Sommer RJ, Macek B: **Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models**. *Genome Res* 2010, **20**(6):837-846.

85. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: **What is a gene, post-ENCODE? History and updated definition**. *Genome research* 2007, **17**(6):669-681.

86. Brent MR: **Steady progress and recent breakthroughs in the accuracy of automated genome annotation**. *Nature reviews Genetics* 2008, **9**(1):62-73.

87. Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD, Tatusova T: **Solving the Problem: Genome Annotation Standards before the Data Deluge**. *Stand Genomic Sci* 2011, **5**(1):168-193.

88. Brister JR, Bao Y, Kuiken C, Lefkowitz EJ, Le Mercier P, Leplae R, Madupu R, Scheuermann RH, Schobel S, Seto D *et al*: **Towards Viral Genome Annotation Standards, Report from the 2010 NCBI Annotation Workshop**. *Viruses* 2010, **2**(10):2258-2268.

89. Madupu R, Brinkac LM, Harrow J, Wilming LG, Bohme U, Lamesch P, Hannick LI: **Meeting report: a workshop on Best Practices in Genome Annotation**. *Database (Oxford)* 2010, **2010**:baq001.

90. Lu KH: **An analysis of the caries process by finite absorbing Markov chains**. *J Dent Res* 1966, **45**(4):998-1015.

91. Brent MR: **How does eukaryotic gene prediction work?** *Nature biotechnology* 2007, **25**(8):883-885.

92. Cortes C, Vapnik V: **Support-vector networks**. *Machine learning* 1995, **20**(3):273-297.

93. Guigo R, Knudsen S, Drake N, Smith T: **Prediction of gene structure**. *Journal of molecular biology* 1992, **226**(1):141-157.

94. Solovyev VV, Salamov AA, Lawrence CB: **The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:354-362.

95. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *Journal of molecular biology* 1997, **268**(1):78-94.

96. Reese MG, Kulp D, Tammana H, Haussler D: **Genie--gene finding in Drosophila melanogaster**. *Genome research* 2000, **10**(4):529-538.

97. Do CB, Woods DA, Batzoglou S: **CONTRAfold: RNA secondary structure prediction without physics-based models**. *Bioinformatics* 2006, **22**(14):e90-98.

98. Bernal A, Crammer K, Hatzigeorgiou A, Pereira F: **Global discriminative learning for higher-accuracy computational gene prediction**. *PLoS Comput Biol* 2007, **3**(3):e54.

99. DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE: **Conrad: gene prediction using conditional random fields**. *Genome research* 2007, **17**(9):1389-1398.

100. Gross SS, Do CB, Sirota M, Batzoglou S: **CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction**. *Genome biology* 2007, **8**(12):R269.

101. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19 Suppl 2**:ii215-225.

102. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources**. *BMC Bioinformatics* 2006, **7**:62.

103. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17 Suppl 1**:S140-148.

104. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA**. *Genome research* 2000, **10**(4):516-522.

105. Souvorov A, Kapustin Y, Kiryutin B, Chetvernin V, Tatusova T, Lipman D: **Gnomon-NCBI eukaryotic gene prediction tool**. *National Center for Biotechnology Information* 2010.

106. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59.

107. Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, Philips P, De Bona F, Hartmann L, Bohlen A *et al*: **mGene: accurate SVM-based gene finding with an application to nematode genomes**. *Genome research* 2009, **19**(11):2133-2143.

108. Allen JE, Salzberg SL: **JIGSAW: integration of multiple sources of evidence for gene prediction**. *Bioinformatics* 2005, **21**(18):3596-3603.

109. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments**. *Genome biology* 2008, **9**(1):R7.

110. Howe KL, Chothia T, Durbin R: **GAZE: a generic framework for the integration of gene-prediction data by dynamic programming**. *Genome research* 2002, **12**(9):1418-1427.

111. Coghlan A, Durbin R: **Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure**. *Bioinformatics* 2007, **23**(12):1468-1475.

112. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set**. *Genome biology* 2007, **8**(1):R13.

113. Liu Q, Mackey AJ, Roos DS, Pereira FC: **Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction**. *Bioinformatics* 2008, **24**(5):597-605.

114. Issac B, Raghava GP: **EGPred: prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches**. *Genome Res* 2004, **14**(9):1756-1766.

115. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence**. *Genome Res* 1998, **8**(9):967-974.

116. Mott R: **EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA**. *Comput Appl Biosci* 1997, **13**(4):477-478.

117. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment**. *P Natl Acad Sci USA* 1996, **93**(17):9061-9066.

118. Wheelan SJ, Church DM, Ostell JM: **Spidey: a tool for mRNA-to-genomic alignments**. *Genome Res* 2001, **11**(11):1952-1957.

119. Usuka J, Zhu W, Brendel V: **Optimal spliced alignment of homologous cDNA to a genomic DNA template**. *Bioinformatics* 2000, **16**(3):203-211.

120. Rinner O, Morgenstern B: **AGenDA: gene prediction by comparative sequence analysis**. *In Silico Biol* 2002, **2**(3):195-205.

121. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

122. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.

123. Birney E, Durbin R: **Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:56-64.

124. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome research* 2004, **14**(5):988-995.

125. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC Bioinformatics* 2005, **6**:31.

126. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biology* 2013, **14**(4):R36.

127. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

128. Tchourbanov A, Adviser-Ali H, Adviser-Deogun J: **Signal based Bayesian framework for gene structural prediction**: University of Nebraska at Lincoln; 2006.

129. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics* 2005, **21**(9):1859-1875.

130. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinformatics* 2010, **26**(7):873-881.

131. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S *et al*: **The Vertebrate Genome Annotation (Vega) database**. *Nucleic acids research* 2005, **33**(Database issue):D459-465.

132. Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K, Spannagl M, Mayer KF, Stumpflen V *et al*: **MIPS: curated databases and comprehensive secondary data resources in 2010**. *Nucleic acids research* 2011, **39**(Database issue):D220-224.

133. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I *et al*: **The integrated microbial genomes (IMG) system**. *Nucleic Acids Res* 2006, **34**(Database issue):D344-348.

134. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M *et al*: **The RAST Server: rapid annotations using subsystems technology**. *BMC Genomics* 2008, **9**:75.

135. Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, Theil S, Reboux S, Amano N, Flutre T, Pelegrin C *et al*: **TriAnnot: A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes**. *Front Plant Sci* 2012, **3**:5.

136. Seaver SM, Gerdes S, Frelin O, Lerma-Ortiz C, Bradbury LM, Zallot R, Hasnain G, Niehaus TD, El Yacoubi B, Pasternak S *et al*: **High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource**. *P Natl Acad Sci USA* 2014, **111**(26):9645-9650.

137. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects**. *BMC Bioinformatics* 2011, **12**:491.

138. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.

139. Koski LB, Gray MW, Lang BF, Burger G: **AutoFACT: an automatic functional annotation and classification tool**. *BMC Bioinformatics* 2005, **6**:151.

140. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucleic acids research* 2003, **31**(19):5654-5666.

141. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system**. *Genome Res* 2004, **14**(5):942-950.

142. Kitts P: **The NCBI handbook**. In. Edited by McEntyre J, Ostell J, 2nd edn; 2002.

143. **HAVANA** [http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/]

144. Poptsova MS, Gogarten JP: **Using comparative genome analysis to identify problems in annotated microbial genomes**. *Microbiology* 2010, **156**(Pt 7):1909-1917.

145. Mann M, Kulak NA, Nagaraj N, Cox J: **The coming age of complete, accurate, and ubiquitous proteomes**. *Molecular cell* 2013, **49**(4):583-590.

146. de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, Mann M: **Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system**. *Genome biology* 2006, **7**(6):R50.

147. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.

148. Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R: **Generating and navigating proteome maps using mass spectrometry**. *Nat Rev Mol Cell Biol* 2010, **11**(11):789-801.

149. Gstaiger M, Aebersold R: **Applying mass spectrometry-based proteomics to genetics, genomics and network biology**. *Nat Rev Genet* 2009, **10**(9):617-627.

150. Patterson SD, Aebersold RH: **Proteomics: the first decade and beyond**. *Nat Genet* 2003, **33 Suppl**:311-323.

151. Washburn MP, Wolters D, Yates JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology**. *Nature biotechnology* 2001, **19**(3):242-247.

152. Douglas DJ, Frank AJ, Mao D: **Linear ion traps in mass spectrometry**. *Mass Spectrom Rev* 2005, **24**(1):1-29.

153. Elias JE, Haas W, Faherty BK, Gygi SP: **Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations**. *Nat Methods* 2005, **2**(9):667-675.

154. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome**. *J Proteome Res* 2003, **2**(1):43-50.

155. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R: **The Orbitrap: a new mass spectrometer**. *J Mass Spectrom* 2005, **40**(4):430-443.

156. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF: **Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry**. *Proc Natl Acad Sci U S A* 2004, **101**(26):9528-9533.

157. Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP: **Optimization and use of peptide mass measurement accuracy in shotgun proteomics**. *Mol Cell Proteomics* 2006, **5**(7):1326-1337.

158. Chernushevich IV, Loboda AV, Thomson BA: **An introduction to quadrupole-time-of-flight mass spectrometry**. *Journal of mass spectrometry : JMS* 2001, **36**(8):849-865.

159. Heeren RM, Kleinnijenhuis AJ, McDonnell LA, Mize TH: **A mini-review of mass spectrometry using high-performance FTICR-MS methods**. *Anal Bioanal Chem* 2004, **378**(4):1048-1058.

160. Clauser KR, Baker P, Burlingame AL: **Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching**. *Analytical chemistry* 1999, **71**(14):2871-2882.

161. Zubarev R, Mann M: **On the proper use of mass accuracy in proteomics**. *Mol Cell Proteomics* 2007, **6**(3):377-381.

162. O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins**. *J Biol Chem* 1975, **250**(10):4007-4021.

163. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR: **Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides**. *Analytical chemistry* 2003, **75**(5):1155-1163.

164. Pappin DJ, Hojrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass fingerprinting**. *Curr Biol* 1993, **3**(6):327-332.

165. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**(18):3551-3567.

166. Clauser KR, Hall SC, Smith DM, Webb JW, Andrews LE, Tran HM, Epstein LB, Burlingame AL: **Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE**. *P Natl Acad Sci USA* 1995, **92**(11):5072-5076.

167. Wilkins M, Gasteiger E, Sanchez J, Bairoch A, Appel R, Dunn M, Hochstrasser D: **Proteomics and two-dimensional electrophoresis-Multiple parameter cross-species protein identification using Multident-A world-wide web accessible tool**. *Electrophoresis-An International Journal* 1998, **19**(18):3199-3206.

168. Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information**. *Analytical chemistry* 2000, **72**(11):2482-2489.

169. Tuloup M, Hernandez C, Coro I, Hoogland C, Binz P, Appel R: **Aldente and BioGraph: An improved peptide mass fingerprinting protein identification environment**. In: *Proceedings of the Swiss Proteomics Society 2003 Congress: Understanding Biological Systems through Proteomics: 2003*; 2003: 174-176.

170. Schibeci D, Potter R, Wathen-Dunn K, Jones M, Bellgard M: **Applying artificial neural networks to the classification of wheat varieties processed via MALDI-TOF mass spectrometry**. 2001.

171. Sørensen HA, Sperotto MM, Petersen M, Keşmir C, Radzikowski L, Jacobsen S, Søndergaard I: **Variety identification of wheat using mass spectrometry with neural networks and the influence of mass spectra processing prior to neural network analysis**. *Rapid Commun Mass Sp* 2002, **16**(12):1232-1237.

172. Bellgard M, Taplin R, Chapman B, Livk A, Wellington C, Hunter A, Lipscombe R: **Classification of fish samples via an integrated proteomics and bioinformatics approach**. *Proteomics* 2013, **13**(21):3124-3130.

173. Liotta LA, Ferrari M, Petricoin E: **Clinical proteomics: written in blood**. *Nature* 2003, **425**(6961):905.

174. Boersema PJ, Geiger T, Wisniewski JR, Mann M: **Quantification of the N-glycosylated secretome by super-SILAC during breast cancer progression and in human blood samples**. *Molecular & cellular proteomics : MCP* 2013, **12**(1):158-171.

175. Wisniewski JR, Dus K, Mann M: **Proteomic workflow for analysis of archival formalin-fixed and paraffin-embedded clinical samples to a depth of 10 000 proteins**. *Proteomics Clin Appl* 2013, **7**(3-4):225-233.

176. Zanivan S, Meves A, Behrendt K, Schoof EM, Neilson LJ, Cox J, Tang HR, Kalna G, van Ree JH, van Deursen JM *et al*: **In vivo SILAC-based proteomics reveals phosphoproteome changes during mouse skin carcinogenesis**. *Cell Rep* 2013, **3**(2):552-566.

177. Domon B, Aebersold R: **Mass spectrometry and protein analysis**. *Science* 2006, **312**(5771):212-217.

178. Biemann K: **Contributions of mass spectrometry to peptide and protein structure**. *Biomed Environ Mass Spectrom* 1988, **16**(1-12):99-111.

179. Roepstorff P, Fohlman J: **Proposal for a common nomenclature for sequence ions in mass spectra of peptides**. *Biomed Mass Spectrom* 1984, **11**(11):601.

180. Johnson RS, Martin SA, Biemann K: **Collision-Induced Fragmentation of (M+H)+Ions of Peptides - Side-Chain Specific Sequence Ions**. *International Journal of Mass Spectrometry and Ion Processes* 1988, **86**:137-154.

181. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA: **De novo peptide sequencing via tandem mass spectrometry**. *J Comput Biol* 1999, **6**(3-4):327-342.

182. Taylor JA, Johnson RS: **Sequence database searches via de novo peptide sequencing by tandem mass spectrometry**. *Rapid Commun Mass Sp* 1997, **11**(9):1067-1075.

183. Edman P: **Method for determination of the amino acid sequence in peptides**. *Acta chem scand* 1950, **4**(283-293):7.

184. Opiteck GJ, Jorgenson JW: **Two-dimensional SEC/RPLC coupled to mass spectrometry for the analysis of peptides**. *Analytical chemistry* 1997, **69**(13):2283-2291.

185. Opiteck GJ, Ramirez SM, Jorgenson JW, Moseley MA, 3rd: **Comprehensive two-dimensional high-performance liquid chromatography for the isolation of overexpressed proteins and proteome mapping**. *Analytical biochemistry* 1998, **258**(2):349-361.

186. Figeys D, Ducret A, Yates JR, 3rd, Aebersold R: **Protein identification by solid phase microextraction-capillary zone electrophoresis-microelectrospray-tandem mass spectrometry**. *Nature biotechnology* 1996, **14**(11):1579-1583.

187. Tong W, Link A, Eng JK, Yates JR, 3rd: **Identification of proteins in complexes by solid-phase microextraction/multistep elution/capillary electrophoresis/tandem mass spectrometry**. *Analytical chemistry* 1999, **71**(13):2270-2278.

188. Chen J, Balgley BM, DeVoe DL, Lee CS: **Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis**. *Analytical chemistry* 2003, **75**(13):3145-3152.

189. Cargile BJ, Bundy JL, Freeman TW, Stephenson JL, Jr.: **Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification**. *Journal of proteome research* 2004, **3**(1):112-119.

190. Spahr CS, Davis MT, McGinley MD, Robinson JH, Bures EJ, Beierle J, Mort J, Courchesne PL, Chen K, Wahl RC *et al*: **Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest**. *Proteomics* 2001, **1**(1):93-107.

191. Yi EC, Marelli M, Lee H, Purvine SO, Aebersold R, Aitchison JD, Goodlett DR: **Approaching complete peroxisome characterization by gas-phase fractionation**. *Electrophoresis* 2002, **23**(18):3205-3216.

192. Blonder J, Hale ML, Lucas DA, Schaefer CF, Yu LR, Conrads TP, Issaq HJ, Stiles BG, Veenstra TD: **Proteomic analysis of detergent-resistant membrane rafts**. *Electrophoresis* 2004, **25**(9):1307-1318.

193. Blonder J, Rodriguez-Galan MC, Lucas DA, Young HA, Issaq HJ, Veenstra TD, Conrads TP: **Proteomic investigation of natural killer cell microsomes using gas-phase fractionation by mass spectrometry**. *Biochimica et biophysica acta* 2004, **1698**(1):87-95.

194. Guina T, Wu M, Miller SI, Purvine SO, Yi EC, Eng J, Goodlett DR, Aebersold R, Ernst RK, Lee KA: **Proteomic analysis of Pseudomonas aeruginosa grown under magnesium limitation**. *J Am Soc Mass Spectr* 2003, **14**(7):742-751.

195. Utleg AG, Yi EC, Xie T, Shannon P, White JT, Goodlett DR, Hood L, Lin B: **Proteomic analysis of human prostasomes**. *Prostate* 2003, **56**(2):150-161.

196. Olsen JV, Ong SE, Mann M: **Trypsin cleaves exclusively C-terminal to arginine and lysine residues**. *Molecular & cellular proteomics : MCP* 2004, **3**(6):608-614.

197. Rodriguez J, Gupta N, Smith RD, Pevzner PA: **Does trypsin cut before proline?** *Journal of proteome research* 2008, **7**(1):300-305.

198. Thelen JJ, Miernyk JA: **The proteomic future: where mass spectrometry should be taking us**. *Biochem J* 2012, **444**(2):169-181.

199. Altelaar AF, Mohammed S, Brans MA, Adan RA, Heck AJ: **Improved identification of endogenous peptides from murine nervous tissue by multiplexed peptide extraction methods and multiplexed mass spectrometric analysis**. *J Proteome Res* 2009, **8**(2):870-876.

200. Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JE, Coon JJ: **Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors**. *Anal Chem* 2007, **79**(2):477-485.

201. Good DM, Wirtala M, McAlister GC, Coon JJ: **Performance characteristics of electron transfer dissociation mass spectrometry**. *Mol Cell Proteomics* 2007, **6**(11):1942-1951.

202. Toorn HWPvd: **Targeted SCX Based Peptide Fractionation for Optimal Sequencing by Collision Induced, and Electron Transfer Dissociation**. *Journal of proteomics & bioinformatics* 2008, **01**(08):379-388.

203. Mikesh LM, Ueberheide B, Chi A, Coon JJ, Syka JE, Shabanowitz J, Hunt DF: **The utility of ETD mass spectrometry in proteomic analysis**. *Biochim Biophys Acta* 2006, **1764**(12):1811-1822.

204. Wiesner J, Premsler T, Sickmann A: **Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications**. *Proteomics* 2008, **8**(21):4466-4483.

205. Boersema PJ, Mohammed S, Heck AJ: **Phosphopeptide fragmentation and analysis by mass spectrometry**. *J Mass Spectrom* 2009, **44**(6):861-878.

206. Swaney DL, McAlister GC, Coon JJ: **Decision tree-driven tandem mass spectrometry for shotgun proteomics**. *Nat Methods* 2008, **5**(11):959-964.

207. Zubarev RA, Zubarev AR, Savitski MM: **Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet?** *J Am Soc Mass Spectr* 2008, **19**(6):753-761.

208. Frese CK, Altelaar AF, Hennrich ML, Nolting D, Zeller M, Griep-Raming J, Heck AJ, Mohammed S: **Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos**. *J Proteome Res* 2011, **10**(5):2377-2388.

209. Molina H, Matthiesen R, Kandasamy K, Pandey A: **Comprehensive comparison of collision induced dissociation and electron transfer dissociation**. *Anal Chem* 2008, **80**(13):4825-4835.

210. Sobott F, Watt SJ, Smith J, Edelmann MJ, Kramer HB, Kessler BM: **Comparison of CID versus ETD based MS/MS fragmentation for the analysis of protein ubiquitination**. *J Am Soc Mass Spectrom* 2009, **20**(9):1652-1659.

211. Scott NE, Parker BL, Connolly AM, Paulech J, Edwards AV, Crossett B, Falconer L, Kolarich D, Djordjevic SP, Hojrup P *et al*: **Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of Campylobacter jejuni**. *Mol Cell Proteomics* 2011, **10**(2):M000031-MCP000201.

212. Wenner BR, Lynn BC: **Factors that affect ion trap data-dependent MS/MS in proteomics**. *J Am Soc Mass Spectr* 2004, **15**(2):150-157.

213. Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM: **Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks**. *P Natl Acad Sci USA* 2007, **104**(14):5860-5865.

214. Wu L, Han DK: **Overcoming the dynamic range problem in mass spectrometry-based shotgun proteomics**. *Expert Rev Proteomics* 2006, **3**(6):611-619.

215. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM: **Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies**. *Journal of proteome research* 2010, **9**(8):4152-4160.

216. Luethy R, Kessner DE, Katz JE, Maclean B, Grothe R, Kani K, Faca V, Pitteri S, Hanash S, Agus DB *et al*: **Precursor-ion mass re-estimation improves peptide identification on hybrid instruments**. *Journal of proteome research* 2008, **7**(9):4031-4039.

217. Simpson RJ, Connolly LM, Eddes JS, Pereira JJ, Moritz RL, Reid GE: **Proteomic analysis of the human colon carcinoma cell line (LIM 1215): development of a membrane protein database**. *Electrophoresis* 2000, **21**(9):1707-1732.

218. Nesvizhskii AI, Aebersold R: **Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS**. *Drug Discov Today* 2004, **9**(4):173-181.

219. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: **A statistical model for identifying proteins by tandem mass spectrometry**. *Analytical chemistry* 2003, **75**(17):4646-4658.

220. Nesvizhskii AI, Aebersold R: **Interpretation of shotgun proteomic data: the protein inference problem**. *Molecular & cellular proteomics : MCP* 2005, **4**(10):1419-1440.

221. Loo JA, Edmonds CG, Smith RD: **Primary sequence information from intact proteins by electrospray ionization tandem mass spectrometry**. *Science* 1990, **248**(4952):201-204.

222. Reid GE, McLuckey SA: **'Top down' protein characterization via tandem mass spectrometry**. *Journal of mass spectrometry : JMS* 2002, **37**(7):663-675.

223. Sze SK, Ge Y, Oh H, McLafferty FW: **Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue**. *P Natl Acad Sci USA* 2002, **99**(4):1774-1779.

224. Dorrestein PC, Zhai H, Taylor SV, McLafferty FW, Begley TP: **The biosynthesis of the thiazole phosphate moiety of thiamin (vitamin B1): the early steps catalyzed by thiazole synthase**. *Journal of the American Chemical Society* 2004, **126**(10):3091-3096.

225. Whitelegge J, Halgand F, Souda P, Zabrouskov V: **Top-down mass spectrometry of integral membrane proteins**. *Expert Rev Proteomics* 2006, **3**(6):585-596.

226. Dorrestein PC, Van Lanen SG, Li W, Zhao C, Deng Z, Shen B, Kelleher NL: **The bifunctional glyceryl transferase/phosphatase OzmB belonging to the HAD superfamily that diverts 1,3-bisphosphoglycerate into polyketide biosynthesis**. *Journal of the American Chemical Society* 2006, **128**(32):10386-10387.

227. McLafferty FW, Breuker K, Jin M, Han X, Infusini G, Jiang H, Kong X, Begley TP: **Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics**. *Febs J* 2007, **274**(24):6256-6268.

228. Siuti N, Kelleher NL: **Decoding protein modifications using top-down mass spectrometry**. *Nature methods* 2007, **4**(10):817-821.

229. Whitelegge JP, Zabrouskov V, Halgand F, Souda P, Bassilian S, Yan W, Wolinsky L, Loo JA, Wong DT, Faull KF: **Protein-Sequence Polymorphisms and Post-translational Modifications in Proteins from Human Saliva using Top-Down Fourier-transform Ion Cyclotron Resonance Mass Spectrometry**. *Int J Mass Spectrom* 2007, **268**(2-3):190-197.

230. Zabrouskov V, Whitelegge JP: **Increased coverage in the transmembrane domain with activated-ion electron capture dissociation for top-down Fourier-transform mass spectrometry of integral membrane proteins**. *Journal of proteome research* 2007, **6**(6):2205-2210.

231. Parks BA, Jiang L, Thomas PM, Wenger CD, Roth MJ, Boyne MT, 2nd, Burke PV, Kwast KE, Kelleher NL: **Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers**. *Analytical chemistry* 2007, **79**(21):7984-7991.

232. Roth MJ, Parks BA, Ferguson JT, Boyne MT, 2nd, Kelleher NL: **"Proteotyping": population proteomics of human leukocytes using top down mass spectrometry**. *Analytical chemistry* 2008, **80**(8):2857-2866.

233. Garcia BA: **What does the future hold for Top Down mass spectrometry?** *J Am Soc Mass Spectr* 2010, **21**(2):193-202.

234. **Bottom-up proteomics** [http://en.wikipedia.org/wiki/Bottom-up_proteomics]

235. Meng F, Cargile BJ, Patrie SM, Johnson JR, McLoughlin SM, Kelleher NL: **Processing complex mixtures of intact proteins for direct analysis by mass spectrometry**. *Anal Chem* 2002, **74**(13):2923-2929.

236. Meng F, Du Y, Miller LM, Patrie SM, Robinson DE, Kelleher NL: **Molecular-level description of proteins from saccharomyces cerevisiae using quadrupole FT hybrid mass spectrometry for top down proteomics**. *Anal Chem* 2004, **76**(10):2852-2858.

237. Patrie SM, Ferguson JT, Robinson DE, Whipple D, Rother M, Metcalf WW, Kelleher NL: **Top down mass spectrometry of < 60-kDa proteins from Methanosarcina acetivorans using quadrupole FRMS with automated octopole collisionally activated dissociation**. *Mol Cell Proteomics* 2006, **5**(1):14-25.

238. Sharma S, Simpson DC, Tolic N, Jaitly N, Mayampurath AM, Smith RD, Pasa-Tolic L: **Proteomic profiling of intact proteins using WAX-RPLC 2-D separations and FTICR mass spectrometry**. *J Proteome Res* 2007, **6**(2):602-610.

239. Han X, Jin M, Breuker K, McLafferty FW: **Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons**. *Science* 2006, **314**(5796):109-112.

240. Shen Y, Hixson KK, Tolic N, Camp DG, Purvine SO, Moore RJ, Smith RD: **Mass spectrometry analysis of proteome-wide proteolytic post-translational degradation of proteins**. *Anal Chem* 2008, **80**(15):5819-5828.

241. Shen Y, Tolic N, Hixson KK, Purvine SO, Anderson GA, Smith RD: **De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins**. *Anal Chem* 2008, **80**(20):7742-7754.

242. Wynne C, Fenselau C, Demirev PA, Edwards N: **Top-down identification of protein biomarkers in bacteria with unsequenced genomes**. *Anal Chem* 2009, **81**(23):9633-9642.

243. Tsai YS, Scherl A, Shaw JL, MacKay CL, Shaffer SA, Langridge-Smith PR, Goodlett DR: **Precursor ion independent algorithm for top-down shotgun proteomics**. *J Am Soc Mass Spectrom* 2009, **20**(11):2154-2166.

244. Vellaichamy A, Tran JC, Catherman AD, Lee JE, Kellie JF, Sweet SM, Zamdborg L, Thomas PM, Ahlf DR, Durbin KR *et al*: **Size-sorting combined with improved nanocapillary liquid chromatography-mass spectrometry for identification of intact proteins up to 80 kDa**. *Anal Chem* 2010, **82**(4):1234-1244.

245. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M *et al*: **Mapping intact protein isoforms in discovery mode using top-down proteomics**. *Nature* 2011, **480**(7376):254-258.

246. Wu C, Tran JC, Zamdborg L, Durbin KR, Li M, Ahlf DR, Early BP, Thomas PM, Sweedler JV, Kelleher NL: **A protease for 'middle-down' proteomics**. *Nature methods* 2012, **9**(8):822-824.

247. Savaryn JP, Catherman AD, Thomas PM, Abecassis MM, Kelleher NL: **The emergence of top-down proteomics in clinical research**. *Genome Med* 2013, **5**(6):53.

248. Giglione C, Boularot A, Meinnel T: **Protein N-terminal methionine excision**. *Cell Mol Life Sci* 2004, **61**(12):1455-1474.

249. Compton PD, Kelleher NL: **Spinning up mass spectrometry for whole protein complexes**. *Nature methods* 2012, **9**(11):1065-1066.

250. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR: **Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra**. *Nature methods* 2004, **1**(1):39-45.

251. Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ: **Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry**. *Analytical chemistry* 2010, **82**(3):833-841.

252. Carvalho PC, Han X, Xu T, Cociorva D, Carvalho Mda G, Barbosa VC, Yates JR, 3rd: **XDIA: improving on the label-free data-independent analysis**. *Bioinformatics* 2010, **26**(6):847-848.

253. Blackburn K, Mbeunkui F, Mitra SK, Mentzel T, Goshe MB: **Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation**. *Journal of proteome research* 2010, **9**(7):3621-3637.

254. Bateman RH, Carruthers R, Hoyes JB, Jones C, Langridge JI, Millar A, Vissers JP: **A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation**. *J Am Soc Mass Spectr* 2002, **13**(7):792-803.

255. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ: **Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition**. *Molecular & cellular proteomics : MCP* 2006, **5**(1):144-156.

256. Silva JC, Denny R, Dorschel C, Gorenstein MV, Li GZ, Richardson K, Wall D, Geromanos SJ: **Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale**. *Molecular & cellular proteomics : MCP* 2006, **5**(4):589-607.

257. Purvine S, Eppel JT, Yi EC, Goodlett DR: **Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer**. *Proteomics* 2003, **3**(6):847-850.

258. Niggeweg R, Kocher T, Gentzel M, Buscaino A, Taipale M, Akhtar A, Wilm M: **A general precursor ion-like scanning mode on quadrupole-TOF instruments compatible with chromatographic separation**. *Proteomics* 2006, **6**(1):41-53.

259. Geiger T, Cox J, Mann M: **Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation**. *Molecular & cellular proteomics : MCP* 2010, **9**(10):2252-2261.

260. Weisbrod CR, Eng JK, Hoopmann MR, Baker T, Bruce JE: **Accurate peptide fragment mass analysis: Multiplexed peptide identification and quantification**. *Journal of proteome research* 2012, **11**(3):1621-1632.

261. Andrews GL, Simons BL, Young JB, Hawkridge AM, Muddiman DC: **Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600)**. *Analytical chemistry* 2011, **83**(13):5442-5446.

262. Liu Y, Huttenhain R, Surinova S, Gillet LC, Mouritsen J, Brunner R, Navarro P, Aebersold R: **Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS**. *Proteomics* 2013, **13**(8):1247-1256.

263. Held JM, Schilling B, D'Souza AK, Srinivasan T, Behring JB, Sorensen DJ, Benz CC, Gibson BW: **Label-Free Quantitation and Mapping of the ErbB2 Tumor Receptor by Multiple Protease Digestion with Data-Dependent (MS1) and Data-Independent (MS2) Acquisitions**. *Int J Proteomics* 2013, **2013**:791985.

264. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R: **Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis**. *Molecular & cellular proteomics : MCP* 2012, **11**(6):O111 016717.

265. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V *et al*: **Multiplexed MS/MS for improved data-independent acquisition**. *Nature methods* 2013, **10**(8):744-746.

266. Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI, Goodlett DR: **Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean**. *Analytical chemistry* 2009, **81**(15):6481-6488.

267. Panchaud A, Jung S, Shaffer SA, Aitchison JD, Goodlett DR: **Faster, quantitative, and accurate precursor acquisition independent from ion count**. *Analytical chemistry* 2011, **83**(6):2250-2257.

268. Law KP, Lim YP: **Recent advances in mass spectrometry: data independent analysis and hyper reaction monitoring**. *Expert Rev Proteomics* 2013, **10**(6):551-566.

269. Chapman JD, Goodlett DR, Masselon CD: **Multiplexed and data-independent tandem mass spectrometry for global proteome profiling**. *Mass spectrometry reviews* 2014, **33**(6):452-470.

270. Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R: **ProbIDtree: an automated software program capable of identifying multiple peptides from a**

single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* 2005, **5**(16):4096-4106.

271. Wang J, Perez-Santiago J, Katz JE, Mallick P, Bandeira N: **Peptide identification from mixture tandem mass spectra**. *Molecular & cellular proteomics : MCP* 2010, **9**(7):1476-1485.

272. Wang J, Bourne PE, Bandeira N: **Peptide identification by database search of mixture tandem mass spectra**. *Molecular & cellular proteomics : MCP* 2011, **10**(12):M111 010017.

273. Wang J, Bourne PE, Bandeira N: **MixGF: spectral probabilities for mixture spectra from more than one peptide**. *Mol Cell Proteomics* 2014.

274. Mcdonald RS, Wilks PA: **Jcamp-Dx - a Standard Form for Exchange of Infrared-Spectra in Computer Readable Form**. *Appl Spectrosc* 1988, **42**(1):151-162.

275. **ASTM E1947 - 98(2014) Standard Specification for Analytical Data Interchange Protocol for Chromatographic Data** [http://www.astm.org/Standards/E1947.htm]

276. **Network Common Data Form (NetCDF)** [http://www.unidata.ucar.edu/software/netcdf/docs/index.html]

277. Orchard S, Montechi-Palazzi L, Deutsch EW, Binz PA, Jones AR, Paton N, Pizarro A, Creasy DM, Wojcik J, Hermjakob H: **Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23-25, 2007 Ecole Nationale Superieure (ENS), Lyon, France**. *Proteomics* 2007, **7**(19):3436-3440.

278. **HUPO-PSI mzData** [http://www.psidev.info/mzdata]

279. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R *et al*: **A common open representation of mass spectrometry data and its application to proteomics research**. *Nat Biotechnol* 2004, **22**(11):1459-1466.

280. Lin SM, Zhu L, Winter AQ, Sasinowski M, Kibbe WA: **What is mzXML good for?** *Expert Rev Proteomics* 2005, **2**(6):839-845.

281. Deutsch E: **mzML: a single, unifying data format for mass spectrometer output**. *Proteomics* 2008, **8**(14):2776-2777.

282. **HUPO-PSI mzML** [http://www.psidev.info/mzml]

283. McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, Johnson JR, Cociorva D, Yates JR, 3rd: **MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications**. *Rapid communications in mass spectrometry : RCM* 2004, **18**(18):2162-2168.

284. **ReAdW** [http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW]

285. **mzWIFF** [http://tools.proteomecenter.org/wiki/index.php?title=Software:mzWiff]

286. **msConvert** [http://proteowizard.sourceforge.net/tools/msconvert.html]

287. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J *et al*: **A cross-platform toolkit for mass spectrometry and proteomics**. *Nature biotechnology* 2012, **30**(10):918-920.

288. Pedrioli PG: **Trans-proteomic pipeline: a pipeline for proteomic analysis**. *Methods Mol Biol* 2010, **604**:213-238.

289. Griss J, Reisinger F, Hermjakob H, Vizcaino JA: **jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats**. *Proteomics* 2012, **12**(6):795-798.

290. Cote RG, Reisinger F, Martens L: **jmzML, an open-source Java API for mzML, the PSI standard for MS data**. *Proteomics* 2010, **10**(7):1332-1335.

291. Reisinger F, Krishna R, Ghali F, Rios D, Hermjakob H, Vizcaino JA, Jones AR: **jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data**. *Proteomics* 2012, **12**(6):790-794.

292. Kim S, Adviser-Pevzner PA: **Generating functions of tandem mass spectra and their applications for peptide identifications**: University of California at San Diego; 2012.

293. Zhang Z, Marshall AG: **A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra**. *J Am Soc Mass Spectr* 1998, **9**(3):225-233.

294. Du P, Angeletti RH: **Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution**. *Analytical chemistry* 2006, **78**(10):3385-3392.

295. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA: **Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach**. *Molecular & cellular proteomics : MCP* 2010, **9**(12):2772-2782.

296. Slawski M, Hussong R, Tholey A, Jakoby T, Gregorius B, Hildebrandt A, Hein M: **Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching**. *BMC Bioinformatics* 2012, **13**:291.

297. Rejtar T, Chen HS, Andreev V, Moskovets E, Karger BL: **Increased identification of peptides by enhanced data processing of high-resolution MALDI TOF/TOF mass spectra prior to database searching**. *Analytical chemistry* 2004, **76**(20):6017-6028.

298. Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt A: **High-accuracy peak picking of proteomics data using wavelet techniques**. *Pac Symp Biocomput* 2006:243-254.

299. Coombes KR, Baggerly KA, Morris JS: **Pre-processing mass spectrometry data**. In: *Fundamentals of Data Mining in Genomics and Proteomics.* Springer; 2007: 79-102.

300. Zhang J, He S, Ling CX, Cao X, Zeng R, Gao W: **PeakSelect: preprocessing tandem mass spectra for better peptide identification**. *Rapid communications in mass spectrometry : RCM* 2008, **22**(8):1203-1212.

301. Bern M, Goldberg D, McDonald WH, Yates JR, 3rd: **Automatic quality assessment of peptide tandem mass spectra**. *Bioinformatics* 2004, **20 Suppl 1**:i49-54.

302. Frank AM: **A ranking-based scoring function for peptide-spectrum matches**. *J Proteome Res* 2009, **8**(5):2241-2252.

303. Flikka K, Martens L, Vandekerckhoe J, Gevaert K, Eidhammer I: **Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering**. *Proteomics* 2006, **6**(7):2086-2094.

304. Wu FX, Gagne P, Droit A, Poirier GG: **Quality assessment of peptide tandem mass spectra**. *BMC Bioinformatics* 2008, **9 Suppl 6**:S13.

305. Junqueira M, Spirin V, Santana Balbuena T, Waridel P, Surendranath V, Kryukov G, Adzhubei I, Thomas H, Sunyaev S, Shevchenko A: **Separating the wheat from the chaff: unbiased filtering of background tandem mass spectra improves protein identification**. *Journal of proteome research* 2008, **7**(8):3382-3395.

306. Ma ZQ, Chambers MC, Ham AJ, Cheek KL, Whitwell CW, Aerni HR, Schilling B, Miller AW, Caprioli RM, Tabb DL: **ScanRanker: Quality assessment of tandem mass spectra via sequence tagging**. *Journal of proteome research* 2011, **10**(7):2896-2904.

307. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR, 3rd: **Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility**. *Analytical chemistry* 2003, **75**(10):2470-2477.

308. Tabb DL, Thompson MR, Khalsa-Moyers G, VerBerkmoes NC, McDonald WH: **MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra**. *J Am Soc Mass Spectr* 2005, **16**(8):1250-1261.

309. Beer I, Barnea E, Ziv T, Admon A: **Improving large-scale proteomics by clustering of mass spectrometry data**. *Proteomics* 2004, **4**(4):950-960.

310. Ramakrishnan SR, Mao R, Nakorchevskiy AA, Prince JT, Willard WS, Xu W, Marcotte EM, Miranker DP: **A fast coarse filtering method for peptide identification by mass spectrometry**. *Bioinformatics* 2006, **22**(12):1524-1531.

311. Dutta D, Chen T: **Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search**. *Bioinformatics* 2007, **23**(5):612-618.

312. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, Pevzner PA: **Clustering millions of tandem mass spectra**. *Journal of proteome research* 2008, **7**(1):113-122.

313. Mann M, Wilm M: **Error-tolerant identification of peptides in sequence databases by peptide sequence tags**. *Anal Chem* 1994, **66**(24):4390-4399.

314. Frank A, Pevzner P: **PepNovo: de novo peptide sequencing via probabilistic network modeling**. *Anal Chem* 2005, **77**(4):964-973.

315. Mo L, Dutta D, Wan Y, Chen T: **MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry**. *Anal Chem* 2007, **79**(13):4870-4878.

316. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G: **PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 2003, **17**(20):2337-2342.

317. Chi H, Sun RX, Yang B, Song CQ, Wang LH, Liu C, Fu Y, Yuan ZF, Wang HP, He SM *et al*: **pNovo: de novo peptide sequencing and identification using HCD spectra**. *Journal of proteome research* 2010, **9**(5):2713-2724.

318. Datta R, Bern M: **Spectrum fusion: using multiple mass spectra for de novo Peptide sequencing**. *Journal of computational biology : a journal of computational molecular cell biology* 2009, **16**(8):1169-1182.

319. He L, Ma B: **ADEPTS: advanced peptide de novo sequencing with a pair of tandem mass spectra**. *J Bioinform Comput Biol* 2010, **8**(6):981-994.

320. Liu X, Shan B, Xin L, Ma B: **Better score function for peptide identification with ETD MS/MS spectra**. *BMC Bioinformatics* 2010, **11 Suppl 1**:S4.

321. Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M: **Higher-energy C-trap dissociation for peptide modification analysis**. *Nature methods* 2007, **4**(9):709-712.

322. Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA: **Proteomics-grade de novo sequencing approach**. *Journal of proteome research* 2005, **4**(6):2348-2354.

323. Bern M, Cai Y, Goldberg D: **Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry**. *Anal Chem* 2007, **79**(4):1393-1400.

324. Kim S, Gupta N, Bandeira N, Pevzner PA: **Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra**. *Mol Cell Proteomics* 2009, **8**(1):53-69.

325. Jeong K, Kim S, Bandeira N, Pevzner PA: **Gapped spectral dictionaries and their applications for database searches of tandem mass spectra**. *Mol Cell Proteomics* 2011, **10**(6):M110 002220.

326. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: **InsPecT: identification of posttranslationally modified peptides from tandem mass spectra**. *Anal Chem* 2005, **77**(14):4626-4639.

327. Tabb DL, Saraf A, Yates JR, 3rd: **GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model**. *Analytical chemistry* 2003, **75**(23):6415-6421.

328. Searle BC, Dasari S, Turner M, Reddy AP, Choi D, Wilmarth PA, McCormack AL, David LL, Nagalla SR: **High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results**. *Analytical chemistry* 2004, **76**(8):2220-2230.

329. Sunyaev S, Liska AJ, Golod A, Shevchenko A: **MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry**. *Analytical chemistry* 2003, **75**(6):1307-1315.

330. Day RM, Borziak A, Gorin A: **PPM-chain - De novo peptide identification program comparable in performance to sequest**. *2004 Ieee Computational Systems Bioinformatics Conference, Proceedings* 2004:505-508.

331. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA: **The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra**. *Molecular & cellular proteomics : MCP* 2007, **6**(9):1638-1655.

332. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL: **TagRecon: high-throughput mutation identification through sequence tagging**. *Journal of proteome research* 2010, **9**(4):1716-1726.

333. Mackey AJ, Haystead TA, Pearson WR: **Getting More from Less Algorithms for Rapid Protein Identification with Multiple Short Peptide Sequences**. *Molecular & Cellular Proteomics* 2002, **1**(2):139-147.

334. Alves G, Yu YK: **Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with de novo based statistics**. *Bioinformatics* 2005, **21**(19):3726-3732.

335. Kim S, Bandeira N, Pevzner PA: **Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification**. *Molecular & cellular proteomics : MCP* 2009, **8**(6):1391-1400.

336. Jeong K, Kim S, Pevzner PA: **UniNovo: a universal tool for de novo peptide sequencing**. *Bioinformatics* 2013, **29**(16):1953-1962.

337. Guthals A, Clauser KR, Bandeira N: **Shotgun protein sequencing with meta-contig assembly**. *Molecular & cellular proteomics : MCP* 2012, **11**(10):1084-1096.

338. Guthals A, Clauser KR, Frank AM, Bandeira N: **Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides**. *Journal of proteome research* 2013, **12**(6):2846-2857.

339. Liu X, Dekker LJ, Wu S, Vanduijn MM, Luider TM, Tolic N, Kou Q, Dvorkin M, Alexandrova S, Vyatkina K *et al*: **De novo protein sequencing by combining top-down and bottom-up tandem mass spectra**. *Journal of proteome research* 2014, **13**(7):3241-3248.

340. Blueggel M, Chamrad D, Meyer HE: **Bioinformatics in proteomics**. *Curr Pharm Biotechno* 2004, **5**(1):79-88.

341. Eng JK, Mccormack AL, Yates JR: **An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database**. *J Am Soc Mass Spectr* 1994, **5**(11):976-989.

342. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database**. *Analytical chemistry* 1995, **67**(8):1426-1436.

343. Steen H, Mann M: **The ABC's (and XYZ's) of peptide sequencing**. *Nature reviews Molecular cell biology* 2004, **5**(9):699-711.

344. Hernandez P, Muller M, Appel RD: **Automated protein identification by tandem mass spectrometry: issues and strategies**. *Mass spectrometry reviews* 2006, **25**(2):235-254.

345. Nesvizhskii AI, Vitek O, Aebersold R: **Analysis and validation of proteomic data generated by tandem mass spectrometry**. *Nat Methods* 2007, **4**(10):787-797.

346. Nesvizhskii AI: **Protein identification by tandem mass spectrometry and sequence database searching**. *Methods in molecular biology* 2007, **367**:87-119.

347. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**(9):1466-1467.

348. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**(5):958-964.

349. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A: **The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data**. *Molecular & cellular proteomics : MCP* 2004, **3**(6):531-533.

350. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS *et al*: **An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis**. *Proteomics* 2005, **5**(13):3475-3490.

351. Adamski M, Blackwell T, Menon R, Martens L, Hermjakob H, Taylor C, Omenn GS, States DJ: **Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project**. *Proteomics* 2005, **5**(13):3246-3261.

352. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM: **Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study**. *Nat Biotechnol* 2006, **24**(3):333-338.

353. Hermjakob H: **The HUPO proteomics standards initiative--overcoming the fragmentation of proteomics data**. *Proteomics* 2006, **6 Suppl 2**:34-38.

354. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas project**. *Nucleic Acids Res* 2006, **34**(Database issue):D655-658.

355. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide identifications for the proteomics community**. *Nucleic Acids Res* 2006, **34**(Database issue):D659-663.

356. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N *et al*: **ProteomeXchange provides globally coordinated proteomics data submission and dissemination**. *Nat Biotechnol* 2014, **32**(3):223-226.

357. Kim S, Gupta N, Pevzner PA: **Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases**. *J Proteome Res* 2008, **7**(8):3354-3363.

358. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, Blair IA, FitzGerald GA, Grosser T: **EBP, a program for protein identification using multiple tandem mass spectrometry datasets**. *Molecular & cellular proteomics : MCP* 2007, **6**(3):527-536.

359. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW *et al*: **IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering**. *Journal of proteome research* 2009, **8**(8):3872-3881.

360. Zhang B, Chambers MC, Tabb DL: **Proteomic parsimony through bipartite graph analysis improves accuracy and transparency**. *Journal of proteome research* 2007, **6**(9):3549-3557.

361. Grobei MA, Qeli E, Brunner E, Rehrauer H, Zhang R, Roschitzki B, Basler K, Ahrens CH, Grossniklaus U: **Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function**. *Genome research* 2009, **19**(10):1786-1800.

362. Qeli E, Ahrens CH: **PeptideClassifier for protein inference and targeted quantitative proteomics**. *Nature biotechnology* 2010, **28**(7):647-650.

363. Li J, Zimmerman LJ, Park BH, Tabb DL, Liebler DC, Zhang B: **Network-assisted protein identification and data interpretation in shotgun proteomics**. *Mol Syst Biol* 2009, **5**:303.

364. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T *et al*: **Computational prediction of proteotypic peptides for quantitative proteomics**. *Nature biotechnology* 2007, **25**(1):125-131.

365. Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, Reilly JP, Radivojac P: **A computational approach toward label-free protein quantification using predicted peptide detectability**. *Bioinformatics* 2006, **22**(14):e481-488.

366. Alves P, Arnold RJ, Novotny MV, Radivojac P, Reilly JP, Tang H: **Advancement in protein inference from shotgun proteomics using peptide detectability**. *Pac Symp Biocomput* 2007:409-420.

367. Meyer-Arendt K, Old WM, Houel S, Renganathan K, Eichelberger B, Resing KA, Ahn NG: **IsoformResolver: A peptide-centric algorithm for protein inference**. *Journal of proteome research* 2011, **10**(7):3060-3075.

368. Zhou A, Zhang F, Chen JY: **PEPPI: a peptidomic database of human protein isoforms for proteomics experiments**. *BMC Bioinformatics* 2010, **11 Suppl 6**:S7.

369. Marx H, Lemeer S, Klaeger S, Rattei T, Kuster B: **MScDB: a mass spectrometry-centric protein sequence database for proteomics**. *Journal of proteome research* 2013, **12**(6):2386-2398.

370. Horn DM, Zubarev RA, McLafferty FW: **Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules**. *J Am Soc Mass Spectr* 2000, **11**(4):320-332.

371. Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL: **New and automated MSn approaches for top-down identification of modified proteins**. *J Am Soc Mass Spectr* 2005, **16**(12):2027-2038.

372. LeDuc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL: **ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry**. *Nucleic acids research* 2004, **32**(Web Server issue):W340-345.

373. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL: **ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry**. *Nucleic acids research* 2007, **35**(Web Server issue):W701-706.

374. Frank AM, Pesavento JJ, Mizzen CA, Kelleher NL, Pevzner PA: **Interpreting top-down mass spectra using spectral alignment**. *Analytical chemistry* 2008, **80**(7):2499-2505.

375. Karabacak NM, Li L, Tiwari A, Hayward LJ, Hong P, Easterling ML, Agar JN: **Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry**. *Molecular & cellular proteomics : MCP* 2009, **8**(4):846-856.

376. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA: **Protein identification using top-down**. *Molecular & cellular proteomics : MCP* 2012, **11**(6):M111 008524.

377. Liu X, Hengel S, Wu S, Tolic N, Pasa-Tolic L, Pevzner PA: **Identification of ultramodified proteins using top-down tandem mass spectra**. *Journal of proteome research* 2013, **12**(12):5830-5838.

378. **PNNL Algorithm Development** [http://omics.pnl.gov/algorithm-development]

379. Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ: **Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures**. *Proteomics* 2009, **9**(6):1696-1719.

380. Kwon J, Park SH, Park C, Kwon S-O, Choi J-S: **Analysis of membrane proteome by data-dependent LC-MS/MS combined with data-independent LC-MSE technique**. *J Anal Sci Technol* 2010, **1**(1):78-85.

381. Shliaha PV, Bond NJ, Gatto L, Lilley KS: **Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies**. *Journal of proteome research* 2013, **12**(6):2323-2339.

382. Bond NJ, Shliaha PV, Lilley KS, Gatto L: **Improving qualitative and quantitative performance for MSE-based label-free proteomics**. *Journal of proteome research* 2013, **12**(6):2340-2353.

383. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: **Andromeda: a peptide search engine integrated into the MaxQuant environment**. *Journal of proteome research* 2011, **10**(4):1794-1805.

384. Tsou C: **Computational strategies of data independent acquisition for high-throughput proteomics**. 2014.

385. **ASMS 2012: Probabilistic E-value Scoring for the Precursor Ion Independent Top-down Algorithm** [https://persephone.rxlab.umaryland.edu/index/posts/content/ASMS2012_poster/ASMS2012poster_Philip.pdf]

386. Domokos L, Henneberg D, Weimann B: **Computer-Aided Identification of Compounds by Comparison of Mass-Spectra**. *Analytica Chimica Acta* 1984, **165**(Nov):61-74.

387. Yates JR, 3rd, Morgan SF, Gatlin CL, Griffin PR, Eng JK: **Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis**. *Analytical chemistry* 1998, **70**(17):3557-3565.

388. Zong NC, Li H, Lam MP, Jimenez RC, Kim CS, Deng N, Kim AK, Choi JH, Zelaya I, Liem D *et al*: **Integration of cardiac proteome biology and medicine by a specialized knowledgebase**. *Circulation research* 2013, **113**(9):1043-1053.

389. **NIST Libraries of Peptide Tandem Mass Spectra** [http://peptide.nist.gov/]

390. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ: **Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries**. *Analytical chemistry* 2006, **78**(16):5678-5684.

391. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R: **Development and validation of a spectral library searching method for peptide identification from MS/MS**. *Proteomics* 2007, **7**(5):655-667.

392. Stein SE, Scott DR: **Optimization and Testing of Mass-Spectral Library Search Algorithms for Compound Identification**. *J Am Soc Mass Spectr* 1994, **5**(9):859-866.

393. Craig R, Cortens JC, Fenyo D, Beavis RC: **Using annotated peptide mass spectrum libraries for protein identification**. *Journal of proteome research* 2006, **5**(8):1843-1849.

394. Dasari S, Chambers MC, Martinez MA, Carpenter KL, Ham AJ, Vega-Montoto LJ, Tabb DL: **Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment**. *J Proteome Res* 2012, **11**(3):1686-1695.

395. Falkner JA, Falkner JW, Yocum AK, Andrews PC: **A spectral clustering approach to MS/MS identification of post-translational modifications**. *J Proteome Res* 2008, **7**(11):4614-4622.

396. **MSplit-DIA** [http://proteomics.ucsd.edu/software-tools/msplit-dia/]

397. Wang M, Bandeira N: **Spectral library generating function for assessing spectrum-spectrum match significance**. *J Proteome Res* 2013, **12**(9):3944-3951.

398. Frank AM, Monroe ME, Shah AR, Carver JJ, Bandeira N, Moore RJ, Anderson GA, Smith RD, Pevzner PA: **Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra**. *Nature methods* 2011, **8**(7):587-591.

399. Lam H: **Spectral archives: a vision for future proteomics data repositories**. *Nature methods* 2011, **8**(7):546-548.

400. Guthals A, Watrous JD, Dorrestein PC, Bandeira N: **The spectral networks paradigm in high throughput mass spectrometry**. *Mol Biosyst* 2012, **8**(10):2535-2544.

401. Pevzner PA, Mulyukov Z, Dancik V, Tang CL: **Efficiency of database search for identification of mutated and modified proteins via mass spectrometry**. *Genome research* 2001, **11**(2):290-299.

402. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA: **Identification of post-translational modifications via blind search of mass-spectra**. *Proc IEEE Comput Syst Bioinform Conf* 2005:157-166.

403. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search**. *Anal Chem* 2002, **74**(20):5383-5392.

404. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI: **iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates**. *Molecular & cellular proteomics : MCP* 2011, **10**(12):M111 007690.

405. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ: **Semi-supervised learning for peptide identification from shotgun proteomics datasets**. *Nature methods* 2007, **4**(11):923-925.

406. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW: **Combining results of multiple search engines in proteomics**. *Molecular & cellular proteomics : MCP* 2013, **12**(9):2383-2393.

407. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJR, Pevzner PA: **The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search**. *Molecular & Cellular Proteomics* 2010, **9**(12):2840-2852.

408. Fenyo D, Beavis RC: **A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes**. *Analytical chemistry* 2003, **75**(4):768-774.

409. Eriksson J, Chait BT, Fenyo D: **A statistical basis for testing the significance of mass spectrometric protein identification results**. *Analytical chemistry* 2000, **72**(5):999-1005.

410. Sadygov RG, Yates JR, 3rd: **A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases**. *Analytical chemistry* 2003, **75**(15):3792-3798.

411. Kall L, Storey JD, MacCoss MJ, Noble WS: **Assigning significance to peptides identified by tandem mass spectrometry using decoy databases**. *J Proteome Res* 2008, **7**(1):29-34.

412. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**. *Nat Methods* 2007, **4**(3):207-214.

413. Elias JE, Gygi SP: **Target-decoy search strategy for mass spectrometry-based proteomics**. *Methods Mol Biol* 2010, **604**:55-71.

414. Moore RE, Young MK, Lee TD: **Qscore: an algorithm for evaluating SEQUEST database search results**. *J Am Soc Mass Spectr* 2002, **13**(4):378-386.

415. Klammer AA, MacCoss MJ: **Effects of modified digestion schemes on the identification of proteins from complex mixtures**. *Journal of proteome research* 2006, **5**(3):695-700.

416. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. *J Roy Stat Soc B Met* 1995, **57**(1):289-300.

417. Choi H, Nesvizhskii AI: **False discovery rates and related statistical concepts in mass spectrometry-based proteomics**. *J Proteome Res* 2008, **7**(1):47-50.

418. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.

419. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes analysis of a microarray experiment**. *J Am Stat Assoc* 2001, **96**(456):1151-1160.

420. Choi H, Ghosh D, Nesvizhskii AI: **Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling**. *Journal of proteome research* 2008, **7**(1):286-292.

421. Kall L, Storey JD, MacCoss MJ, Noble WS: **Posterior error probabilities and false discovery rates: two sides of the same coin**. *Journal of proteome research* 2008, **7**(1):40-44.

422. Storey JD, Akey JM, Kruglyak L: **Multiple locus linkage analysis of genomewide expression in yeast**. *PLoS Biol* 2005, **3**(8):e267.

423. Alves G, Yu YK: **Statistical Characterization of a 1D Random Potential Problem - with applications in score statistics of MS-based peptide sequencing**. *Physica A* 2008, **387**(26):6538-6544.

424. Gupta N, Bandeira N, Keich U, Pevzner PA: **Target-decoy approach and false discovery rate: when things may go wrong**. *J Am Soc Mass Spectr* 2011, **22**(7):1111-1120.

425. Gupta N, Pevzner PA: **False discovery rates of protein identifications: a strike against the two-peptide rule**. *J Proteome Res* 2009, **8**(9):4173-4181.

426. Nesvizhskii AI: **A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics**. *J Proteomics* 2010, **73**(11):2092-2123.

427. Jeong K, Kim S, Bandeira N: **False discovery rates in spectral identification**. *BMC Bioinformatics* 2012, **13 Suppl 16**:S2.

428. Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, Seymour SL, Griffin TJ: **A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies**. *Proteomics* 2013, **13**(8):1352-1357.

429. Branca RM, Orre LM, Johansson HJ, Granholm V, Huss M, Perez-Bercoff A, Forshed J, Kall L, Lehtio J: **HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics**. *Nat Methods* 2014, **11**(1):59-62.

430. Craig R, Beavis RC: **A method for reducing the time required to match protein sequences with tandem mass spectra**. *Rapid communications in mass spectrometry : RCM* 2003, **17**(20):2310-2316.

431. Na S, Bandeira N, Paek E: **Fast multi-blind modification search through tandem mass spectrometry**. *Molecular & cellular proteomics : MCP* 2012, **11**(4):M111 010199.

432. Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M: **The Ensembl analysis pipeline**. *Genome Res* 2004, **14**(5):934-941.

433. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **UniProt: the Universal Protein knowledgebase**. *Nucleic acids research* 2004, **32**(Database issue):D115-119.

434. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its new supplement TREMBL**. *Nucleic acids research* 1996, **24**(1):21-25.

435. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic acids research* 2005, **33**(Database issue):D501-504.

436. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD: **Proteogenomics: needs and roles to be filled by proteomics in genome annotation**. *Brief Funct Genomic Proteomic* 2008, **7**(1):50-62.

437. Castellana N, Bafna V: **Proteogenomics to discover the full coding content of genomes: a computational perspective**. *J Proteomics* 2010, **73**(11):2124-2135.

438. Renuse S, Chaerkady R, Pandey A: **Proteogenomics**. *Proteomics* 2011, **11**(4):620-630.

439. Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies**. *Nature methods* 2014, **11**(11):1114-1125.

440. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation**. *Proteomics* 2004, **4**(1):59-77.

441. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O *et al*: **A high-quality catalog of the Drosophila melanogaster proteome**. *Nat Biotechnol* 2007, **25**(5):576-583.

442. Tress ML, Bodenmiller B, Aebersold R, Valencia A: **Proteomics studies confirm the presence of alternative protein isoforms on a large scale**. *Genome Biol* 2008, **9**(11):R162.

443. Payne SH, Huang ST, Pieper R: **A proteogenomic update to Yersinia: enhancing genome annotation**. *BMC Genomics* 2010, **11**:460.

444. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS *et al*: **Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome**. *Genome Res* 2011, **21**(5):756-767.

445. Christie-Oleza JA, Miotello G, Armengaud J: **High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade**. *BMC Genomics* 2012, **13**:73.

446. Christie-Oleza JA, Pina-Villalonga JM, Bosch R, Nogales B, Armengaud J: **Comparative proteogenomics of twelve Roseobacter exoproteomes reveals different adaptive strategies among these marine bacteria**. *Mol Cell Proteomics* 2012, **11**(2):M111 013110.

447. Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD *et al*: **Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation**. *Genome Res* 2007, **17**(9):1362-1377.

448. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J *et al*: **Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes**. *Genome Res* 2008, **18**(7):1133-1142.

449. Risk BA, Spitzer WJ, Giddings MC: **Peppy: proteogenomic search software**. *Journal of proteome research* 2013, **12**(6):3019-3025.

450. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al*: **A draft map of the human proteome**. *Nature* 2014, **509**(7502):575-581.

451. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.

452. Hartmann EM, Armengaud J: **N-terminomics and proteogenomics, getting off to a good start**. *Proteomics* 2014, **14**(23-24):2637-2646.

453. Yates JR, 3rd, Eng JK, McCormack AL: **Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases**. *Anal Chem* 1995, **67**(18):3202-3210.

454. Link AJ, Hays LG, Carmack EB, Yates JR, 3rd: **Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143**. *Electrophoresis* 1997, **18**(8):1314-1334.

455. Kuster B, Mortensen P, Andersen JS, Mann M: **Mass spectrometry allows direct identification of proteins in large genomes**. *Proteomics* 2001, **1**(5):641-650.

456. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Interrogating the human genome using uninterpreted mass spectrometry data**. *Proteomics* 2001, **1**(5):651-667.

457. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry**. *Genome research* 2007, **17**(2):231-239.

458. Edwards NJ: **Novel peptide identification from tandem mass spectra using ESTs and sequence database compression**. *Mol Syst Biol* 2007, **3**:102.

459. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al*: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome research* 2012, **22**(9):1775-1789.

460. Ning K, Nesvizhskii AI: **The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment**. *BMC Bioinformatics* 2010, **11 Suppl 11**:S14.

461. Sheynkman GM, Shortreed MR, Frey BL, Smith LM: **Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq**. *Molecular & cellular proteomics : MCP* 2013, **12**(8):2341-2353.

462. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nature biotechnology* 2010, **28**(5):511-515.

463. Wu P, Zhang H, Lin W, Hao Y, Ren L, Zhang C, Li N, Wei H, Jiang Y, He F: **Discovery of novel genes and gene isoforms by integrating transcriptomic and proteomic profiling from mouse liver**. *Journal of proteome research* 2014, **13**(5):2409-2419.

464. Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappe J, Gevaert K, Van Damme P: **Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events**. *Molecular & cellular proteomics : MCP* 2013, **12**(7):1780-1790.

465. Koch A, Gawron D, Steyaert S, Ndah E, Crappe J, De Keulenaer S, De Meester E, Ma M, Shen B, Gevaert K *et al*: **A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites**. *Proteomics* 2014.

466. Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, Griffin TJ, Smith LM: **Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations**. *BMC Genomics* 2014, **15**:703.

467. Wang X, Zhang B: **customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search**. *Bioinformatics* 2013, **29**(24):3235-3237.

468. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic acids research* 2001, **29**(1):308-311.

469. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic acids research* 2005, **33**(Database issue):D514-517.

470. Nishikawa K, Ishino S, Takenaka H, Norioka N, Hirai T, Yao T, Seto Y: **Protocol - Constructing a Protein Mutant Database**. *Protein Eng* 1994, **7**(5):733-733.

471. Picardi E, Pesole G: **REDItools: high-throughput RNA editing detection made easy**. *Bioinformatics* 2013, **29**(14):1813-1814.

472. Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, MacCoss M, Bafna V: **Proteogenomic database construction driven from large scale RNA-seq data**. *J Proteome Res* 2014, **13**(1):21-28.

473. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M *et al*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat Genet* 2011, **43**(5):491-498.

474. Woo S, Cha SW, Na S, Guest C, Liu T, Smith RD, Rodland KD, Payne S, Bafna V: **Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data**. *Proteomics* 2014, **14**(23-24):2719-2730.

475. Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S: **ECgene: genome annotation for alternative splicing**. *Nucleic acids research* 2005, **33**(Database issue):D75-79.

476. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M: **Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation**. *Nucleic acids research* 2007, **35**(Database issue):D55-60.

477. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y: **NONCODEv4: exploring the world of long non-coding RNA genes**. *Nucleic acids research* 2014, **42**(Database issue):D98-103.

478. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes Dev* 2011, **25**(18):1915-1927.

479. Frenkel-Morgenstern M, Gorohovski A, Lacroix V, Rogers M, Ibanez K, Boullosa C, Andres Leon E, Ben-Hur A, Valencia A: **ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data**. *Nucleic acids research* 2013, **41**(Database issue):D142-151.

480. Junqueira M, Spirin V, Balbuena TS, Thomas H, Adzhubei I, Sunyaev S, Shevchenko A: **Protein identification pipeline for the homology-driven proteomics**. *J Proteomics* 2008, **71**(3):346-356.

481. Castellana NE, Pham V, Arnott D, Lill JR, Bafna V: **Template proteogenomics: sequencing whole proteins using an imperfect database**. *Molecular & cellular proteomics : MCP* 2010, **9**(6):1260-1270.

482. Castellana N: **Proteogenomics: applications of mass spectrometry at the interface of genomics and proteomics**. 2012.

483. Blakeley P, Overton IM, Hubbard SJ: **Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies**. *Journal of proteome research* 2012.

484. Krug K, Carpy A, Behrends G, Matic K, Soares NC, Macek B: **Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments**. *Molecular & cellular proteomics : MCP* 2013, **12**(11):3420-3430.

485. Li Y, Chi H, Wang LH, Wang HP, Fu Y, Yuan ZF, Li SJ, Liu YS, Sun RX, Zeng R *et al*: **Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing**. *Rapid Commun Mass Spectrom* 2010, **24**(6):807-814.

486. Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS: **Rapid and accurate peptide identification from tandem mass spectra**. *J Proteome Res* 2008, **7**(7):3022-3027.

487. Li D, Gao W, Ling CX, Wang X, Sun R, He S: **IndexToolkit: an open source toolbox to index protein databases for high-throughput proteomics**. *Bioinformatics* 2006, **22**(20):2572-2573.

488. Sun J, Chen B, Wu FX: **An improved peptide-spectral matching algorithm through distributed search over multiple cores and multiple CPUs**. *Proteome Sci* 2014, **12**:18.

489. Quandt A, Masselot A, Hernandez P, Hernandez C, Maffioletti S, Appel RD, Lisacek F: **SwissPIT: An workflow-based platform for analyzing tandem-MS spectra using the Grid**. *Proteomics* 2009, **9**(10):2648-2655.

490. Halligan BD, Geiger JF, Vallejos AK, Greene AS, Twigger SN: **Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms**. *J Proteome Res* 2009, **8**(6):3148-3153.

491. Roos FF, Jacob R, Grossmann J, Fischer B, Buhmann JM, Gruissem W, Baginsky S, Widmayer P: **PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra**. *Bioinformatics* 2007, **23**(22):3016-3023.

492. Venter E, Smith RD, Payne SH: **Proteogenomic analysis of bacteria and archaea: a 46 organism case study**. *PLoS One* 2011, **6**(11):e27587.

493. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R: **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* 2006, **5**(4):652-670.

494. Ning K, Fermin D, Nesvizhskii AI: **Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets**. *Proteomics* 2010, **10**(14):2712-2718.

495. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics**. *Genes Cells* 2012, **17**(8):633-644.

496. Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, Mandyam RA, Nones K, Pearson JV, Grimmond SM: **PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization**. *Journal of proteome research* 2015, **14**(5):2255-2266.

497. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences**. *Proc Int Conf Intell Syst Mol Biol* 1999:138-148.

498. Blakeley P: **Computational proteomics for genome annotation**. University of Manchester; 2013.

499. Abraham P, Adams RM, Tuskan GA, Hettich RL: **Moving away from the reference genome: evaluating a peptide sequencing tagging approach for single amino acid polymorphism identifications in the genus Populus**. *Journal of proteome research* 2013, **12**(8):3642-3651.

500. Li J, Su Z, Ma ZQ, Slebos RJ, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B: **A bioinformatics workflow for variant peptide detection in shotgun proteomics**. *Molecular & cellular proteomics : MCP* 2011, **10**(5):M110 006536.

501. Armirotti A, Millo E, Damonte G: **How to discriminate between leucine and isoleucine by low energy ESI-TRAP MSn**. *J Am Soc Mass Spectr* 2007, **18**(1):57-63.

502. Kumar D, Yadav AK, Kadimi PK, Nagaraj SH, Grimmond SM, Dash D: **Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline**. *Mol Cell Proteomics* 2013, **12**(11):3388-3397.

503. Seifert J, Herbst FA, Halkjaer Nielsen P, Planes FJ, Jehmlich N, Ferrer M, von Bergen M: **Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities**. *Proteomics* 2013, **13**(18-19):2786-2804.

504. Amann RI, Ludwig W, Schleifer KH: **Phylogenetic identification and in situ detection of individual microbial cells without cultivation**. *Microbiol Rev* 1995, **59**(1):143-169.

505. Raes J, Bork P: **Molecular eco-systems biology: towards an understanding of community function**. *Nat Rev Microbiol* 2008, **6**(9):693-699.

506. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S *et al*: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry**. *Genome biology* 2005, **6**(1):R9.

507. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments**. *Proteomics* 2004, **4**(7):1985-1988.

508. Shadforth I, Xu W, Crowther D, Bessant C: **GAPP: a fully automated software for the confident identification of human peptides from tandem mass spectra**. *Journal of proteome research* 2006, **5**(10):2849-2852.

509. Shadforth I, Dunkley T, Lilley K, Crowther D, Bessant C: **Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds**. *Rapid Commun Mass Spectrom* 2005, **19**(22):3363-3368.

510. Chepanoske CL, Richardson BE, von Rechenberg M, Peltier JM: **Average peptide score: a useful parameter for identification of proteins derived from database searches of liquid chromatography/tandem mass spectrometry data**. *Rapid Commun Mass Spectrom* 2005, **19**(1):9-14.

511. Allmer J, Markert C, Stauber EJ, Hippler M: **A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases**. *FEBS Lett* 2004, **562**(1-3):202-206.

512. Ferro M, Tardif M, Reguer E, Cahuzac R, Bruley C, Vermat T, Nugues E, Vigouroux M, Vandenbrouck Y, Garin Jrm: **PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences**. *Journal of proteome research* 2008, **7**(5):1873-1883.

513. Holmes MR, Giddings MC: **Using GFS to identify encoding genomic loci from protein mass spectral data**. *Curr Protoc Bioinformatics* 2008, **Chapter 13**:Unit 13 19.

514. Sanders WS, Wang N, Bridges SM, Malone BM, Dandass YS, McCarthy FM, Nanduri B, Lawrence ML, Burgess SC: **The proteogenomic mapping tool**. *BMC Bioinformatics* 2011, **12**:115.

515. McCarthy FM, Cooksey AM, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a whole organ using proteomics: the avian bursa of Fabricius**. *Proteomics* 2006, **6**(9):2759-2771.

516. Pertea M, Lin X, Salzberg SL: **GeneSplicer: a new computational method for splice site prediction**. *Nucleic Acids Res* 2001, **29**(5):1185-1190.

517. Pang CN, Tay AP, Aya C, Twine NA, Harkness L, Hart-Smith G, Chia SZ, Chen Z, Deshpande NP, Kaakoush NO *et al*: **Tools to Co-visualize and Co-analyse Proteomic data with Genomes and Transcriptomes: validation of genes and alternative mRNA splicing**. *Journal of proteome research* 2013.

518. Jones AR, Siepen JA, Hubbard SJ, Paton NW: **Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines**. *Proteomics* 2009, **9**(5):1220-1229.

519. Yadav AK, Kumar D, Dash D: **MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry**. *Journal of proteome research* 2011, **10**(5):2154-2160.

520. Eng JK, Jahan TA, Hoopmann MR: **Comet: an open-source MS/MS sequence database search tool**. *Proteomics* 2013, **13**(1):22-24.

521. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K *et al*: **OpenMS - an open-source software framework for mass spectrometry**. *BMC Bioinformatics* 2008, **9**:163.

522. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome research* 2009, **19**(9):1639-1645.

523. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Brief Bioinform* 2013, **14**(2):178-192.

524. Tovchigrechko A, Venepally P, Payne SH: **PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations**. *Bioinformatics* 2014.

525. **ProteoSAFe** [http://proteomics2.ucsd.edu/ProteoSAFe/]

526. Altintas I, Berkley C, Jaeger E, Jones M, Ludascher B, Mock S: **Kepler: an extensible system for design and execution of scientific workflows**. In: *Scientific and Statistical Database Management, 2004 Proceedings 16th International Conference on: 2004*: IEEE; 2004: 423-424.

527. Deelman E, Singh G, Su M-H, Blythe J, Gil Y, Kesselman C, Mehta G, Vahi K, Berriman GB, Good J: **Pegasus: A framework for mapping complex scientific workflows onto distributed systems**. *Scientific Programming* 2005, **13**(3):219-237.

528. Elhai J, Taton A, Massar JP, Myers JK, Travers M, Casey J, Slupesky M, Shrager J: **BioBIKE: a Web-based, programmable, integrated biological knowledge base**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W28-32.

529. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A *et al*: **Taverna: a tool for the composition and enactment of bioinformatics workflows**. *Bioinformatics* 2004, **20**(17):3045-3054.

530. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W729-732.

531. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al*: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome research* 2005, **15**(10):1451-1455.

532. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol* 2010, **11**(8):R86.

533. Hunter AA, Macgregor AB, Szabo TO, Wellington CA, Bellgard MI: **Yabi: An online research environment for grid, high performance and cloud computing**. *Source Code Biol Med* 2012, **7**(1):1.

534. Dinov ID, Torri F, Macciardi F, Petrosyan P, Liu Z, Zamanyan A, Eggert P, Pierce J, Genco A, Knowles JA *et al*: **Applications of the pipeline environment for visual informatics and genomics computations**. *BMC Bioinformatics* 2011, **12**:304.

535. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P *et al*: **myExperiment: a repository and social**

network for the sharing of bioinformatics workflows. *Nucleic acids research* 2010, **38**(Web Server issue):W677-682.

536. Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, Wang Y, Shenykman GM, Bandhakavi S, Smith LM, Griffin TJ: **Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework**. *Journal of proteome research* 2014, **13**(12):5898-5908.

537. Ylonen T, Lonvick C: **The secure shell (SSH) authentication protocol**. 2006.

538. Allcock W, Bresnahan J, Kettimuthu R, Link M, Dumitrescu C, Raicu I, Foster I: **The Globus striped GridFTP framework and server**. In: *Proceedings of the 2005 ACM/IEEE conference on Supercomputing: 2005*: IEEE Computer Society; 2005: 54.

539. **Amazon Simple Storage Service** [http://aws.amazon.com/s3/]

540. Staples G: **TORQUE resource manager**. In: *Proceedings of the 2006 ACM/IEEE conference on Supercomputing: 2006*: ACM; 2006: 8.

541. Nitzberg B, Schopf JM, Jones JP: **PBS Pro: Grid computing and scheduling attributes**. In: *Grid resource management.* Springer; 2004: 183-190.

542. Gremme G, Steinbiss S, Kurtz S: **GenomeTools: a comprehensive software library for efficient processing of structured genome annotations**. *IEEE/ACM Trans Comput Biol Bioinform* 2013, **10**(3):645-656.

543. Andrews S: **FastQC: A quality control tool for high throughput sequence data**. *Reference Source* 2010.

544. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014.

545. Dobin A, Gingeras TR: **Comment on "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions" by Kim et al**. *bioRxiv* 2013.

546. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

547. Piehowski PD, Petyuk VA, Sandoval JD, Burnum KE, Kiebel GR, Monroe ME, Anderson GA, Camp DG, 2nd, Smith RD: **STEPS: a grid search methodology for optimized peptide identification filtering of MS/MS database search results**. *Proteomics* 2013, **13**(5):766-770.

548. Castellana NE, Shen Z, He Y, Walley JW, Cassidy CJ, Briggs SP, Bafna V: **An Automated Proteogenomic Method Utilizes Mass Spectrometry to Reveal Novel Genes in Zea mays**. *Molecular & cellular proteomics : MCP* 2013.

549. Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, Maccoss M, Bafna V: **Proteogenomic Database Construction Driven from Large Scale RNA-seq Data**. *Journal of proteome research* 2013.

550. Kuster B, Schirle M, Mallick P, Aebersold R: **Scoring proteomes with proteotypic peptide probes**. *Nat Rev Mol Cell Biol* 2005, **6**(7):577-583.

551. Lange V, Picotti P, Domon B, Aebersold R: **Selected reaction monitoring for quantitative proteomics: a tutorial**. *Mol Syst Biol* 2008, **4**:222.

552. Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, Mandyam RA, Nones K, Pearson JV, Grimmond SM: **PGTools: a software suite for proteogenomics data analysis and visualization**. *Journal of proteome research* 2015.

553. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: **Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia**. *Science* 1985, **230**(4732):1350-1354.

554. Mali P, Esvelt KM, Church GM: **Cas9 as a versatile tool for engineering biology**. *Nat Methods* 2013, **10**(10):957-963.

555. Bertelli C, Greub G: **Rapid bacterial genome sequencing: methods and applications in clinical microbiology**. *Clin Microbiol Infect* 2013, **19**(9):803-813.

556. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S, Watanabe A, Idesawa K, Iriguchi M, Kawashima K *et al*: **Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110**. *DNA Res* 2002, **9**(6):189-197.

557. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer**. *Bioinformatics* 2007, **23**(6):673-679.

558. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**:119.

559. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: the proteomics identifications database**. *Proteomics* 2005, **5**(13):3537-3545.

560. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases**. *Bioinformatics* 2001, **17**(3):282-283.

561. Fonseca MM, Harris DJ, Posada D: **Origin and length distribution of unidirectional prokaryotic overlapping genes**. *G3 (Bethesda)* 2014, **4**(1):19-27.

562. Palleja A, Harrington ED, Bork P: **Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions?** *BMC Genomics* 2008, **9**:335.

563. Oliver JL, Marin A: **A relationship between GC content and coding-sequence length**. *J Mol Evol* 1996, **43**(3):216-223.

564. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic acids research* 2004, **32**(5):1792-1797.

565. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

566. Bellgard MI, Gojobori T: **Identification of a ribonuclease H gene in both Mycoplasma genitalium and Mycoplasma pneumoniae by a new method for exhaustive identification of ORFs in the complete genome sequences**. *FEBS letters* 1999, **445**(1):6-8.

567. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome biology* 2004, **5**(2):R12.

568. Panagiotakos DB, Pitsavos C, Polychronopoulos E, Chrysohoou C, Zampelas A, Trichopoulou A: **Can a Mediterranean diet moderate the development and clinical**

progression of coronary heart disease? A systematic review. *Med Sci Monit* 2004, **10**(8):RA193-198.

569. Burns J, Gardner PT, O'Neil J, Crawford S, Morecroft I, McPhail DB, Lister C, Matthews D, MacLean MR, Lean ME *et al*: **Relationship among antioxidant activity, vasodilation capacity, and phenolic content of red wines**. *Journal of agricultural and food chemistry* 2000, **48**(2):220-230.

570. Agarwal B, Baur JA: **Resveratrol and life extension**. *Annals of the New York Academy of Sciences* 2011, **1215**:138-143.

571. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J *et al*: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PLoS One* 2007, **2**(12):e1326.

572. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**(7161):463-467.

573. Forcato C: **Gene prediction and functional annotation in the Vitis vinifera genome**. 2010.

574. Vizcaino JA, Cote R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H, Martens L: **The Proteomics Identifications database: 2010 update**. *Nucleic acids research* 2010, **38**(Database issue):D736-742.

575. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: a web server for gene finding in eukaryotes**. *Nucleic acids research* 2004, **32**(Web Server issue):W309-312.

576. Smit A, Hubley R, Green P: **1996–2010. RepeatMasker Open-3.0**. *URL: http://www repeatmasker org*.

577. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Res* 1999, **27**(2):573-580.

578. Grimplet J, Van Hemert J, Carbonell-Bejerano P, Diaz-Riquelme J, Dickerson J, Fennell A, Pezzotti M, Martinez-Zapater JM: **Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences**. *BMC Res Notes* 2012, **5**:213.

579. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: **Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information**. *Nucleic acids research* 1996, **24**(17):3439-3452.

580. Arnesen T: **Towards a functional understanding of protein N-terminal acetylation**. *PLoS Biol* 2011, **9**(5):e1001074.

581. Simpson GG, Laurie RE, Dijkwel PP, Quesada V, Stockwell PA, Dean C, Macknight RC: **Noncanonical translation initiation of the Arabidopsis flowering time and alternative polyadenylation regulator FCA**. *The Plant cell* 2010, **22**(11):3764-3777.

582. Christensen AC, Lyznik A, Mohammed S, Elowsky CG, Elo A, Yule R, Mackenzie SA: **Dual-domain, dual-targeting organellar protein presequences in Arabidopsis can use non-AUG start codons**. *The Plant cell* 2005, **17**(10):2805-2816.

583. Depeiges A, Degroote F, Espagnol MC, Picard G: **Translation initiation by non-AUG codons in Arabidopsis thaliana transgenic plants**. *Plant Cell Rep* 2006, **25**(1):55-61.

584. Riechmann JL, Ito T, Meyerowitz EM: **Non-AUG initiation of AGAMOUS mRNA translation in Arabidopsis thaliana**. *Molecular and cellular biology* 1999, **19**(12):8505-8512.

585. Perseke M, Hetmank J, Bernt M, Stadler PF, Schlegel M, Bernhard D: **The enigmatic mitochondrial genome of Rhabdopleura compacta (Pterobranchia) reveals insights into selection of an efficient tRNA system and supports monophyly of Ambulacraria**. *BMC Evol Biol* 2011, **11**:134.

586. Yokobori S, Ueda T, Feldmaier-Fuchs G, Paabo S, Ueshima R, Kondow A, Nishikawa K, Watanabe K: **Complete DNA sequence of the mitochondrial genome of the ascidian Halocynthia roretzi (Chordata, Urochordata)**. *Genetics* 1999, **153**(4):1851-1862.

587. Hirose T, Ideue T, Wakasugi T, Sugiura M: **The chloroplast infA gene with a functional UUG initiation codon**. *FEBS letters* 1999, **445**(1):169-172.

588. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015.

589. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.

590. Myers R, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison R, Bernstein B, Gingeras T, Kent W, Birney E, Wold B *et al*: **A user's guide to the encyclopedia of DNA elements (ENCODE)**. *PLoS Biol* 2011, **9**(4):e1001046.

591. Risk BA, Edwards NJ, Giddings MC: **A peptide-spectrum scoring system based on ion alignment, intensity, and pair probabilities**. *J Proteome Res* 2013, **12**(9):4240-4247.

592. Tran JC, Doucette AA: **Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation**. *Analytical chemistry* 2008, **80**(5):1568-1573.

593. Wisniewski JR, Zougman A, Nagaraj N, Mann M: **Universal sample preparation method for proteome analysis**. *Nature methods* 2009, **6**(5):359-362.

594. Yu Y, Xie L, Gunawardena HP, Khatun J, Maier C, Spitzer W, Leerkes M, Giddings MC, Chen X: **GOFAST: an integrated approach for efficient and comprehensive membrane proteome analysis**. *Analytical chemistry* 2012, **84**(21):9008-9014.

595. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D *et al*: **Analysis of the bread wheat genome using whole-genome shotgun sequencing**. *Nature* 2012, **491**(7426):705-710.

596. Shewry PR: **Wheat**. *J Exp Bot* 2009, **60**(6):1537-1553.

597. Gibb S: **cleaver: Cleavage of Polypeptide Sequences.** In.: R package version 1.4.0, https://github.com/sgibb/cleaver/; 2014.

598. Borodovsky M, Mcininch J: **Genmark - Parallel Gene Recognition for Both DNA Strands**. *Comput Chem* 1993, **17**(2):123-133.

599. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding**. *Nucleic Acids Research* 1998, **26**(4):1107-1115.

600. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA**. *Nucleic Acids Res* 2009, **37**(18):e123.

601. Xu H, Freitas MA: **MassMatrix: a database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data**. *Proteomics* 2009, **9**(6):1548-1555.

602. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome research* 2003, **13**(11):2498-2504.

603. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms**. *Nucleic acids research* 2005, **33**(Database issue):D433-437.

604. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2008, **4**(1):44-57.

605. Noyce AB, Smith R, Dalgleish J, Taylor RM, Erb KC, Okuda N, Prince JT: **Mspire-Simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data**. *Journal of proteome research* 2013, **12**(12):5742-5749.