Computational Proteomics for Genome Annotation

Paul David Blakeley
January 2013
List of Contents

List of Figures ......................................................................................................................... 5
List of Tables ........................................................................................................................... 6
Abstract ..................................................................................................................................... 7
Declaration ................................................................................................................................. 8
Copyright ................................................................................................................................. 9
Acknowledgements .................................................................................................................. 10
The Author ............................................................................................................................... 11
Rationale for an Alternative Format Thesis .............................................................................. 12
Abbreviations .......................................................................................................................... 13

Chapter 1: Introduction .......................................................................................................... 14
  1.1. Genome annotation ......................................................................................................... 15
      1.1.1. Accurate genome annotation requires experimental evidence .............................. 16
      1.1.2. Genome annotation pipelines map experimental data to the genome .................... 19
      1.1.3. The need for protein-level data .............................................................................. 21
  1.2. Mapping the proteome ................................................................................................... 23
      1.2.1. High-throughput proteomics ................................................................................. 24
      1.2.2. Targeted Proteomics ............................................................................................. 26
      1.2.3. Peptide Identification via database searching .......................................................... 28
  1.3. Proteogenomics ............................................................................................................. 30
      1.3.1. Introducing proteogenomic approaches to genome annotation ............................... 31
  1.4. Overview ......................................................................................................................... 36
  1.5. References ...................................................................................................................... 38

Chapter 2: Investigating protein isoforms via proteomics: a feasibility study ....................... 49
  Abstract .................................................................................................................................... 50
  2.1. Introduction ...................................................................................................................... 51
  2.2. Materials and Methods .................................................................................................. 53
      2.2.1. Genome and proteome sequences ........................................................................... 53
      2.2.2. Peptide identifications ........................................................................................... 54
      2.2.3. Chicken samples and data analysis ........................................................................ 55
      2.2.4. Peptide mapping .................................................................................................... 55
      2.2.5. Database comparison with Swiss-Prot HPI ............................................................... 56
      2.2.6. QconCAT design ................................................................................................... 57
  2.3. Results and discussion ..................................................................................................... 58
      2.3.1. Proteomic data and alternate splicing ...................................................................... 58
      2.3.2. Examples of alternate splicing with supporting proteomic evidence ....................... 67
      2.3.3. Protein Isoforms in UniProtKB/Swiss-Prot ............................................................... 69
      2.3.4. Implications for design ........................................................................................... 71
      2.3.5. Isoform detection and quantification using QconCATs and SRM ............................. 74

1 WORD COUNT: 39,225
Chapter 3: Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies

3.1. Introduction

3.2. Materials and Methods

3.2.1. EST dataset

3.2.2. Preparation of chicken samples and mass spectrometry

3.2.3. Databases

3.2.4. EORF description

3.2.5. Mass spectrometry database searching

3.2.6. Statistical methods for validating PSMs

3.2.7. Estimating the proportion of correct PSMs

3.3. Results and Discussion

3.3.1. Searching against Six-frame or redundant databases affects sensitivity

3.3.2. The target-decoy approach over-estimates the q-value/PEP for the six-frame database search

3.3.3. Six-frame databases confound the target-decoy assumption

3.3.4. Identification of single amino acid polymorphisms (SAAPs)

3.4. Conclusion

3.5. References

Chapter 4: Improving Genome Annotation using Shotgun Proteomics and De Novo-Assembled Transcripts

4.1. Introduction

4.2. Materials and Methods

4.2.1. EST dataset

4.2.2. Conceptual translations

4.2.3. Preparation of chicken samples and mass spectrometry

4.2.4. Databases

4.2.5. Mass spectrometry database searching

4.2.6. Coverage of proteins predicted from ESTs

4.2.7. Identification of mismatches in peptide sequences

4.2.8. Validation of high-confidence SNPs

4.2.9. Finding novel protein-coding regions

4.2.10. Finding partial and missing genes

4.2.11. Functional annotation of the novel LMAN2 gene

4.3. Results and Discussion

4.3.1. EST translation approaches for genome annotation

4.3.2. Identification rates using the different databases

4.3.3. Mapping ESTs and their PSMs to the chicken genome

4.3.4. Identification of single amino acid polymorphisms (SAAPs)

4.3.5. Identification of novel genes
4.3.6. Identification of a translated pseudogene..................................................154
4.3.7. Partial and ‘missing’ genes ........................................................................154
4.3.8. Discovery of chicken Aldolase C.................................................................158
4.3.9. Functional annotation of novel genes ..........................................................161

4.4. Conclusion ......................................................................................................162
4.5. References .......................................................................................................163
Appendixes.............................................................................................................178
List of Figures

Figure 1.1. Different types of alternative splicing.................................................................18
Figure 1.2. Mass spectrometry-based proteomics workflow.................................................26
Figure 1.3. Proteogenomics is an iterative process that benefits the “Omics” sciences...........30
Figure 1.4. Database search strategies used for proteogenomics........................................33
Figure 2.1. Exon and Peptide Mapping nomenclature............................................................57
Figure 2.2. Ensembl exon classification.................................................................................61
Figure 2.3 Predicted and experimental peptides that span introns........................................63
Figure 2.4. Relative classification of peptide sets from metazoan proteomes in terms of detection of specific protein isoforms...........................................................................64
Figure 2.5. Peptide evidence for alternative protein isoforms encoded by a hypothetical gene....67
Figure 2.6. Peptide evidence for exon-intron boundaries in the Chicken TPM3 gene.............68
Figure 2.7. Correction of an incomplete model of human NDUFV3 gene.................................72
Figure 2.8. Splicing complexity of the Drosophila para gene.................................................75
Figure 3.1 Schematic of EST translation for target:decoy database generation.........................97
Figure 3.2. Overlap of peptides identified in pairwise database searches..............................100
Figure 3.3. Variation of search statistics with Mascot scores................................................103
Figure 3.4. Estimating the proportion of true positive PSMs identified in the six-frame database search......................................................................................................................105
Figure 3.5. Mascot ion score distribution for target and decoy PSMs.......................................108
Figure 3.6. Percentage of target PSMs of all PSMs reported at a given Mascot score, comparing 6-frame vs. ESTScan2 searches....................................................................................111
Figure 3.7. The six-frame database leads to a greater number of decoy hits, while the target hits remain almost unchanged.................................................................112
Figure 3.8. Effect of database size on FDR of the six-frame PSMs.........................................113
Figure 3.9. Comparison of equivalent PEPs from standard six-frame searches against alternate database searches..................................................................................................................115
Figure 3.10. Amino acid composition of the translated EST contig reading frames...............117
Figure 3.11. EST contig directionality......................................................................................119
Figure 4.1. Proteogenomics workflow summary....................................................................142
Figure 4.2. Exonerate alignment of the signal recognition particle (54kDa) protein sequence predicted by ESTScan2.................................................................................................143
Figure 4.3. Distribution of PSM-identified EST contigs compared with known Ensembl genes located on chicken chromosome 1..........................................................................................145

Figure 4.4. Identification of single amino acid polymorphisms (SAAPs)........................................147

Figure 4.5. Novel peptide identifications improve genome annotation in chicken.........................151

Figure 4.6. Different classes of novel peptides improve genome annotation.................................153

Figure 4.7. EST contig-to-genome mappings by chromosome.........................................................155

Figure 4.8. Alignment of the novel Aldolase C protein sequence with chicken aldolase B and human aldolase B and C..................................................................................................................160

List of Tables

Table 2.1. Peptide data sets calculated or used in this study...........................................................58

Table 2.2. Protein isoforms in Ensembl 48..........................................................................................60

Table 2.3. Relative fractions of gene sets with associated peptide evidence......................................66

Table 2.4. Suitability for alternative peptidases for AS characterisation in human..........................70

Table 3.1. Unique peptide identifications at q-value/PEP < 0.01 for different database searches..................................................................................................................98

Table 3.2. Hypothetical target and decoy PSMs accepted at a fixed score in a standard and inflated database....................................................................................................................109

Table 3.3. Unique peptide identifications at different Quality PEP cut-offs derived from searches over different databases........................................................................................................115

Table 3.4. Kullback-Leibler divergence shows amino acid distribution for the correct reading frame is similar to annotated proteins..........................................................................................117

Table 4.1. BLAST comparisons of EST contig data sets to Ensembl genes..................................141

Table 4.2. Gene models confirmed using BLASTX to map PSM-confirmed-EST contigs to Ensembl protein sequences......................................................................................................144

Table 4.3. Amino acid polymorphisms identified by mapping PSMs identified from ESTs, to Ensembl protein sequences...........................................................................................................148

Table 4.4. Confirmation of SAAPs...................................................................................................149

Table 4.5. Examples of genes missing from the chicken genome assembly......................................157

Table 4.6. Examples of genes only partially present in the chicken genome assembly.......................158
Abstract

Name: Paul David Blakeley
Institute: University of Manchester
Degree title: Doctor of Philosophy
Thesis title: Computational Proteomics for Genome Annotation
Date: January 2013

The field of proteogenomics operates at the interface between proteomics and genomics, and has emerged during the past decade to exploit the vast quantities of high-throughput sequence data. A range of different proteogenomics approaches have been developed, which integrate mass spectrometry data with genome sequence data to provide empirical evidence for protein-coding genes. However, current methods may not be optimized as they do not fully consider the splicing complexity in eukaryotes and there is currently no best practice method. To address this, we investigate the level of proteomics support for Ensembl gene models in human, and a selection of model organisms. We find a disparity between the number of splice variants confirmed by extant data, and the number that can theoretically be confirmed using current proteomics technologies. We then go on to investigate EST-based proteogenomics methods, which enabled the discovery of novel peptide sequences in the chicken genome, which represent hitherto unannotated genes, amended gene models, polymorphisms, and genes missing from the genome assembly. Different approaches for searching mass spectrometry data against transcript sequences are explored, and we show that searching mass spectra against protein sequences predicted by the EORF and ESTScan2 translation tools results in the best sensitivity.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and s/he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

Firstly, I would like to thank my supervisor Prof. Simon Hubbard for giving me the opportunity to conduct this research and providing support and advice throughout the PhD. I would also like to thank past and present members of the Hubbard group: Jennifer Siepen, Lisa Lee, Haslina Hashim, William Rowe, Craig Lawless, David Wedge, King Wai Lau and David Lee who were always willing to help and share their expertise and knowledge.

I thank Jullian Selley for IT support and Mascot server administration. Further, I would like to thank members of the Faculties IT support team who have endeavoured to keep the computer clusters up and running and providing assistance with general IT queries. I am extremely grateful to the BBSRC for funding this research.

Finally, I would like to give special thanks to my family and friends who have been highly supportive throughout the PhD.
The Author

After completing my BSc (Hons) in Biochemistry at The University of Liverpool, I undertook a PhD in Bioinformatics at The University of Manchester. I joined the Welcome Trust Sanger Institute in September 2012, where I am now working on the ENCODE project.
Rationale for an Alternative Format Thesis

The alternative format is the most logical way to present my work because much of the results chapters have been incorporated into papers which have been published or are due to be submitted. The thesis consists of chapters which largely correspond to a publishable paper: an original research paper has been published in Proteomics, and another has recently been published in Journal of Proteome Research. I also plan to submit a third paper within a few months.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARF</td>
<td>Alternative Reading Frame</td>
</tr>
<tr>
<td>AS</td>
<td>Alternative Splicing</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding DNA Sequence</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tag</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>Mass Spectrometry</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>PEP</td>
<td>Posterior Error Probability</td>
</tr>
<tr>
<td>PFF</td>
<td>Peptide Fragment Fingerprinting</td>
</tr>
<tr>
<td>PMF</td>
<td>Peptide Mass Fingerprinting</td>
</tr>
<tr>
<td>pre-mRNA</td>
<td>Precursor Messenger RNA</td>
</tr>
<tr>
<td>PSM</td>
<td>Peptide Spectrum Match</td>
</tr>
<tr>
<td>SAAP</td>
<td>Single Amino Acid Polymorphism</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SRM</td>
<td>Selected Reaction Monitoring</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcriptional Start Site</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non Coding RNA</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Genome sequencing arguably underpins all biological research, and its value is recognised in a range of other disciplines including Earth and environmental sciences (Shinzato et al., 2011), palaeontology (Organ et al., 2007), psychology (Ng et al., 2009) and medicine (Ng et al., 2009). The post-genomic era of biology has witnessed some remarkable achievements. It has permitted the quantitative analysis of an entire proteome (De Godoy et al., 2008), and has helped unravel the molecular basis of diseases such as cancer (Pleasance et al., 2010). The availability of genome-wide data has triggered an increase in data-driven research and systems-level approaches to biology. However, the string of base-pairs that constitutes a genome sequence must first be interpreted in order to obtain useful biological information. Without annotation, interpretation and experimental validation it remains a blueprint for the potential expression of a whole range of biomolecules, including protein-coding genes. Experimental techniques are therefore required to validate, augment and refine the annotated genome and understand how function is produced, and understand how dynamic responses are elicited from essentially a static molecule.

This chapter will examine how genome sequences can be annotated and the role which techniques such proteomics play in annotating and understanding the biological function that the genome encodes.
1.1. Genome annotation

Since the birth of DNA sequencing more than 35 years ago, advances in sequencing technologies have led to an explosion in the number of sequenced genomes, with 2315 different organisms having their genome sequenced (http://www.ebi.ac.uk/genomes/). The total number of DNA base-pairs deposited in the nucleotide sequence databases has increased exponentially over the past decade (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). This rate of growth is set to increase with the advent of next-generation sequencing platforms, which can sequence entire genome sequences in a matter of days/weeks. Making sense of the influx of sequencing data and relating it to other types of biological data is a major focus of research.

Once a genome sequence is available, biochemical and biological information can be assigned to the sequence of nucleotides – the process of genome annotation. This can be achieved by using methods that predict genes based on (i) statistical analysis of features and signals contained within the genome sequence (i.e. ab initio), (ii) relating experimental derived sequence data with the genome sequence, and (iii) comparative approaches using data from evolutionary related species. Gene annotations often use multiple methods, whereby information from one method can improve the accuracy or sensitivity of other methods. Gene annotation can be automated (high-throughput) or can be achieved manually (high fidelity), but ideally should be automated for consistency and speed while achieving a level of accuracy comparable with manual annotations. The process of genome annotation is twofold, involving both structural and functional annotation. First, the genomic sequence is surveyed to locate genes (structural annotation). The task of locating genes is complicated by the fact that only ~1% of the human genome is protein-coding (ENCODE Project Consortium, 2007), and much of the DNA is apparently non-functional (or having an unknown function). Once a gene is located, the exon-intron structure for all its putative mRNA transcripts is predicted to generate a gene model. It is highly important that initial gene model predictions are accurate because annotations that are based on an existing incorrect gene models will result in a propagation of errors, something that would have implications for future experiments that require a priori knowledge of gene models.
1.1.1. Accurate genome annotation requires experimental evidence

Determining the exact structure of a gene is often more difficult than gene finding. Computational predictions have found most of the human protein-coding genes, but it is estimated that as many as 50% of these have an incorrect genomic structure (Guigo et al., 2006). Gene boundary, open reading frame and exon-intron boundary predictions have improved during the post genomic era, but there are still many gene models that are incomplete or erroneous, even in well annotated genomes such as human or mouse. In addition, genes that are short in length or have atypical codon usage are often missed by gene prediction algorithms (Warren et al., 2010).

The presence of non-coding introns within most eukaryotic genes complicates the task of predicting gene models. In the simple unicellular eukaryote, *S. cerevisiae*, only ~4% of genes are interrupted by an intron (Spingola et al., 1999). But the proportion of interrupted genes increases to ~94% in mammals (Wang et al., 2008). The introns are transcribed, along with the coding exons, to generate pre-mRNA transcripts. During the maturation of a pre-mRNA transcript, a complex set of protein-protein, protein-RNA, and RNA-RNA interactions result in the removal of introns and ligation of exons, either during or after transcription. The splicesomal complex initiates this process by recognition of signals within the intron, including the 5' and 3' splice sites, which are usually 5'-GU-3' and 5'-AG-3' respectively. Computational prediction of splice sites using *ab initio* methods can be tricky because sequences that are identical to the splice sites occur frequently in eukaryotic genomes. For example, 5' splice sites can occur ~400 million times in the haploid human genome of ~3 billion nucleotides (Sonnenburg et al., 2007). Adjacent 5’ and 3’ splice sites are not always used and different combinations of splice sites can be used through alternative splicing to produce a range of different mRNAs from the same pre-mRNA. Furthermore, splice sites are not absolutely defined, and are instead marked by consensus sequences. Splice site usage depends on a combination of multiple ‘weak’ signals from interactions between *cis*-acting DNA elements and *trans*-acting factors, which can be categorised into serine-arginine-rich (SR) proteins and heterogeneous ribonucleoprotein particles (hnRNPs) (Mayeda and Krainer, 1992; Zhang and Chasin, 2004; Zahler and Roth, 1995; Smith and Valcárcel, 2000). This can result in several different types of splicing patterns occurring for a given gene model (Figure 1.1). The mechanisms responsible for choosing splice site combinations are not fully understood,
and the choice of splice site pairs depends on stochastic processes (Melamud and Moult, 2009). Therefore it would be naive to rely solely on \textit{ab initio} gene model predictions. There has recently been a concerted effort at deciphering the splicing code for the prediction of tissue-specific exon skipping (Barash \textit{et al}., 2010). However, this is ultimately reliant on a set of accurate exon coordinates, and its utility is restricted to the prediction of mammalian exon skipping events. This highlights the requirement for experimentally determined exon boundaries to generate accurate gene models.

Alternative splicing of pre-mRNAs introduces an enormous amount of diversity into the transcriptome. A recent estimate suggests that approximately 95\% of human multi-exon genes undergo alternative splicing (Pan \textit{et al}., 2008), which means splice variants are much more prevalent than the current sequence databases would suggest.

The conundrum of deciding which pairs of splice sites are used to generate the mature mRNAs complicates the task of genome annotation. Indeed, predicting gene models is not a trivial endeavour because in extreme cases a single eukaryotic gene can potentially encode up to \(~38,016\) distinct mRNA transcripts (Schmucker \textit{et al}., 2000), and a surprisingly large combination of features are thought to be responsible for splice site selection (Barash \textit{et al}., 2010). Gene prediction algorithms can correctly identify an exon more than 95\% of the time, but show poor specificity in terms of identifying the exact exon boundaries (Reese \textit{et al}., 2000). Moreover, other forms of post-transcriptional regulation complicate matters, such as alternative polyadenylation (Zhang \textit{et al}., 2005), alternative transcriptional start sites (Roni \textit{et al}., 2007), and RNA editing (Hodges \textit{et al}., 1991).
Figure 1.1. Different types of alternative splicing.

a) Use of an alternative 5’ splice site demonstrated in the *Drosophila fruitless* gene. The female pattern of splicing involves the use of a downstream 5’ splice site, whereas males use an upstream 5’ splice site which reveals an upstream start codon. Red region represent exonic splicing enhancers. b) Exon skipping demonstrated in the human *PIGF* gene. Exon 6 may not be recognised, resulting in its exclusion. c) Intron retention in the 5’UTR demonstrated in the *Drosophila msl-2* gene. The intron is not removed in females, thus facilitating RNA export into the cytosol. d) Mutually exclusive exon skipping in human *FGFR2* results in isoforms with different ligand specificity. e) Complex intron retention in the human *AIF-1* gene. Figure adapted from (Smith and Valcárcel, 2000; Hiller et al., 2005; Sammeth et al., 2008)).
Analogous to splice sites, sequences resembling start and stop codons occur frequently in the genome, but most are never bound by ribosomes. Hence, a predicted ORF is not always equivalent to a coding DNA sequence (CDS) (Hillman et al., 2004) and annotating true protein coding ORFs from transcriptome data is not trivial either. For example, it is generally held that longer ORFs are more likely to encode proteins, but there are some cases where shorter ORFs are functional (Kondo et al., 2010). Hence gene prediction algorithms are usually biased against short ORFs (Warren et al., 2010), which instead require detection and validation through experiments. The correct identification of the start and end positions of a gene presents an even greater challenge. Difficulties in predicting untranslated regions (UTRs) have meant that gene predictors ignore UTRs, and instead tend to predict CDSs (Kong et al., 2007). The position of a transcriptional start site (TSS) for instance, is dependent on the superstructure of the chromatin, which in turn is dependent on histone modification/cytosine methylation patterns (ENCODE Project Consortium, 2007). Hence, gene models and alternatively spliced transcripts cannot, at present, be determined by analysis of DNA sequence signals/features alone and the task of genome annotation requires data from additional high-throughput experiments. The next section will consider some of the commonly used genome annotation pipelines, all of which rely on experimental data.

1.1.2. Genome annotation pipelines map experimental data to the genome

The major genome annotation projects utilise transcriptome sequence data as well as protein homology information to build gene models. The manually curated Vertebrate Genome Annotation (VEGA) database (Wilming et al., 2008) creates high quality gene annotations based on full-length sequences of cDNA inserts, Expressed Sequence Tags (ESTs) and protein alignments. All genes within VEGA are manually annotated and are therefore of the highest quality, but are limited to a subset of genes in human, mouse and zebrafish. Automated genome annotation pipelines are more comprehensive and typically use \textit{ab initio} gene prediction as a framework from which the final gene model is mapped using experimental data. Augustus (Stanke et al., 2006) is one of the most accurate gene prediction tools available (Guigo et al., 2006), which can optionally incorporate experimental data to improve the gene model prediction. The Ensembl database (Potter et al., 2004), the National Center for Biotechnology Information (NCBI) browser (Pruitt et al.) and the University of California Santa Cruz (UCSC) database (Kent et al., 2002) use
automated annotation pipelines which rely on transcript sequence-based evidence from various databases in conjunction with gene prediction algorithms. Ensembl provides a genome annotation resource for 58 different organisms (as of release 67) and displays gene models with associated metadata on an intuitive genome browser. Ensembl does not utilise ESTs owing to their high error content. Instead it uses both the Genewise (Birney et al., 2004) and Exonerate (Slater and Birney, 2005) algorithms, which are used to predict gene models by aligning protein and full-length cDNA sequences with the genome sequence. Ensembl also displays “GENCODE” gene models, resulting from the ENCODE projects work in characterising all functional elements of the human genome (ENCODE Project Consortium, 2007). GENCODE is a subsidiary project of ENCODE, and reconstructs high quality gene models through targeted experiments and manual curation (Harrow et al., 2006). GENCODE provided a reference set of genes, which were used to evaluate gene models through the ENCODE genome annotation assessment project (EGASP) (Guigo et al., 2006). One of the main conclusions of EGASP was that programs based on mRNA and protein sequence alignments produce the most accurate gene annotations (Guigo et al., 2006).

Initially, full-length cDNA sequences were considered to be the ‘gold standard’ resource for genome annotation (Brent, 2008). However, such cDNA databases are relatively small, especially for non-human genomes (Guigo et al., 2006) which somewhat restricts their utility. On the other hand, the NCBI dbEST contains over 8.5 million human, 4.8 million mouse, and 0.6 million chicken EST entries (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Providing they are assembled accurately into transcripts with a stringent quality control step, ESTs remain a valuable potential resource for producing gene models. Moreover, RNA-Seq data is now routinely acquired and exploited for the purpose of genome annotation, whereby short reads are mapped to the genome to build novel gene models (Denoeud et al., 2008). The superior coverage and depth provided by RNA-Seq is now being exploited by the ENCODE project, and such data can be visualised in genome browsers (‘ENCODE whole genome data in the UCSC genome browser’). In light of this, the RNA-Seq genome annotation assessment project (RGASP) has been initiated which aims to assess the quality of transcript predictions based on the mapping and assembly of reads against the genome sequence (http://www.gencodegenes.org/rgasp/). The whole process of genegenome

20
annotation remains one of continual improvement and flux as novel technologies become available and genome scientists continue to develop better integrated approaches, learning the lessons from the various benchmarking exercises and large-scale projects such as those described above.

Gene annotation pipelines (e.g. Ensembl) also use evidence from protein sequence databases. The sequencing of full and partial length cDNAs and their subsequent alignment to the genome can predict the location and structure of a protein coding gene. However, protein evidence from the same species or a closely related species can unequivocally confirm whether or not a gene is protein-coding, while also improving the gene model. The protein evidence used by Ensembl is from the UniProtKB/Swiss-Prot manually curated database (Hubbard et al., 2007). The Swiss-Prot portion of the UniProt database contains non-redundant protein sequences based on experimental evidence acquired over many years, including more traditional biochemical approaches such as Edman sequencing, which in turn is curated manually by experts (The UniProt Consortium, 2008). Hence, Swiss-Prot sequences are considered to be of the highest accuracy, but the database is relatively small compared with other protein sequence databases such as Entrez Protein and is therefore confined to a subset of well-studied genes. This highlights the potential utility of high-throughput proteomics data, which in conjunction with computational pipelines, can be used to expand protein-level verification to gene models in a completely automated way.

1.1.3. The need for protein-level data

The field of proteomics has much to offer for genome annotation because it can directly confirm protein-coding genes and detect novel protein-coding genes missed by other approaches. Differentiating between genes that encode proteins and those that do not, is a fundamental aspect of gene annotation which is not as simple as it may seem and is fundamentally important when considering genome complexity. The human genome contains between 10,000 to 20,000 non-functional genomic sequences derived from the duplication of functional genes or from the retrotransposition of mRNAs and subsequent integration of the cDNA into a new genomic locus (Zhang et al., 2003) (Torrents et al., 2003) (Zhang and Gerstein, 2004). These are called, respectively, non-processed and processed pseudogenes. Pseudogenes are copies of protein-coding genes that are no longer
translated into functional proteins, but in some cases regulate the activity of their homologous functional counterpart(s) (Hirotsune et al., 2003). Unprocessed or duplicated pseudogenes also include unitary pseudogenes, which lack functioning counterparts, but have functional orthologs in related genomes (Zhang et al., 2010). All of these create a problem for gene annotators because they can share high sequence similarity with functional gene sequences and can be mistakenly identified as genes in both ab initio based and experimental based annotation pipelines (Zhang et al., 2003) (Van Baren and Brent, 2006; Mounsey et al., 2002).

Even cDNA-based evidence is insufficient when trying to distinguish genes from pseudogenes because there is growing evidence that suggests transcription of pseudogenes is widespread (Poliseno et al., 2010; Watanabe et al., 2008). It turns out that some pseudogene RNAs and small-interfering RNAs derived from pseudogene RNAs have biological roles in gene expression, possibly operating through RNA interference (Hirotsune et al., 2003; Tam et al., 2008). Hence the distinction between genes and pseudogenes is not always clear-cut. Indeed, functional genes can be mistakenly annotated as pseudogenes, something which has been observed even in well-annotated genomes such as mouse (Brosch et al., 2011). To illustrate the utility of a proteomics approach using mass spectrometry, it was shown that 33 different C.elegans pseudogenes are in fact likely to be protein-coding genes (Merrihew et al., 2008).

Further problems for annotation pipelines arise from the fact that the majority of the human genome is believed to be transcribed into RNA (ENCODE Project Consortium, 2007). Non-coding RNAs (ncRNAs) constitutes a significant proportion of the transcriptome and have diverse roles in gene expression including the chemical modification of RNA, translational regulation, RNA splicing and genomic imprinting. Some ncRNAs can resemble mRNAs in terms of their length, chemical modification (i.e. polyadenylation) and features such as ORFs (Tupy et al., 2005). Moreover, ncRNAs are present as “contaminants” in cDNA libraries (Tupy et al., 2005), which can lead to ncRNAs being incorrectly assigned as protein-coding genes. Indeed, it has been shown that ~10% of Swiss-Prot protein sequences are actually derived from ncRNAs (Frith et al., 2006). Since annotation pipelines such as Ensembl use cDNA and protein evidence (e.g. from Swiss-Prot), this implies that protein sequence databases contain translations from non-coding RNA sequences.
Distinguishing between protein-coding and non-coding sequence is important for genome annotation. Protein sequence evidence could replace previous strategies used to distinguish mRNAs from ncRNAs such as ORF length and ORF conservation. Indeed, the latter strategies are not always reliable (reviewed in (Dinger et al., 2008)) and a targeted direct approach using mass spectrometry to identify the protein products would be preferable. It should be pointed out that the absence of a predicted peptide or protein in a mass spectrometry experiment, however, is not proof that the parent molecular species is not present for a whole variety of reasons (sequencing errors, polymorphism, post-translational modifications, sensitivity, non-expression in given conditions etc).

Another important role for proteomics is to provide evidence for the translation of splice-variants. The inclusion/exclusion of cassette exons, or the use of alternate 5’/3’ splice sites can introduce a premature stop codon (PSC) into the mRNA. The stop codon may be located within the alternative exon itself, or an alternative exon may introduce a frameshift, thus changing all downstream codons. If the PSC lies sufficiently upstream from the stop codon, and the mRNA is translated, then it may become targeted by the nonsense mediated decay (NMD) pathway, and subsequently degraded (Stoilov et al., 2004). Therefore alternative splicing can regulate the abundance of mRNA transcripts (and encoded proteins) by generating transcripts which serve only regulatory functions. The low abundance of mRNAs targeted by NMD means that such transcripts are underrepresented in cDNA libraries, but they may still be incorporated into gene models. Therefore proteomics data is needed in order to confirm that low abundance transcripts are indeed precursors to functional protein products. Chapter 2 addresses this by confirming the translation of splice variants that exist in the Ensembl gene annotations for human and several model organisms.

1.2. Mapping the proteome

The main role of the genome is to encode the information necessary to generate functional entities that carry out biological function, most commonly proteins, which are the main coordinators of biological processes. The aim of proteomics is to study the full set of proteins expressed under a particular biological condition or locality within the organism, ideally in a quantitative fashion. Until recently, only a subset of most proteomes
could be studied in a single experiment via mass spectrometry. But technical innovations in mass spectrometry and separation science, such as improvements in accuracy and resolution (Mann and Kelleher, 2008), have meant that near comprehensive (De Godoy et al., 2008; Picotti et al., 2009) and reproducible (Picotti et al., 2009; Picotti and Aebersold, 2012) analysis of the proteome is now achievable. The integration of these data with the genome increases the comprehensiveness of the proteome analysis by allowing the identification of proteins that are not present in existing protein sequence databases, which typically have stringent acceptance criteria when building gene models.

1.2.1. High-throughput proteomics

Mass spectrometry-based proteomics does not usually involve analysis of the proteins directly. Instead mass spectrometers (referred to as ‘instruments’ hereon in) are used to detect surrogate peptides resulting from an enzymatic digest, usually carried out using proteolytic enzymes such as trypsin. In high-throughput proteomics experiments, the high complexity of the digested peptide mixture means that multidimensional protein identification technology (MudPIT) (Washburn et al., 2001) is often applied to the mixture prior to mass spectrometry analysis. MudPIT increases sensitivity by separating out the complex peptide mixture based on some physicochemical property such as charge or hydropathy to reduce the complexity of the analyte mixture entering the MS stage and improving the chances of detection of low-abundance proteins. In most modern instruments, the eluting peptide mixture is introduced into the instrument through either electrospray ionisation (Fenn et al., 1989) or Matrix-Assisted Laser Desorption Ionisation (Spengler et al., 1992). Figure 1.2 describes a generic mass-spectrometry workflow. After the ionisation step, the resulting precursor ions are then analysed using one (or a combination) of three main methodologies: peptide mass fingerprinting (PMF), peptide fragment fingerprinting (PFF), and de novo sequencing. PMF involves measuring the mass-to-charge (m/z) ratio of each precursor ion (i.e. peptide) and is normally applied to samples cut from SDS-PAGE gels, which represent a single or simple mixture of proteins that have been subsequently treated with proteolytic enzyme. Robotic systems have been deployed to speed up this process to conduct high-throughput experiments. However, a single precursor ion mass measurement often fails to differentiate between candidate
peptides which have similar theoretical masses, even when instruments capable of high mass accuracy measurements are used. Although PMF will usually capture several precursor ions from a single protein, this is generally not considered to be a reliable high-throughput technique for proteomics. Instead, in modern high-throughput proteomic experiments, LC-MS/MS is more typically employed. In this case, a PFF stage is performed in a second mass analyser where the precursor ion is fragmented into its constituent ions using either Electron Capture Dissociation (ECD) or Collision-induced Dissociation (CID). Each fragmentation technique promotes the generation of ions via cleavage at specific bonds in the polypeptide chain to generate a characteristic ion series. For example, CID generally promotes the generation of b and y ions which are subtended at the amino and carboxy terminus of the peptide ion respectively. These ion groups of m/z peaks represent the experimental spectra (tandem MS spectra) that are compared to theoretical spectra derived from a database of protein sequences using a database search algorithm, such as Mascot (Perkins et al., 1999), which result in candidate peptide-spectrum matches (PSMs). This database search approach is the most common method for peptide identification. However new algorithms are being developed to combine de novo peptide sequencing with database searching by generating short sequence tags directly from spectra, which are then used to limit the database search space (Zhang et al., 2012; Bern et al., 2007).
Figure 1.2. Mass spectrometry-based proteomics workflow.

A sample of proteins is obtained via the fractionation or affinity purification (1). Optionally, a subset of proteins can be extracted using gel-electrophoresis. The sample is then digested with an endopeptidase, usually trypsin (2). Peptides are separated by multidimensional liquid chromatography and introduced into a MALDI or Electrospray ion source (3). An electrical field is applied in the mass analyser, and the time taken for an ion to reach the detector is used to calculate the mass-to-charge ratio (4). In tandem mass spectrometry (5), ions are introduced either stochastically (data-independent) or selectively (data-dependant) into a collision cell where they undergo fragmentation by high energy collision with a gas. The fragment ions are introduced into a second mass analyser, which typically measures the time-of-flight or oscillation frequency (e.g. Orbitrap, FTICR) to record the mass-to-charge ratios (5). Spectra obtained in both (3) and (4) are then used to identify the peptides present in the sample. (Adapted from (Aebersold and Mann, 2003)).

1.2.2. Targeted Proteomics

Ideally, high-throughput proteomics should identify all of the proteins present in the sample. However, typically only a subset of the sample is identified. This can be attributed to factors such as low expression levels (e.g. transcription factors, splice
variants), inaccessible or hydrophobic proteins (e.g. transmembrane proteins), instrument performance (De Godoy et al., 2006), and a lack of detectable ‘proteotypic’ peptides (Mallick et al., 2007). Generally, sampling of the proteome via high-throughput proteomics is biased towards the most abundant proteins because precursor ions derived from such proteins are more likely to be selected for fragmentation at the MS2 stage. Most experimental set-ups in even the most sophisticated laboratories are currently struggling to acquire high quality data for the majority of the proteins expressed in a cell from a complex metazoan cell, though technology is continually improving. In contrast to high-throughput proteomics, targeted proteomics based on selected reaction monitoring (SRM) can sample the full dynamic range of a eukaryotic proteome (Picotti et al., 2009). By utilising triple quadrupole instruments, SRM can provide measurements of low-abundance proteins by selecting a specific precursor and product ion(s) to monitor through the three quadrupoles. However, it is likely that there will still be some proteins that cannot be detected using this approach (Eyers et al., 2011) because their constituent peptides have certain physiochemical properties that render them unsuitable for ionisation (Li et al., 1999) or they still remain beyond the reach of current sensitivity, which is approximately 50 copies per cell (Picotti et al., 2009).

However, for most proteins there are some peptides that are consistently observed over a number of mass spectrometry experiments, and therefore act as reproducible surrogates for the protein in question. These have been termed ‘proteotypic’ peptides (Mallick et al., 2007). They would be expected to be unique to the protein in question, to ionize well in the mass spectrometer, and to generate a distinct set of fragment ions that lead to a good identification. Ideally, they should also be devoid of post-translational modifications. A number of classifiers have been developed to predict whether or not a peptide is likely to be proteotypic (Mallick et al., 2007; Eyers et al., 2011; Fusaro et al., 2009). CONSeQuence (Eyers et al., 2011), which combines random forest, SVM, neural network and genetic programming based classifiers, was applied to the yeast proteome and predicted that in most cases there are less than ten proteotypic peptides per protein (Eyers et al., 2011). Moreover, the majority of these proteotypic peptides are not unique to specific splice-variants, thereby rendering them useless for the targeted analysis of protein isoforms.
Targeted approaches are useful for focusing on these proteotypic peptides to enable sensitive and reliable protein identification and these prediction algorithms are extremely informative when designing targeted proteomics experiments. Chapter 2 investigates the prevalence of proteotypic peptides in different organisms, and specifically addresses the problem of identifying splice variants through targeted proteomics experiments.

### 1.2.3. Peptide Identification via database searching

Database search algorithms (Tanner et al., 2005; Eng et al., 1994; Perkins et al., 1999; Craig and Beavis, 2004) are essential for high-throughput proteomics as they allow the assignment of thousands or even millions of spectra collected from tandem MS experiments to peptide sequences resulting from an in silico enzymatic digest of a protein sequence database. The identity of proteins present in the sample is inferred from the set of peptide sequences with spectral matches. The candidate peptides are scanned using the experimentally calculated precursor mass, and are matched if the precursor mass matches the theoretical mass within a fixed tolerance window. Statistical or probabilistic methods are then used to measure the similarity between the observed tandem spectra and theoretical spectra. Typically only the best matches are reported for each spectrum. For example Mascot (Perkins et al., 1999) calculates the probability that the match between the calculated masses and the theoretical masses is a random event, whereas Sequest (Eng et al., 1994) calculates the cross-correlation between the two masses. Although such scores are useful for relative PSM ranking, they fail to properly classify correct and incorrect PSMs if used directly, and cannot be reliably compared across experiments (States et al., 2006). This is also true when transformed into apparently equivalent metrics such as expectation values (E-values) that are typically reported by most search engines. Therefore, the error rates cannot be inferred directly from the search algorithm-specific score and instead additional statistical measures have become popular and are applied, namely the False Discovery Rate (FDR) and the Posterior Error Probability (PEP).

The p-value is one of the most widely used measures of statistical significance in biology. However proteomics, like other disciplines which handle large-scale data, tend to rely on FDRs to correct for multiple testing. The FDR in proteomics is defined as the expected proportion of incorrect PSMs within a set of PSMs. However, because the FDR is
not a property of an individual peptide score, it is usually converted into a q-value (Storey, 2002) which ensures that increasing the peptide score will decrease the FDR. The FDR can be estimated using the target-decoy approach (Elias and Gygi, 2007; Kall et al., 2008a). This typically involves searching the spectra against a decoy database containing randomised, shuffled, or reversed sequences, along with the standard target database, to estimate the proportion of incorrect PSMs based on the number of hits to the decoy database. Spectra can be searched against either a concatenated target-decoy database, or separately searched against the target and decoy databases. In either case, a FDR can be calculated from the resulting lists of PSMs taking the top-scoring match for each spectrum.

Another useful property of the FDR is that it allows for multiple testing and represents a single confidence estimate for a collection of PSMs. Even though p-values can be corrected to deal with multiple comparisons, this becomes an overly conservative measure of significance when large datasets are used. However, one problem with the FDR is that all PSMs accepted at a specified FDR will be treated equally, even though some will have a greater chance of being (in)correct than others depending on the specific score assigned to them by the search engine. Therefore when a specific PSM is of interest, in validating a gene model for instance, then researchers need to know the likelihood that a particular PSM has been correctly assigned. In this case the PEP, or ‘local FDR’ is useful, which is defined as the probability that a PSM is incorrect (Kall et al., 2008b). Qvality (Kall et al., 2009) and PeptideProphet (Keller et al., 2002) are two commonly used programs used to calculate the PEP, both of which use a mixture modelling approach to assign PEPs calculated from the distribution of target and decoy PSM scores.

Controlling the error rate is important to ensure that maximum sensitivity is achieved whilst maintaining sufficiently high accuracy and not accepting large numbers of false positives. This is crucial when searching proteomics data against large genome or transcript databases, which can be highly redundant, error-prone and littered with non-coding regions. Indeed, excessively large databases can influence the scoring scheme by increasing the number of candidate decoy peptides (Granholm and Kall, 2011). In chapter 3, we investigate how database composition can influence the statistical modelling for different types of EST-based proteogenomic databases, and propose an optimal approach for searching proteomics data against transcript sequences.
1.3. Proteogenomics

As has already been discussed, proteomics is a rich source of data that can be used to improve the annotation of protein-coding genes. The two sources of data (proteomics and genomics) are in principle constantly ‘evolving’ as proteomics data can be used to refine gene predictions, and these more accurate gene models can allow more protein identification to be made (Figure 1.3). However, there is a complex relationship between the two since it is impossible to discover novel gene structure via a database-directed proteomics approach if the new sequences are not in the database being searched. Additionally, given the generally conservative nature of most gene prediction pipelines and the reluctance to use transcriptional data such as ESTs directly, proteogenomics has taken a while to catch on and be taken up by the community. The next section describes some of the approaches that have been taken.

Figure 1.3. Proteogenomics is an iterative process that benefits the “Omic” sciences.
1.3.1. Introducing proteogenomic approaches to genome annotation

After a relatively slow start, proteomics data is now becoming more widely used for genome annotation (Desiere et al., 2005; Gupta et al., 2007; Tanner et al., 2007; Castellana et al., 2008b; Merrihew et al., 2008). There are three main ways in which MS-based proteomics can be applied in the genome annotation process. One approach is to confirm existing genome annotations (Buza et al., 2007; Bringans et al., 2009), which can also resolve conflicts between competing gene models (Wright et al., 2009). A second approach involves correcting previously-annotated gene models, typically involving the identification of peptides within introns or UTRs (Merrihew et al., 2008), as well as the identification of novel exon skipping events (Mo et al., 2008). Finally, a third approach involves the identification of completely novel genes (Findlay et al., 2009).

Several established proteomics analysis pipelines and data repositories, such as the Genome Annotating Proteomics Pipeline (GAPP) (Shadforth et al., 2006) and PeptideAtlas (Desiere et al., 2005), can map peptide identifications onto a genome, allowing them to be displayed as an additional track of visual evidence in the Ensembl genome browser (Potter et al., 2004). These data provide experimental validation of existing annotations. PeptideAtlas is the most widely used and comprehensive of the two pipelines, and provides a proteogenomic framework in which statistically significant peptide identifications are retained after processing PSM lists using PeptideProphet (Keller et al., 2002). Significant peptide identifications are then compared with protein sequence databases (e.g. Ensembl) and inherit genome coordinates based on this match. Several proteomes have been mapped in this way including human, yeast (Saccharomyces cerevisiae), Drosophila melanogaster and Caenorhabditis elegans (King et al., 2006; Desiere et al., 2005; Merrihew et al., 2008; Schrimpf et al., 2009)). More recently, Cow (Bislev et al., 2012) and Honey Bee (Chan et al., 2012) builds were introduced, with the latter being the first PeptideAtlas build to incorporate novel peptides identified directly from the genome sequence.

Computational gene predictions, in conjunction with cDNA/EST alignments, have been successful in identifying the majority of protein-coding genes present in sequence databases. However, although such methods can be accurate, they do not always provide a definitive proof of gene structure. Transcriptional evidence can only imply that a gene is likely to be protein-coding, and as mentioned earlier, gene models often contain errors.
Hence, proteomics is useful as it can provide additional evidence to validate predicted gene models. Ideally a gene model is typically validated by high-confidence peptides within an acceptable FDR (less than 0.01) or with a low probability of being incorrect (PEP<0.05 or better). In eukaryotes, intron-spanning peptides are particularly useful since they can confirm predicted splicing events and delineate the complete splicing patterns responsible for splice variants. Intron-spanning peptides typically constitute ~25% of trypsin-digested peptides, and in human, more than 11,000 splice events were confirmed by mapping these against putative exons (Tanner et al., 2007). A recent study in Drosophila confirmed the presence of alternative protein isoforms in more than 100 genes (Tress et al., 2008). If alternative splicing is as prevalent as the transcriptomic studies suggest, then targeted proteomics experiments could potentially serve an important role to confirm thousands of putative protein isoforms.

Proteomics can also be used to rectify existing gene models. A recent analysis of the human transcriptome showed that 8% of RNA-Seq reads for polyadenylatated RNAs, map within introns (Ameur et al., 2011). Whilst most reads mapping to intronic sequences are likely to correspond to novel exons, many can also represent unspliced mRNAs (Wetterbom et al., 2010), or even intragenic ncRNAs (Huang et al., 2010). Therefore, additional evidence is needed to confirm the expression of intragenic regions. Proteome peptide identifications that map to intragenic regions are useful in this respect, as they can identify novel protein-coding exons. Hence, there is an important role for proteomics in correcting existing gene models by searching MS data against genome sequences (Kalume et al., 2005; Merrihew et al., 2008; Schrimpf et al., 2009). Chapter 4 examines how proteogenomics can identify intragenic peptides, and exemplifies this with the discovery of novel exons in the chicken NAA25 gene.

As stated earlier in the chapter, the annotation of pseudogenes can benefit from proteomics data and this has been demonstrated in C. elegans (Merrihew et al., 2008). The C. elegans genome was the first multicellular eukaryote to have its genome fully sequenced and is at a mature stage in the annotation process. However, thirty-three annotated pseudogenes in C. elegans were shown to be translated protein-coding genes (Merrihew et al., 2008). Because pseudogenes can share similar sequence features with protein-coding genes, there are likely to be many more false negative predictions that can be corrected using proteomics data. Other proteogenomic studies in Yersiniae (Schrimpe-
Rutledge *et al.*, 2012), Arabidopsis (Castellana *et al.*, 2008a) and mouse (Brosch *et al.*, 2011) have detected translation of pseudogenes, indicating that either the “resurrection” of pseudogenes has occurred, or alternatively that genome annotation pipelines are prone to the misclassification of pseudogenes.

Searching spectra against existing gene models restricts the discovery potential of proteomics. There are a number of ways to identify novel peptides and this largely depends on the nature of the database that is searched. Figure 1.4 shows a number of alternative databases used in proteogenomics. There is a tradeoff between discovery potential and search sensitivity, and choosing the database that gives the best balance is crucial in proteogenomics due to the large search spaces involved. At one extreme, one can search spectra against six-frame translations of the entire genome, which represents arguably the largest possible database. These searches are expected to offer the largest loss of search sensitivity. At the other extreme, searches against existing gene models provide no possibility of discovering novel genes or novel gene structure.

![Figure 1.4. Database search strategies used for proteogenomics.](image)

At one end of the scale, searching proteomics data against known gene model can confirm the translation and provide functional ‘protein level’ annotation (e.g. protein localization). At the other end of the scale, six-frame translation of an entire genome offers to discover completely novel genes, but at the expense of search sensitivity at a given false discovery rate.
A common role for proteogenomics is the discovery of novel genes by searching proteomics data against the entire genome. This is usually performed by translating the genome sequence in all six reading frames, and is therefore not biased towards existing, potentially erroneous, gene annotations. Such methods have been demonstrated to be highly effective in annotating bacterial genomes (Gupta et al., 2007), probably reflecting the fact that these genomes are relatively small and are not subject to complex post-transcriptional processing. These methods have also been successfully applied to the discovery genes expressed in seminal fluid in *Drosophila*, where false positives were eliminated through the experimental verification of novel gene models (Findlay et al., 2009). However, without thorough experimental verification, such methods may not be reliable for large metazoan genomes because the massive search spaces involved can reduce sensitivity at a specific FDR. This may be resolved through the use of positional proteomics, by collecting only mass spectra for N- or C- terminal peptides and searching a database containing potential terminal peptide sequences (Mcdonald et al., 2005), which can dramatically reduce the search space. An alternative strategy for reducing the search space involves using *de novo* gene finding algorithms to consider all putative genes, or by searching against transcriptome sequences.

### 1.3.2. Proteogenomics using EST databases

Most gene/genome annotation pipelines currently use physical evidence in the form of full-length cDNA inserts and/or expressed sequence tags (ESTs) to augment the main evidence derived from protein sequence alignments to predict gene models (Ashurst et al., 2005). ESTs can be utilized to define exon boundaries and predict alternative splicing patterns. They provide a useful proxy for the gene and protein sequence, since they represent the intermediate molecule, which is translated into protein. However, ESTs also contain a noncoding UTR sequence at their 5′ and 3′ ends, and are subject to high error rates. This can take the form of single base changes, inserts, deletions, or completely artefactual chimeric fusions of two unrelated sequences. These errors can be eradicated by merging ESTs into consensus sequences to achieve greater coverage of the transcript (Huang and Madan, 1999), a process that itself can create more problems by removing biologically important sequence variants. Keeping such errors to a minimum is important.
because even a single base change can completely alter the translated amino acid sequence and eradicate potential PSMs. Such errors can also be reduced by aligning the ESTs to the genome, either as part of the assembly process, or after assembly into transcripts using tools such as EXONERATE (Slater and Birney, 2005) or the BLAST-Like Alignment Tool (BLAT) (Kent, 2002). The mapped ESTs can then be compared against the genome sequence allowing the detection of artifacts (e.g. chimeric transcripts). This mapping step also allows the assignment of genomic coordinates for peptides identified within the ESTs, which can be used to classify novel peptides as either intergenic or intragenic (see Chapter 4).

Searching spectra against a six-frame nucleotide translation of an entire genome is generally not applicable to the large, intron-rich genomes of animals. In these cases, searching against EST and cDNA libraries represents a useful intermediate approach. The alignment of transcript sequences onto the genome can significantly reduce the search space because mRNAs mostly correspond with protein-coding regions of the genome. This is especially true of vertebrate genes, which are composed of relatively short exons and large introns. Hence, by searching mass spectra against transcript sequences, false positives are likely to be reduced because the database is largely composed of protein-coding sequences. This is explored in chapter 4, which details our annotation of the chicken (Gallus gallus) genome by searching against genome-mapped transcripts.

In order for transcript sequences to be useful for MS-based proteomics, they must be converted into protein sequences. This is complicated by the fact that the transcript sequence may contain errors that could prevent spectra being matched with the resultant protein sequence. Programs have been designed to overcome such problems, such as ESTScan2 (Iseli et al., 1999), which are able to read low-quality DNA sequences to generate the most likely protein translation. When spectra are searched against these protein sequences, novel splice variants, polymorphisms, and even genes can be discovered. For example, this method has been successfully applied to the animal parasite Fasciola hepatica (Robinson et al., 2009). Proteins involved in migration through host tissues were investigated by searching MS/MS data against EST data, uncovering a range of proteolytic enzymes, including novel trypsin-like serine proteases and carboxypeptidases.
The use of transcript sequences as a platform for proteomics-based genome annotation was explored in human sequences as early as 2001 (Choudhary et al., 2001). This feasibility study demonstrated the advantages of searching complex LC-MS/MS data against an extensive collection of human ESTs and exemplified their utility for annotating newly sequenced (poorly annotated) genomes. For example, ESTs are often highly redundant, which could allow the identification of SNPs. One recent proteogenomics study identified peptide variants in the transcriptome of a colorectal cancer cell line that represent SNPs that were absent from the dbSNP (Wang et al., 2011).

ESTs can also be used to identify novel splice variants, something which is difficult when searching against raw genome translations due to the inability to capture intron-spanning peptides. Moreover, ESTs are often derived from cDNA libraries representing many different tissues and developmental stages (Bonnet et al., 2008), a useful feature when studying splice variants with restricted expression patterns. One proteogenomic study searched MS/MS data against ESTs derived from human platelets to provide a catalogue of “in-frame” exon skip events that are relevant at the protein level (Power et al., 2009). These methods give concurrent protein and transcript-level support to improve gene models. Indeed, the proteomics data can provide orthogonal evidence to compensate for the sequencing and assembly errors present in EST data.

1.4. Overview

Gene models are dynamic and are constantly revised when more high-throughput data becomes available. It is inevitable that automated genome annotation will introduce errors due to the reasons mentioned in the previous sections, and proteomics data can play a vital role in rectifying these. Therefore in Chapter 2, we use proteomics data to validate (potentially erroneous) gene models for human and several model organisms and assess the quality of current genes sets. Emphasis is given on the identification of alternatively spliced isoforms and the biological relevance of these is highlighted with specific examples. The identified peptides are then classified according to the information content each possesses for the validation of gene models, where highly informative peptides can differentiate between splice variants. Chapter 2 also investigates the discrepancy between
the current level of proteomics support for gene models, and the level that could theoretically be achieved through targeted proteomics experiments.

The work in Chapter 2 is mainly dependant on Ensembl gene models. However, as discussed previously, the automated Ensembl pipeline is somewhat rigid as it does not consider ESTs and proteomics data when predicting gene models. Therefore, Chapters 3 and 4 explore methods for integrating these underused datasets to improve genome annotation in chicken with the aim of finding novel genes and transcripts that were missed by Ensembl. However, prior to the annotation of the chicken genome (Chapter 4), a benchmarking study was undertaken in Chapter 3 to investigate the optimum method for searching proteomics data against transcriptome data. Given that high-throughput transcriptomics and proteomics are being used more frequently for genome annotation, it is crucial to get this right so that a best practise method can become established in the field. Hence, Chapter 3 compares different approaches for the in-silico translation of ESTs and investigates how differences in database composition can affect peptide identification sensitivity. Specifically, we evaluate the six-frame translation database in terms of its ability to generate statistically significant PSMs and we examine how database composition influences the statistical modelling.

The evaluation of different database types and statistical models indicated the best approach to use for a large-scale proteogenomics study. In Chapter 4, we use this refined approach to annotate the chicken genome to see if mapping proteomics and transcriptomics data to the genome could improve gene models. We show the advantages of using de novo-assembled ESTs in proteogenomics to enable the identification of novel peptides that are not easily discernible using other database search strategies. The novel peptides are categorised according to the type of genomic feature they represent and we discuss how these can improve our understanding of avian biology.
1.5. References


Guigo, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V. B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras,


Genome annotation underpins almost every aspect of biology and the success of many experiments depends on the accurate construction of gene models. However, as mentioned in Chapter 1, such models often contain errors and proteomics data can be used to confirm whether predictions based on gene prediction algorithms and cDNA alignments represent translated proteins. Alternative splicing of pre-mRNAs presents a challenge to the field of proteomics because in some cases only a few peptides per protein can differentiate between splice-variants. It is unknown whether shotgun proteomics, in its current state, is capable of detecting the full repertoire of splice-variants across the genome, or whether new proteomics strategies are required to target discriminative peptides in a gene-centric approach.

In this chapter, we investigate the proportion of splice-variants that have been identified through shotgun proteomics experiments for a range of eukaryotic model organisms. We assess the discrepancy between the current level of proteomics support for splice variants, and the level of coverage that is theoretically achievable. The work presented here can assist future targeted proteomics experiments by categorising peptides according to their discriminative power and detectability. We then go on to investigate alternative proteomics strategies to detect and measure protein isoforms and outline the design of preliminary, gene-centric proteomics experiments.

The contents of this chapter were published as a research article in Proteomics (Blakeley, P., Siepen, J. A., Lawless, C. & Hubbard, S. J. (2010). Investigating protein isoforms via proteomics: a feasibility study. Proteomics, 10(6),1127-40.). The author of this thesis was the first author of the research article and was responsible generating the results and figures. Jennifer Siepen and Craig Lawless kindly provided scripts for submitting jobs to the mascot server and extracting Swiss-Prot protein protein isoforms respectively. Simon Hubbard directed the work and formulated most of the main ideas.
Abstract

Alternative splicing (AS) and processing of pre-messenger RNAs partially explains the *discrepancy* between the number of genes and proteome complexity in multicellular eukaryotic organisms. However, relatively few alternative protein isoforms have been experimentally identified, particularly at the protein level. In this study, we assess the ability of proteomics to inform on differently spliced protein isoforms in human and four other model eukaryotes. The number of Ensembl-annotated genes for which proteomic data exists that informs on AS exceeds 33% of the alternately spliced genes in the human and worm genomes. Examining AS in chicken *via* proteomics for the first time, we find support for over 600 genes. However, although peptide identifications support only a small fraction of alternative protein isoforms that are annotated in Ensembl, many more variants are amenable to proteomic identification. There remains a sizeable gap between these existing identifications (10-52% of AS genes) and those that are theoretically feasible (90-99%). We also compare annotations between Swiss-Prot and Ensembl, recommending use of both to maximise coverage of AS. We propose that targeted proteomic experiments using selected reactions and standards are essential to uncover further alternative isoforms and discuss the issues surrounding these strategies.
2.1. Introduction

Mass spectrometry-based proteomics promises to provide a unique insight into many biological processes. Proteins are key determinants of cellular function and their dynamic interactions reveal much more about cell-type specific processes than a static genome sequence. However, the fields of proteomics and genomics are highly complementary to one another, a fact which is being exploited in the emerging field of proteogenomics (Castellana et al., 2008; Desiere et al., 2005; Gupta et al., 2008; Gupta et al., 2007; Savidor et al., 2006; Tanner et al., 2007; Wright et al., 2009; Merrihew et al., 2008). The rapid expansion of proteomics in the post-genomic era can be partly attributed to advances in nucleic acid sequencing and genome annotation. This has prompted an increase in putative protein sequences in databases such as IPI (Kersey et al., 2004), Ensembl (Hubbard et al., 2009), TrEMBL (The UniProt Consortium, 2008), and RefSeq (Pruitt et al., 2007). Mass spectrometry data is routinely searched against such databases to identify proteins. But protein identification may be hampered if protein-coding genes have been poorly annotated.

High quality genome annotation relies on physical evidence and orthology with ‘known’ genes (Potter et al., 2004). The former has traditionally been in the form of full-length cDNA libraries which provide direct transcriptomic evidence that a particular gene locus is expressed (Okazaki et al., 2002). Sequencing of such libraries can also generate ESTs which provide partial coverage of their parent cDNA. This method of genome annotation has its limitations, especially when annotating protein-coding genes. ESTs for the most part represent protein-coding regions, but some may be derived from non-coding RNAs which can be incorrectly annotated as mRNAs (Hubbard et al., 2005; Frith et al., 2006). Many will also be mostly untranslated regions (UTRs). Furthermore, the merger of ESTs from different splice variants into the same contig may prevent the unambiguous identification of alternatively spliced transcripts. Such information is important given the important role that different isoforms play in tissue specificity and development. To confirm the translation of these alternative spliced transcripts, proteomic evidence potentially has a role to play, since it offers physical evidence that a gene is translated. In addition, proteomics can reveal the true nature of mature proteins that have undergone a multitude of different post-translational modifications (PTMs), which are difficult to predict from nucleotide sequences alone.
Alternative splicing (AS) of pre-mRNAs is widespread in humans and throughout most eukaryotes, occurring in 40-60% of all genes (Modrek and Lee, 2002; Modrek and Lee, 2003). Indeed, the most recent data suggests as many as ~95% of human multi-exon genes undergo AS (Pan et al., 2008). The high levels of tissue specificity and functional diversity in higher organisms is thought to be consistent with their higher levels of AS (Graveley, 2001). For example, AS is thought to play a prominent role in the regulation of neuronal development (Jiao et al., 2008; Boutz et al., 2007), metabolism (Passetti et al., 2009), immunity (Zikherman and Weiss, 2008), and may modify protein-protein interaction domains to regulate signalling pathways at key switch points (Jiao et al., 2008; Resch et al., 2004). For example, Kruppel transcription factors contain Kruppel associated box domains which appear to be removed by AS, directly affecting their function (Resch et al., 2004).

Although AS is generally held to be responsible for the greatest diversity of gene products expressed from the genome, the full extent of AS at the proteome level remains to be characterised. Some studies have concluded that AS has a major influence on the proteome and that proteins can remain functional after AS induced disruption of structured protein regions (Birzele et al., 2008). In contrast, some argue that there is little evidence to suggest that AS can generate functional protein isoforms on a large scale (Tress et al., 2007). Hence, further proteomic analyses are required to investigate alternative protein isoforms. One recent large-scale study has attempted to do just that, identifying alternative protein isoforms using mass spectrometry data in Drosophila (Tress et al., 2008). In this case, 130 genes were observed to display AS at the protein level, confirmed by peptides with attendant high quality mass spectra.

Despite these recent studies, most splice variants in metazoan genomes have not yet been confirmed by mass spectrometry. Nevertheless, a large body of information exists in the primary literature and sequence databases concerning alternate splicing. The Ensembl database provides a large body of putative AS events in its genebuilds, which are derived principally from automatic annotations using aligned transcripts and proteins to the genome sequence. The Human Proteome Initiative (HPI), the human component of UniProtKB/Swiss-Prot (Boeckmann et al., 2005), provides a large number of arguably the highest quality manual annotation of alternative isoforms, giving in-depth functional
information (e.g. domain structure, PTMs). Currently\(^2\) 7,439 human genes are annotated as expressing multiple isoforms in the HPI set.

In this work, we have examined the potential for proteomics to contribute to the identification and characterisation of alternately spliced protein isoforms, considering this from a theoretical standpoint for a range of metazoan eukaryotic genomes. We have based this around the set of alternately spliced transcripts annotated by Ensembl for five genomes, as well as comparing this with the UniProtKB/Swiss-Prot HPI set of human isoforms. Here we use AS to refer generically to any underlying process that generates an alternative transcript, including novel transcriptional start sites, multiple poly-adenylation sites and RNA editing. These data is cross-referenced with extant peptide identifications found in the PeptideAtlas repository (Desiere et al., 2005), as well as a large data set of chicken mass spectra (Hall et al., 2009). This is the first proteome-wide study of AS in the chicken, and is the first comparative analysis of protein isoforms supported by peptide identifications in several metazoan proteomes. Our data support the assertion that proteomics has a vital role to play in characterising the isoformal-specific proteome. However, the current results are principally derived from serendipitous peptide identifications in “shotgun” style proteomics experiments, and therefore argue the role for a more targeted approach. Our data strongly suggests that targeted proteomics experiments have a valid role to play in characterising alternate splicing and gene structure in the context of genome annotation.

\[2.2. \text{Materials and Methods}\]

\[2.2.1. \text{Genome and proteome sequences}\]

All proteome sequences were taken from the Ensembl 48 build, downloading the “pep.all” Fasta files containing the translations of all known and novel Ensembl gene predictions along with the attendant GTF files mapping CDS and exon elements back to genomic coordinates (http://ensembl.org). This was done for the following species: human, mouse, chicken, Drosophila, and Caenorhabditis elegans. This represents a range of highly annotated eukaryotic species which display varying levels of alternate splicing and

\(^2\) As observed in 2009. There are currently 8396 as of September 2012 in UniProt.
for which sizeable datasets of peptide identifications are available. Version 48 of Ensembl was used throughout this study for consistency. It was originally used for all our chicken searches, and broadly corresponds with the Ensembl versions used to generate the PeptideAtlas peptide identifications. A complete description of the Ensembl gene/transcript annotation pipeline is beyond the scope of this thesis (for a full description, see Potter et al., 2004), but briefly, Ensembl determines genes using automatic annotation. This involves both computer and biological (human) expertise to determine an entire gene set based around reconciled alignments of high-quality protein and transcript data to provide an Ensembl genebuild. Hence, all Ensembl transcripts and genes are supported by protein/mRNA evidence. Only protein isoforms that were non-redundant at the sequence level were considered. Genes with transcripts that differed only at the mRNA level were not considered to be multi-isoform for the purpose of this study. All human UniProtKB/Swiss-Prot entries were taken from the integr8 web portal (http://ebi.ac.uk/integr8).

2.2.2. Peptide identifications

Apart from chicken, all peptide identifications were taken from the PeptideAtlas database (http://peptideatlas.org/builds/). PeptideAtlas organises data into “builds” centred on a particular species or tissue. For this work, we used the human build (http://peptideatlas.org/builds/human/200903/APD_Hs_all.fasta March 2009), containing 104,744 peptides, the Mouse plasma build (http://peptideatlas.org/builds/mouse/plasma/2007_05/PA_MmP_2007-05_Ens43_P0.9_peptides.fasta May 2007) containing 13,779 peptides, Drosophila melanogaster build (July 2004, 4442 peptides, no longer accessible online), and C.elegans build (http://peptideatlas.org/builds/celegans/Celegans200805/APD_Ce_all.fasta May 2008), containing 80,834 peptides. All contained high confidence peptide identifications ($p > 0.9$) after processing with the Trans-Proteome Pipeline (Keller et al., 2005).
### 2.2.3. Chicken samples and data analysis

Peptide identifications were derived from a proteomic analysis of the DT40 Chicken cell line (Hall et al., 2009) which aimed to characterise the localisation of organelle proteins by isotope tagging (LOPIT). DT40 cells were fractionated by density gradient centrifugation and 7 fractions were chosen for analysis. The selected fractions were labelled with four-plex iTRAQ reagents and digested with trypsin. Labelled peptides were pooled together and separated using two-dimensional liquid chromatography. LC-MS/MS was performed using an ultimate-nano-LC system (Dionex) coupled to a QSTAR XL mass spectrometer (Applied Biosystems). A total of 403,820 tandem MS spectra were searched against the Ensembl 48 protein database (http://ensembl.org) using Mascot version 2.2. The following parameters were set; fully tryptic peptides with up to 1 missed cleavage, fixed modifications were iTRAQ-labelled free amines and cysteine methylthiolations, variable modifications of iTRAQ-labelled tyrosines and methionine oxidations. MS/MS fragment and peptide error tolerances were set to 0.8 Da and 2 Da respectively. To estimate false discovery rates, spectra were searched against a decoy database constructed from the target proteome concatenated with the set of reverse sequences. Processing of these results estimated FDRs using the approach of Kall et al (Kall et al., 2008) retaining all peptide identifications at an FDR of 0.01. In total, 14,612 high quality identifications were retained.

### 2.2.4. Peptide mapping

For all species, we performed an in silico tryptic digest of the known and predicted proteins from Ensembl 48, following the standard specificity rules cleaving after all lysine and arginine residues not preceding proline. Using the Ensembl GTF file containing CDS/exon genomic coordinates, all theoretical peptides were subsequently mapped back to specific exons and were also used as a platform to match experimentally determined peptides back to genomic loci. Only experimental peptides with exact and complete 100% matches to theoretical peptides were used, thereby ensuring that they were also fully tryptic. Peptides which mapped to several locations and therefore could not be unambiguously mapped to a single genomic location were removed from further
consideration. Some peptides from PeptideAtlas builds could not be mapped owing to conflicting versions of Ensembl or other sequence databases used for peptide identification searches. To allow mapping of missed cleaved peptides, we performed a second *in silico* tryptic digest which generated all possible tryptic peptides containing up to and including one missed cleavage.

All unambiguously mapped peptides were categorised on the basis of the number of exons they covered and on the type of exons they covered (see Figure 2.1.). Exons were considered to be constitutive (present in all isoforms predicted for a given gene), semi-constitutive (present in some but not all isoforms) or unique (present in a single isoform). Peptides that covered two or more exons were classified as intron spanning.

Peptide sets were also subject to two filters to remove peptides unlikely to be seen in mass spectrometry experiments. The first filter excluded peptides that were less than 6 or greater than 40 in amino acids in length. Secondly, the remaining peptides were analysed with the PeptideSieve algorithm, which predicts the likelihood of the peptide being observed in a proteomics experiment, taking into account ionisation and missed cleavage propensity (Mallick *et al.*, 2007). A full list of peptide numbers at each stage is provided in Table 2.1.

### 2.2.5. Database comparison with Swiss-Prot HPI

We also compared the Ensembl human dataset, a largely automated annotation of predicted AS, with the smaller but highly curated set of isoforms contained in the Swiss-Prot HPI set. We first collated all 7,439 multi-protein isoform genes from the HPI UniProtKB/Swiss-Prot dataset and cross-referenced these 14,306 proteins with Ensembl proteins to assess the overlap between the two datasets. Protein isoforms in Swiss-Prot were matched to Ensembl proteins using BLASTP (Altschul *et al.*, 1990) using a simple criterion that retained only the most significant candidate matches (e-value < 1e-12, %identity >99%, aligned coverage >99%). From this, we were able to identify genes that were annotated as encoding multiple isoforms in both databases. Peptides from the Human PeptideAtlas build were matched with the protein isoforms that contained BLAST hits. From this we were able to estimate the proportion of AS isoforms in Swiss-Prot and Ensembl that have supporting proteomics evidence.
2.2.6. QconCAT design

The protein sequences of the full set of AS isoforms for the *Drosophila para* gene were downloaded from FlyBase (FlyBase Consortium, 2002). The protein sequences were trypsin-digested *in silico* using a custom perl script and the resulting peptides were mapped to specific exons. Nineteen peptides were selected to target AS exons, five to target constitutive exons, and as a control, five were also selected to target the housekeeping protein RP49. The peptides were then ordered within the QconCAT to generate the lowest possible average missed cleavage propensity. An SVM based program (Lawless and Hubbard, 2012) was used to calculate the missed cleavage propensities. Only peptides with a missed cleavage propensity lower than 50% were included in the QconCAT.

![Diagram showing exon and peptide mapping nomenclature.]

**Figure 2.1. Exon and Peptide mapping nomenclature.**

Three different transcript models processed from the same pre-mRNA are shown, with introns represented by black lines connecting the exons. Exons are classified by whether are present in all transcripts (C=Constitutive), some transcripts (S=Semi-constitutive), or a single transcript only (U=Unique). Peptides are initially classified by virtue of the number and type of exons that they cover, although this is not always fully informative of their status. For example, peptide A spans a constitutive and unique exon, and is therefore “unique”, whilst peptide B is wholly contained within the unique exon 2, yet is “constitutive” as this sequence is also in-frame and contained within exon 3 present in the other two transcripts.
### Table 2.1. Peptide data sets calculated or used in this study

<table>
<thead>
<tr>
<th>Ensembl species</th>
<th>Tryptic Peptide subsets (including upto 1 missed cleavage)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all unambiguously mapped peptides</td>
<td>length filtered peptides</td>
</tr>
<tr>
<td>Human</td>
<td>1571922</td>
<td>1400266</td>
</tr>
<tr>
<td>Mouse</td>
<td>1156002</td>
<td>810503</td>
</tr>
<tr>
<td>Chicken</td>
<td>1243182</td>
<td>1109471</td>
</tr>
<tr>
<td>Fly</td>
<td>1105525</td>
<td>681090</td>
</tr>
<tr>
<td>Worm</td>
<td>921281</td>
<td>868143</td>
</tr>
</tbody>
</table>

*Predicted to be observed in more than 50% of MS experiments by PeptideSieve.

### 2.3. Results and discussion

#### 2.3.1. Proteomic data and alternative splicing

Alternative splicing can be categorised at the exon level, identifying transcriptional isoforms which use different exonic structures to form a mature transcript. Given that several isoforms may be characterised from a single gene locus, exons can be characterised at the simplest level as either “constitutive” (present in all isoforms), “semi-constitutive” (present in two or more, but not all, isoforms) or “unique” (present in only a single isoform). This is illustrated conceptually in Figure 2.1. Peptide data may be used to characterise this information, mapping the amino acid sequences derived from proteolysis on to this exon structure. Peptides must cover a unique or semi-constitutive exon to allow identification of a specific isoform or a subset of isoforms. However, there is not a simple one-to-one mapping, since peptides can span multiple exons or may be contained entirely within a single exon. Furthermore, a peptide sequence may not be unique within a single set of protein isoforms (i.e. it can be repeated, either in a single isoform or at different genomic sites across several isoforms) or within the entire protein complement (at multiple genomic loci in different genes). Again, Figure 2.1 shows some of the possibilities. Three alternate transcripts are shown, constituted by five unique exons used by this gene. The different peptides are informative about which exons they are consistent with, and
therefore which isoform(s) are present. For example, peptide A spans two exons which are in turn constitutive and unique and so this peptide is also unique since it is only present in transcript 3. However, although peptide B is entirely contained within the unique exon 2 this overlaps with the semi-constitutive exon 3 and the peptide itself is therefore considered to be constitutive. There is therefore no direct relationship between exon status and that of the peptide in terms of information relating to AS and our analyses have taken these situations into account.

These peptide definitions allow us to characterise AS events for individual genes and the multiple protein isoforms they encode. Here, we use a relaxed definition in that any peptide which is either unique or semi-constitutive is informative to the set of predicted isoforms i.e. it is consistent with only a subset of those predictions.

The Ensembl genome annotations were selected as the principal dataset for this study, although the corresponding IPI databases contain more proteins (Kersey et al., 2004), since Ensembl affords a single, consistent mechanism for mapping genes, proteins and peptides back to the genome. It is worth commenting, however, that a sizeable fraction of the Ensembl translations for a given gene are identical at the protein level, despite differing at the transcript level; many AS events only affect UTRs or introns leaving the final protein translation unchanged. Hence, only AS events that lead to different protein products are amenable to MS-based approaches and were considered in this study. The level of AS currently predicted in the Ensembl gene sets, defined at the non-redundant protein level, is listed in Table 2.2a. As would be expected the Human and Mouse proteomes are considerably larger, with a higher level of annotated AS in the human genome, followed by mouse. This highlights differences in the attendant level of experimental supporting data and genome annotation across the different model organisms, as well as differences in the underlying proteome size. Table 2.2 shows that the majority of genes only encode a single protein isoform according to the Ensembl annotations.

Although this is likely to be an underestimate (Modrek and Lee, 2002; Modrek and Lee, 2003) this does represent a consistent set of high quality annotations, ranging from 12-41% of the genome (Table 2.2b). Nevertheless, even though the fly and worm genomes have a relatively low percentage of multiple-isoform genes, there are still a large number of annotated alternative protein isoforms. Indeed, although C. elegans has the lowest fraction of AS genes, the average number of isoforms within this 2337 gene subset is above 2 and
many putative isoforms could potentially be validated at the protein level. Conversely, although the fraction of AS genes in the Chicken is greater than in the invertebrates, there are relatively fewer protein isoforms per gene. Although previous experiments have characterised a significant fraction of the Chicken proteome (Mann, 2007; Mizukami et al., 2008; Buza et al., 2007; Zheng et al., 2008), the majority of the genome still lacks direct proteomic experimental evidence.

Table 2.2. Protein isoforms in Ensembl 48

a) Total genes

<table>
<thead>
<tr>
<th>organism</th>
<th>protein coding genes</th>
<th>proteins</th>
<th>average number of isoforms per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>22,839</td>
<td>41,021</td>
<td>1.80</td>
</tr>
<tr>
<td>Mouse</td>
<td>23,429</td>
<td>35,628</td>
<td>1.52</td>
</tr>
<tr>
<td>Chicken</td>
<td>16,736</td>
<td>21,871</td>
<td>1.31</td>
</tr>
<tr>
<td>Fly</td>
<td>14,039</td>
<td>17,104</td>
<td>1.22</td>
</tr>
<tr>
<td>worm</td>
<td>20,140</td>
<td>23,334</td>
<td>1.16</td>
</tr>
</tbody>
</table>

b) Genes encoding more than one non-redundant protein isoform

<table>
<thead>
<tr>
<th>organism</th>
<th>protein coding genes</th>
<th>proteins</th>
<th>average number of isoforms per gene</th>
<th>% of genes with &gt;1 isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>9,343</td>
<td>27,903</td>
<td>2.99</td>
<td>41</td>
</tr>
<tr>
<td>Mouse</td>
<td>7,280</td>
<td>19,876</td>
<td>2.73</td>
<td>31</td>
</tr>
<tr>
<td>Chicken</td>
<td>3,776</td>
<td>8,990</td>
<td>2.38</td>
<td>23</td>
</tr>
<tr>
<td>Fly</td>
<td>2,079</td>
<td>5,276</td>
<td>2.54</td>
<td>15</td>
</tr>
<tr>
<td>worm</td>
<td>2,337</td>
<td>5,706</td>
<td>2.44</td>
<td>12</td>
</tr>
</tbody>
</table>

This study has focused on characterising AS events in genes that are derived from differential RNA processing leading to different protein sequences. Figure 2.2 shows a breakdown of the exon categories across the five genomes. This follows the same general trend as at the gene level, where the majority of annotated exons are not informative to AS and multiple isoforms. Similarly, the level of informative exons falls further with the less complex and less well annotated genomes. This is unsurprising given the fact that the majority of exons (more than 60 %) in the invertebrate genomes are from single isoform Ensembl genes. The proportion of single isoform genes is also large in the chicken, possibly reflecting the fact the chicken genome is at a relatively early stage in the
annotation process. Together, these data highlights that only a relatively small fraction of even the best annotated genomes are currently amenable for AS characterisation via proteomics, so long as only previously annotated proteins are searched with mass spectrometry data.

**Figure 2.2 Ensembl exon classification.**

The relative numbers of unique, semi-constitutive and constitutive exons are shown across the five genomes under study, distinguishing constitutive exons in genes which only express one isoform (dark brown) from those that don’t (lighter browns). Semi-constitutive exons are expressed in some but not all isoforms. The relative fraction of constitutive exons decreases broadly with organismal complexity. It should be noted that less than 47% of all genes exhibit AS (Table 2.1), which is broadly in line with the percentage of exons which are not constitutive.
Peptides are also potentially informative if they span an intron, confirming the exonic order for at least two exons within a mature mRNA. Previous proteogenomic studies estimate that ~25% of peptides identified in mass spectrometry experiments are intron spanning (Tanner et al., 2007). The relative fraction of tryptic peptides that span introns in a digest of the entire proteome is shown in Figure 2.3. This number holds up around the 20-25% level for vertebrate proteomes in general, slightly less for the nematode worm, and falls off below 10% for the fly genome. However, the level of annotated AS in fly and worm is reduced, which contributes to this result. Of greater interest for proteomic studies is that the level of intron-spanning peptides holds up when experimental factors are considered. As shown in Figure 2.3 the fraction of intron spanning peptides remains broadly constant when filters are applied. The length filter removes short and long peptides whilst the proteotypic filter removes peptides not predicted to be observed in the mass spectrometer (Mallick et al., 2007). In both cases, the level of intron spanning peptides remains roughly constant, similarly for experimentally determined peptides. This is encouraging for MS-based proteomic studies.

Figure 2.4 examines this in more detail, classifying tryptic peptides in the five proteomes in terms of their ability to inform on AS generated isoforms. Similar data were also generated from the exon point of view with virtually identical results (data not shown). The data shows that unique and semi-constitutive peptides make up almost 40% of the theoretical human proteome. This falls off below 30% for mouse and more so in chicken and invertebrates; the fraction of unique and semi-constitutive peptides is only ~10% in C. elegans. This is consistent with the notion that levels of AS have increased during metazoan evolution (Kim et al., 2007), supporting the functional diversification of vertebrate proteomes. It also, in part, reflects the level of annotation currently available for the different proteomes.
Figure 2.3. Predicted and experimental peptides that span introns. The relative fractions of intron-spanning and wholly exon-internal tryptic peptides in five Ensembl proteomics is shown, for four sub categories: “all” refers to a complete digest of the proteome (removing only redundant peptide sequences which cannot be unambiguously placed on the genome), “Length” refers to a further filter where peptides <5 or >40 amino acids are removed, “proteotypic” refers to peptides that are additionally predicted to be observed by PeptideSieve, and “experimental” refers only to peptides with high quality corresponding experimental data (see Methods). The data in this plot includes peptides up to and including 1 missed cleavage. The total number of peptides in each set is display at the end of the row.
Figure 2.4. Relative classifications of peptide sets from metazoan proteomes in terms of detection of specific protein isoforms. Different peptide subsets from the five Ensembl genomes are mapped to the exon structure of their genes and classified as either unique, semi-constitutive or constitutive. The peptide sets are: “all” refers to a complete digest of the proteome (removing only redundant peptide sequences which cannot be unambiguously placed on the genome), “Length” refers to a further filter where peptides <5 or >40 amino acids are removed, “proteotypic” refers to peptides that are additionally predicted to be observed by PeptideSieve, and “experimental” refers only to peptides with high quality corresponding experimental data (see Methods). The data in this plot includes peptides up to and including 1 missed cleavage. The total number of peptides in each set is displayed at the end of the row.

The data suggests there is a greater chance of identifying alternative protein isoforms via mass spectrometry in higher organisms, but assumes that the predicted alternative exons in Ensembl are indeed translated. There is increasing evidence for the coupling of AS with nonsense-mediated mRNA decay, which is unproductive at the protein level (Lewis et al., 2003). Also, some protein isoforms may be expressed at very low levels or only in special circumstances (e.g. during a single developmental phase), or
fail to produce peptides that are compatible with mass spectrometry. To address this latter point, we filtered the peptide sets by length, proteotypic nature, and finally those with experimental MS-based evidence.

Filtering out peptides by length (<6 or >40 amino acids) resulted in a ~12% decrease in the number of peptides for each organism. However, proteotypic peptides are also defined by other physico-chemical properties such as hydrophobicity, secondary structure and overall charge and represent peptides likely to be detected in a mass spectrometer (Mallick et al., 2007). The relative fraction of length-filtered or proteotypic peptides likely to inform on AS remains reasonably high, but the overall numbers involved are clearly smaller. Across all species, there is on average 3.9 proteotypic peptides per alternative exon. An exceptional case is Drosophila (which has 5.5 proteotypic peptides per alternative exon), which has already been studied successfully experimentally (Tress et al., 2008). This information is useful for future proteomic experiments which could target one of several proteotypic peptides for the identification of a specific alternative protein isoform.

The relative fraction of AS-informative peptides in worm and fly does not fall when experimental data is considered. However, we note a small reduction in the relative fraction of AS-informative peptides when solely experimental data is considered for human, mouse and chicken. This is unsurprising when considering that these data were not acquired with the characterisation of AS in mind and the difficult task of sampling the multitude of tissue-specific AS events that occur in these more complex organisms. This represents the serendipitous confirmation of AS from existing experimental data, whereas the proteotypic results show the theoretical maximum. Indeed, this corresponds to around 260,000 unique tryptic peptides in the human proteome, a substantial number. We estimate that under 20,000 (less than 8%) have been observed to date, suggesting there is a big gap to close.

A similar picture emerges when considering the data at the gene level. Table 2.3 shows the relative fraction of genes for which proteomic data is available from the sources described. This is most likely an underestimate, since PeptideAtlas is not the sole public repository (Jones et al., 2008; Mathivanan and Pandey, 2008; Mead et al., 2009) and more (unpublished) data must exist in many individual laboratories. Nevertheless, a sizable
fraction of these genomes is covered by proteomic data in PeptideAtlas, ranging from 6% up to 41% in *C. elegans*. Encouragingly, the sub-fraction of AS genes with peptide-based evidence is of a slightly higher level, ranging from 10 to 52%. This is quite promising given that general proteomics experiments are not targeted at AS characterisation. Equally there remains a large gap to the theoretical set of genes that could be characterised via MS-based approaches, as estimated using all proteotypic peptides, which ranges from 90%-98%. This assumes, however, that all these peptides are amenable to study by proteomics and attendant mass spectrometry, and indeed that the predicted isoforms are correctly annotated, expressed and translated in sufficient quantities in tissues available for study.

The data presented here illustrate the level of AS consistent with proteomic evidence in chicken for the first time, supporting the predicted annotations for 628 genes, around 17% of those predicted to express multiple isoforms.

Table 2.3. Relative fractions of gene sets with associated peptide evidence

<table>
<thead>
<tr>
<th>Species</th>
<th>Total number of genes</th>
<th>genes with peptide evidence(^a)</th>
<th>Genes annotated with multiple protein isoforms</th>
<th>AS genes with peptide evidence for specific isoforms(^b)</th>
<th>AS genes with proteotypic peptides for specific isoforms(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>22,839</td>
<td>7,868 (34%)</td>
<td>9,343</td>
<td>3,059 (33%)</td>
<td>8,376 (90%)</td>
</tr>
<tr>
<td>Mouse</td>
<td>23,429</td>
<td>3,096 (13%)</td>
<td>7,280</td>
<td>747 (10%)</td>
<td>6,874 (94%)</td>
</tr>
<tr>
<td>Chicken</td>
<td>16,736</td>
<td>3,023 (18%)</td>
<td>3,776</td>
<td>628 (17%)</td>
<td>3,710 (98%)</td>
</tr>
<tr>
<td>Fly</td>
<td>14,039</td>
<td>907 (6%)</td>
<td>2,079</td>
<td>211 (10%)</td>
<td>2,020 (97%)</td>
</tr>
<tr>
<td>Worm</td>
<td>20,140</td>
<td>8,256 (41%)</td>
<td>2,337</td>
<td>1,216 (52%)</td>
<td>2,130 (91%)</td>
</tr>
</tbody>
</table>

\(^a\) only high quality peptide identifications from PeptideAtlas or (Hall *et al.*, 2009)

\(^b\) only high quality peptide identifications from PeptideAtlas or (Hall *et al.*, 2009) which are consistent with the expression of a subset of the total annotated isoforms at a single gene locus.

\(^c\) the subset of peptides from a theoretical digest of the entire proteome (including up to 1 missed cleavage) that pass the PeptideSieve and length filters, and which are consistent with the expression of a subset of the total annotated isoforms at a single gene locus.
2.3.2. Examples of alternate splicing with supporting proteomic evidence

Two example cases illustrating the utility of proteomic annotation for AS characterisation are shown in Figs. 2.5 and 2.6. Figure 2.5 shows how the predicted isoforms of the chicken FAM136A gene can be validated through mass spectrometry data. The FAM136A hypothetical protein-coding gene has orthologues across the metazoa, from nematodes to human. In Ensembl it is predicted to encode two alternative protein isoforms resulting from alternative transcriptional initiation. We have mapped six peptides to the two different transcripts and provide evidence for the translation of four alternative exons. However, if the gene prediction made by Ensembl happened to be incorrect, one can imagine a scenario where exons could be incorrectly assigned to a specific transcript. This is where intron spanning peptides are useful in determining transcript structure.

Figure 2.5. Peptide evidence for alternative protein isoforms encoded by a hypothetical gene.

The gene model FAM136A is predicted to encode two different isoforms. Ensembl transcripts (top) and enlarged exons (bottom) are shown in maroon. Mapped tryptic peptides are shown as red and blue blocks below the expanded exon structure. In total, 6 unique peptides were unambiguously mapped to ENSGALT00000022503 and ENSGALT00000040525. All peptides provide evidence for the translation of unique exons.
Figure 2.6. Peptide evidence for exon–intron boundaries in the Chicken TPM3 gene.

The Tropomyosin alpha-3 chain gene is predicted to encode 4 alternative protein isoforms. Ensembl transcripts (top) and enlarged exons (bottom) are shown in maroon. Mapped tryptic peptides are shown as red and blue blocks below the expanded exon structure. In total, 6 unique peptides were unambiguously mapped to ENSGALT00000022043 and ENSGALT00000040695. All peptides cover unique or semi-constitutive exons, providing evidence for differential pre-mRNA processing.

Figure 2.6 illustrates an additional example of RNA splicing by showing the relative positions of exons along transcripts of the chicken Tropomyosin alpha-3 chain (TPM3) gene, which is involved in skeletal and smooth muscle contraction as well as other functions in non-muscle cells. Chicken TPM3 has been annotated by Ensembl to encode four alternative protein isoforms, whereas human TPM3 encodes more than 10, suggesting that AS has evolved to allow diversification of TPM3 function in humans. The chicken TPM3 isoforms were previously lacking proteomic evidence. We have now unambiguously mapped six peptides to two of the Chicken isoforms, and three peptides to a specific alternative exon. The intron spanning peptides 4 and 6 confirm an exon skipping event involving two alternative exons of the ENSGALT00000022043 and ENSGALT00000040695 transcripts (Figure 2.6). A detailed view of the peptide sequences is supplied in an alignment in Appendix 1. Exon skipping events also occur in human
TPM3 isoforms, and these may confer tissue specific functions if the AS events are regulated in a tissue specific manner. Data such as this could lead to hypothesis-driven experiments to characterise the chicken TPM3 protein isoforms in more detail. Because AS often mediates quantitative changes in the levels of splice variants (Marden, 2008), quantitative proteomics is required to characterise these changes in alternative isoform expression, requiring targeted approaches to selectively quantify specific peptides similar to those illustrated in Figs. 2.5 and 2.6.

2.3.3. Protein Isoforms in UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot represents an accurate representation of the human proteome. Because it is manually annotated, Swiss-Prot is widely regarded as the gold standard for high-quality protein sequences. However, a number protein sequences could be missing from Swiss-Prot owing to its stringent annotation process. Therefore, we investigated the overlap of protein isoforms present in the two databases. Table 2.4 shows the similarities between the Swiss-Prot HPI and Ensembl 48 annotations of multiple isoforms in human. Ensembl has a greater number of genes predicted to be multi-isoform, although 5671 are also multiple-isoform in Swiss-Prot. Equally, both resources have genes that are uniquely multi-isoform; 1345 and 3717 genes that are only found in Swiss-Prot and Ensembl respectively. Given that Ensembl uses a high quality but largely automated pipeline it is likely to contain some incorrect isoform annotations, whilst equally it may contain some correct annotations that are not yet included in Swiss-Prot. This is supported by Table 2.4, where significant numbers of peptides have been mapped to genes which are multi-isoform in only one of the two annotations. Moreover, significant numbers of these peptides are “unique” to one of several isoforms in these subsets. There is almost no overlap (only 8 peptides) between the unique peptide sets mapped to unique multi-isoform genes in the two annotations.
Table 2.4. Comparative analysis of different proteome annotations for alternate splicing and proteomic peptide evidence

<table>
<thead>
<tr>
<th>Genes annotated with multiple non-redundant protein isoforms</th>
<th>HPI Swiss-Prot</th>
<th>Ensembl48</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>7,439&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Genes annotated with multiple non-redundant protein isoforms</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Uniquely multi-isoform</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1345</td>
</tr>
<tr>
<td>Total number of associated protein isoforms</td>
<td>3152</td>
<td>17358</td>
</tr>
<tr>
<td>Total number of mapped peptides</td>
<td>4281</td>
<td>23939</td>
</tr>
<tr>
<td>Number of mapped “unique” peptides (present in only a single isoform)</td>
<td>927</td>
<td>2853</td>
</tr>
<tr>
<td>Number of genes with “unique” isoform-specific peptides mapped</td>
<td>271</td>
<td>911</td>
</tr>
</tbody>
</table>

<sup>a</sup> Several Swiss-Prot isoforms could not be unambiguously mapped using our procedure requiring a minimum 99% percent identity and coverage between the two proteins, and hence the number of unique and shared genes do not sum to 7439. 

<sup>b</sup> A few Swiss-Prot genes were equivalence-d to a single Ensembl gene, and vice versa, and hence the shared gene total is not identical in the two annotations.

However, this hides an anomaly in the data. Although these peptides are uniquely assigned to only a single isoform from multiple possibilities in one of the two annotation sets, they are also still detected on most occasions in the corresponding one. Indeed, in spite of gene models varying significantly between Swiss-Prot and Ensembl, we estimate that a small proportion (<2%) of identified peptides are unique to one of the two annotation sets. For example, the human TGFβ gene (Appendix 2) is assigned two isoforms by Ensembl 48 but only one by UniProt (Accession Q15582). Several peptides are “uniquely” mapped to the Ensembl isoforms, but all of them are also present in the single equivalent UniProt protein. Interestingly, Ensembl 55 has revised the ENSP00000306306 isoform and now has three predicted isoforms including one identical to the UniProt isoform. It is not possible from the current data to determine whether one or both of the existing annotation sets are correct from the existing peptide identifications. Examples such as this point to the need for alternative proteomics strategies which can identify specific peptides (such as N-terminal peptides or unique exon spanning peptides) that can distinguish between isoforms in instances similar to that shown in Appendix 2.

Although Swiss-Prot annotations generally contain fewer errors, they can miss isoforms due to their stringent annotation methods. This is demonstrated in Figure 7,
showing the multiple Ensembl isoforms aligned with the HPI UniProt entry for the human NDUFV3 gene. This isoform retains a mitochondrial targeting sequence (in the grey box) but contains an extended internal region derived from an exon inclusion event. Although absent in the HPI UniProt entry, the extended internal region is supported by nine identified peptides including one which overlaps the alternative splice site (peptide 9). Any holistic approach attempting to identify AS events via proteomics in human genes should therefore consider both resources, Ensembl and Swiss-Prot, as novel information can be gained from both. The different philosophies underlying the annotation of these two resources (i.e. semi-automated gene prediction via aligned cDNAs/proteins vs. detailed manual curation) undoubtedly determine these differences.

2.3.4. Implications for design

The data presented here demonstrate the potential for proteomics as a tool to study AS on a genome-wide scale using mass spectrometry, although there is clearly a gap to close. To date, most studies have not been targeted towards this goal and have followed “classical” approaches which reduce the complexity of the digest protein samples using liquid phase columns and/or gels, and use data-directed acquisition approaches in a “shotgun” fashion. Regardless, a considerable number of isoforms have been characterised using these approaches (Castellana et al., 2008; Gupta et al., 2008; Loevenich et al., 2009; Menon et al., 2009; Power et al., 2009; Schrimpf et al., 2009; Tress et al., 2008). However, the data presented here suggest the majority of the alternately spliced isoforms could be characterised by selecting specific peptides (e.g. proteotypic ones) for selected reaction monitoring (Schiess et al., 2009). This also opens up the possibility of a fully quantitative approach, using selected reaction monitoring (sometimes referred to as MRM) approaches or labelled internal standard peptides (AQUA (Gerber et al., 2007; Gerber et al., 2003; Kirkpatrick et al., 2005) or QconCAT (Pratt et al., 2006; Rivers et al., 2007)) to quantify individual isoforms. Although these approaches are not without their problems (Duncan et al., 2009; Pan et al., 2009; Sherman et al., 2009), a rigorous design strategy and careful experimentation can produce useful results.
Figure 2.7. Correction of an incomplete model of the human NDUFV3 gene.

Swiss-Prot contains only a single NADH dehydrogenase [ubiquinone] flavoprotein 3 protein, whereas Ensembl has annotated 3 different isoforms. The multiple sequence alignment shows a large internal region in ENSP00000342895 that is missing in ENSP00000346196 and in equivalent Swiss-Prot entry (P56181). Nine peptides (bold regions) were unambiguously mapped to ENSP00000346196, thus confirming an exon-skipping event that has not been annotated in Swiss-Prot.
Several further considerations are worthy of discussion. Trypsin is the preferred digestive enzyme for most mass spectrometry-based studies although other endoproteinases are often used. For example, a recent high coverage study of the yeast proteome used Lys-C (De Godoy et al., 2008). We calculated the theoretical digested human proteome for three other enzymes in reasonably common use, as shown in Table 2.5. Since the proteotypic predictors are trained for trypsin we applied only the redundancy and length filters to the peptides (>5 and <40 amino acids) and used the relative fraction of intron spanning peptides as a proxy for the ability to inform on AS. The alternate proteases theoretically offer a greater chance of identifying a peptide which spans an exon junction and therefore could inform on a particular splice variant, particularly Lys-C and Arg-C. These peptidases have the advantage of generating larger peptides which have a greater chance of spanning an intron and informing on splicing events. Moreover, they retain a basic amino acid at their C-termini which is consistent with tryptic peptides and is conducive to expected fragmentation patterns in the mass spectrometer. This should be taken into account when designing experiments.

Table 2.5. Suitability for alternative peptidases for AS characterisation in humans

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Number of peptides after length &amp; redundancy filtering</th>
<th>Mean peptide length</th>
<th>% intron spanning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Limit peptides</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trypsin</td>
<td>545670</td>
<td>13.7</td>
<td>20.4%</td>
</tr>
<tr>
<td>Endoproteinase Arg-C</td>
<td>339760</td>
<td>17.2</td>
<td>27.2%</td>
</tr>
<tr>
<td>Endoproteinase Lys C</td>
<td>325073</td>
<td>16.6</td>
<td>25.5%</td>
</tr>
<tr>
<td>Endoproteinase V8</td>
<td>242408</td>
<td>15.9</td>
<td>17.9%</td>
</tr>
<tr>
<td><strong>Peptides upto and including 1 missed cleavage</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trypsin</td>
<td>1400266</td>
<td>15.7</td>
<td>24.4%</td>
</tr>
<tr>
<td>Endoproteinase Arg-C</td>
<td>692072</td>
<td>19.2</td>
<td>29.8%</td>
</tr>
<tr>
<td>Endoproteinase Lys C</td>
<td>566660</td>
<td>18.6</td>
<td>28.7%</td>
</tr>
<tr>
<td>Endoproteinase V8</td>
<td>620864</td>
<td>17.9</td>
<td>24.2%</td>
</tr>
</tbody>
</table>

In this study, we have not restricted our calculations and have therefore used the so-called “one-hit wonders” (Veenstra et al., 2004), where a protein isoform may be characterised by a single peptide identification. This is an inherent problem facing studies
of this type, where only a single peptide could be diagnostic for a particular isoform. Care
must therefore be taken when interpreting such data, and where possible multiple peptides
(perhaps from different proteolytic cleavage agents) could be used to provide higher
confidence. However, when treated properly considerable added value can be gained from
a single PSM (Gupta and Pevzner, 2009).

A further consideration is the database to be searched. A recent HUPO study
highlighted the problems using databases containing redundancy for peptide identification,
leading to inconsistencies between laboratories using the same sample (Bell et al., 2009).
This suggests a gene-centric database is optimal with a single descriptive name of each
gene. This would be acceptable for isoformal proteomics, so long as all the alternately
spliced peptides were represented. Currently, no such database exists and we concur that
greater effort is required to generate a suitable database which contains all the relevant AS
information but is minimally redundant.

2.3.5. Isoform detection and quantification using QconCATs and SRM

QconCAT technology combined with SRM has been successfully applied to
measure levels of low-abundance proteins such as cytokines (Bislev et al., 2012), and to
generate pathway models by measuring the concentrations of the pathway components
(Carroll et al., 2011). This approach is also ideal for the detection and quantification of
splice-variants, which are often lowly expressed, and therefore require highly sensitive
techniques to measure their expression. However, applying this technology to quantify
specific splice variants at the protein level poses new challenges because often splice
variants differ only by a single exon, which greatly reduces the number of discriminative
peptides that are unique to a particular isoform. To assess the feasibility of this, we
designed an experiment to detect and measure the expression of splice variants of the
Drosophila voltage-gated sodium channel DmNav. In Drosophila, this ion channel is
encoded by the paralytic (para) gene which is subject to extensive alternative splicing. A
total of 53 full-length cDNAs representing different splice variants have been sequenced
by two separate studies, which are generated through the inclusion/exclusion of twelve
cassette exons (Lin et al., 2009; Olson et al., 2008). It has been shown that the differential
expression of these splice variants gives individual neurones their characteristic electrical properties and that overexpression of exon L can lead to seizures in fruitfly. Indeed, the human orthologs of para are a key target for anti-epileptic drugs (Mantegazza et al., 2010), making the Drosophila gene an important model for studying epilepsy. Figure 2.8 shows the complexity of splicing in the para gene and the relationship between each alternative exon and DmNav domain architecture.

Figure 2.8. Splicing complexity of the Drosophila para gene.

The para gene encodes fifty-three isoforms of the voltage-gated sodium channel DmNav. The exon-intron structure of each splice variant present in FlyBase is shown, and the alternative exons targeted by the QconCATs are marked. The alternative exons most often result in the addition/deletion of sequences in the cytosolic loops (DmNav protein architecture taken from (Lin et al., 2009)).
Each DmNav splice variant has a unique combination of the twelve alternative exons shown in Figure 2.8. The ratio of each exon in the expressed mRNAs is indicative of the electrical properties of the neuron (Lin et al., 2009). However, exon inclusion ratios inferred from full-length cDNA sequencing only provide an indirect measure of expression at the protein level, and the development of an anti-epileptic drug would benefit from the absolute quantification of DmNav isoforms in epilepsy patients. Therefore, an in silico tryptic digest of all the DmNav isoforms was used to identify a set of peptides that could potentially be used to quantify specific, or groups of protein isoforms. Appendix 3 shows the list of the 29 tryptic peptides chosen to construct a QconCAT for the absolute quantification of para isoforms. None of these peptides were present in PeptideAtlas, which likely reflects the difficulties in detecting transmembrane proteins using shotgun proteomics (Gilmore and Washburn, 2010). The missed cleavage propensity of each tryptic peptide within the context of the QconCAT protein sequence is shown in Appendix 3. Peptides having a 50% chance of being missed cleaved were excluded to give an average of 2.1 peptides per alternative exon. Ideally, a greater number of peptides per exon with lower missed cleavage propensities would be preferable to ensure reliable quantification, but this was precluded by the small size of the targeted AS exons. Moreover, a large percentage of basic residues present in the AS exon regions are likely to preclude a complete digest. This again highlights the need for the use of multiple types of endopeptidase to expand the repertoire of peptides that can be used to target the alternative exons.

Appendix 4 shows a second QconCAT designed for a Lys-c digestion. This increased the number of peptides available to monitor the expression of splice variants, and thus allowing a more reliable quantification of splice variants. However, the current missed cleavage predictor only considers tryptic peptides, and hence further work is needed to develop tools to predict missed cleavages for different types of enzyme. At present, only a single trypsin QconCAT has been synthesised. We hope that this initial work could provide the foundations for a targeted, gene-centric approach to investigate the Drosophila proteome, and specifically uncover the isoformal portion of the proteome.
2.4. Conclusions

Proteomics has a vital part to play in the annotation of genomes, since it is the principal technology applied to characterise proteins on a genome-wide scale. Without this evidence, the mature functional forms of the gene products expressed and translated in cells and tissues remain as “predicted”. To address this, a targeted proteomics strategy is required which selects diagnostic peptides that are able to unambiguously identify given isoforms and are also amenable to mass spectrometric study. Similarly, practitioners are encouraged to consider multiple databases against which to search (i.e. Ensembl and Swiss-Prot) which offer differing levels of annotation quality and coverage of AS events. Alternatively, although not formally evaluated here, the IPI database (http://ebi.ac.uk/IPI/) merges Ensembl, UniProt and RefSeq and offers a database which attempts to capture the coverage of predicted protein sequences available from completed genome sequences with the accuracy of the high quality annotations available from UniProt and RefSeq.

2.5. References


Chapter 3: Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies

As shown in the previous chapter, splicing patterns can be confirmed by mapping peptide identifications to gene models, and the most informative peptides are useful for targeted proteomics experiments. However, this approach to proteogenomic mapping is restricted to existing gene annotations. A wealth of additional information exists in EST and RNA-Seq datasets which can be used to expand or amend gene models, or even uncover entirely new genes. Mass spectrometry database searches against the ESTs can provide an additional layer of evidence, and allows the identification of protein-coding sequences from a pervasively transcribed genome. There are a number of approaches that can be used to map proteomics data to transcript sequences, which typically involve searching against conceptual translations of the EST sequences. The following chapter investigates which of these conceptual translation approaches results in the most sensitive database search in terms of the number of significant peptide identifications generated. Various measures of statistical significance are compared and used to assess the performance of each method. We highlight problems associated with the six-frame translation method which is commonly used in proteogenomics studies. We then go on to suggest modifications to six-frame databases to make them more compatible with the target-decoy approach for measuring statistical significance.

The contents of this chapter were published as a research article in Journal of Proteome Research (Blakeley, P., Overton, I. M., Hubbard, S. J. (2012). Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. Journal of Proteome Research, 11(11), 5221-5234). The author of this thesis was the first author of this research article and was responsible generating the results and figures. The EORF algorithm was written by Ian Overton, who now works at the MRC Human Genetics...
Unit, University of Edinburgh. The EORF protein sequence database was constructed by Ian Overton. Simon Hubbard directed the work and formulated most of the main ideas.

Abstract

Proteogenomics has the potential to advance genome annotation through high quality peptide identifications derived from mass spectrometry experiments, which demonstrate a given gene or isoform is expressed and translated at the protein level. This can advance our understanding of genome function, discovering novel genes and gene structure that have not yet been identified or validated. Because of the high-throughput shotgun nature of most proteomics experiments, it is essential to carefully control for false positives and prevent any potential misannotation. A number of statistical procedures to deal with this are in wide use in proteomics; calculating false discovery rate (FDR) and posterior error probability (PEP) values for groups and individual peptide spectrum matches (PSMs). These methods control for multiple testing and exploit decoy databases to estimate statistical significance. Here, we show that database choice has a major effect on these confidence estimates leading to significant differences in the number of PSMs reported. We note that standard target:decoy approaches using six-frame translations of nucleotide sequences, such as assembled transcriptome data, apparently under-estimate the confidence assigned to the PSMs. The source of this error stems from the unusual nature of the six-frame database, where for every target sequence there exists five “incorrect” targets that are unlikely to code for protein. The over-conservative FDR and PEP estimates lead to fewer PSMs and we show that this effect is a product of the database and statistical modelling and not the search engine. A variety of approaches to limit database size and remove non-coding target sequences are examined and discussed in terms of the altered statistical estimates generated and PSMs reported. These results are of importance to groups carrying out proteogenomics, aiming to maximise the validation and discovery of gene structure in sequenced genomes, whilst still controlling for false positives.
3.1. Introduction

Rapid advances in mass spectrometry-based proteomics have been made possible by improvements in peptide separation techniques, high-resolution instruments and downstream informatic processing. Researchers can now obtain a comprehensive catalogue of the proteome in single-celled organisms (Nagaraj et al., 2011) and a near-comprehensive catalogue in multicellular organisms (Schrimpf et al., 2009). However, the database search method that underpins the majority of proteomic workflows typically requires a high quality set of protein-coding genes. To improve proteome coverage, mass spectrometry (MS) data can be searched against protein sequences inferred from either the genome (Chaerkady et al., 2011; Castellana et al., 2008; Merrihew et al., 2008; Baerenfaller et al., 2008) or transcriptome (Wang et al., 2011; Ching et al., 2011; Edwards, 2007; Robinson et al., 2009; May et al., 2008). Such proteogenomic approaches are not biased towards existing gene annotations and therefore offer scope for novel gene/protein discovery. Indeed, proteogenomics has led to the discovery of thousands of novel gene candidates (Castellana et al., 2008; Merrihew et al., 2008; Adamidi et al., 2011), protein isoforms (Brosch et al.; Tanner et al., 2007; Brosch et al., 2011), amino acid polymorphisms (Kalume et al., 2005; De Souza et al., 2011), and confirmation and correction of gene models (Baudet et al., 2010; Blakeley et al., 2010; Desiere et al., 2005; Findlay et al., 2009; Merrihew et al., 2008; Prasad et al., 2011). A recurring observation in these studies is that current gene models, which are largely computational predictions themselves, are often incomplete and erroneous. For instance, lowly expressed splice variants and non-canonical genes often prove difficult to annotate (Findlay et al., 2009; De Souza et al., 2010).

However, the database size and attendant search space when searching raw genomes, particularly metazoan ones, is usually dramatically inflated. For example, a translation of the human genome in all six reading frames would result in a huge search space: typically thousands of LC-MS/MS spectra would need to be searched against at least $6 \times 10^9$ amino acids. Moreover, eukaryotic genomes are also heavily populated by non-protein-coding regions, introns and hitherto un-annotated splice variants. Although it is possible to search directly against translated genomic sequence, this is clearly a challenging task that requires careful quality control to avoid false positives and integration of PSMs over a genomic locus.
An alternative approach is to search against expressed sequence tags (ESTs) generated from traditional Sanger sequencing (Edwards, 2007; Robinson et al., 2009) or from next-generation sequencing (Wang et al., 2011; Adamidi et al.; Adamidi et al., 2011) since they are by definition transcribed (and likely translated) regions of the genome. For example, this approach has been successfully applied to UniGene clusters by generating a six-frame translation and compressing the resulting protein sequence database to remove redundancy (Edwards, 2007). This allowed the discovery of non-synonymous mutations, splice-variants, micro exons and alternative translation reading frames, some of which could not be identified from a direct genome search. An alternative method for reducing the search space involves translating the ESTs into proteins using probabilistic approaches such as ESTScan2 (Iseli et al., 1999), DECODER (Fukunishi and Hayashizaki, 2001) and FrameDP (Gouzy et al., 2009). For example, Robinson and colleagues (Robinson et al., 2009) identified secretory proteins involved in helminth pathogenesis by searching spectra against protein sequences predicted by ESTScan2, which uses a hidden Markov model to distinguish CDS from untranslated regions (UTRs) and correct potential sequencing errors. Such studies have shown that searching against a concise database enriched in protein-coding sequences can increase sensitivity.

Controlling for false positive identifications is essential in high-throughput proteomics studies and particularly so for proteogenomics where intergenic sequence, introns and UTR together can constitute most of the database. Several database search tools such as Mascot (Perkins et al., 1999), Inspect (Tanner et al., 2005), X-Tandem (Craig and Beavis, 2004) or Sequest (Eng et al., 1994) are commonly used to search spectra against a six-frame translated nucleotide sequence (Sevinsky et al., 2008; Borchert et al., 2010; Bindschedler et al., 2009; Edwards, 2007; Brunner et al., 2007; Gupta et al., 2008; Merrihew et al., 2008; Castellana et al., 2008; Jaffe et al., 2004; Baudet et al., 2010), but the output scores for candidate PSMs cannot be compared directly across experiments. Indeed, it is widely accepted that applying E-value thresholds is anti-conservative (Kall et al., 2008b) and different search engines estimate quite different significance levels (Kapp et al., 2005). Instead, the global error rate is typically estimated by ordering the PSMs according to their E-value or search engine-specific score and calculating the false discovery rate (FDR) (Storey, 2002; Kall et al., 2008a; Elias and Gygi, 2007). The FDR is estimated from the percentage of incorrect PSMs at a given threshold, which is usually
calculated using a target-decoy approach. The exact details of this vary and much discussion in the field exists as to the best approach (Gupta et al., 2011; Gupta and Pevzner, 2009; Nesvizhskii, 2010; Wang et al., 2009; Granholm and Kall, 2011), although the basic principle is well established. This involves searching the spectra against both a target database of ‘real’ sequences and a decoy database containing ‘fake’ sequences produced by reversing or randomising the target sequences; hits to the latter are considered to be false and are used to estimate the level of incorrect target PSMs at a given score threshold. Although the FDR applies globally to a collection of PSMs, individual PSMs can be assigned a q-value (Storey, 2002). The q-value is applied because two or more Mascot scores can give the same FDR. The q-value is the minimum FDR at which a PSM accepted, which ensures that each PSM is uniquely mapped to a FDR.

In contrast to the global error rate, a ‘local’ FDR termed the Posterior Error Probability (PEP) is also frequently estimated as the probability of an individual PSM being incorrect. For example, the software tool Qvality (Kall et al., 2008c; Kall et al., 2009) implements a non-parametric approach for calculating the PEP by estimating the proportion of the target score distribution that is incorrect given a set of p-values or a decoy score distribution. Similarly, PeptideProphet (Keller et al., 2002) can be used to calculate PEPs and can incorporate decoy search results to enable semi-parametric modelling and therefore greater flexibility (Choi and Nesvizhskii, 2008b).

Both the FDR and PEP approaches exploit target/decoy database search results to assign statistical significance to PSMs, presuming that the target database accurately represents genuine protein sequences and the decoy database is of equal (or known) size and similar redundancy (Wang et al., 2009). In this study we examine whether these criteria are indeed met in the context of proteogenomics experiments. Specifically, we highlight the problems associated with the standard target-decoy approach for assessing the statistical significance of PSMs assigned to predicted protein sequences derived from a large collection of Chicken ESTs. We consider searches against six-frame translations and single-frame predicted protein translations, comparing different approaches to estimate the statistical significance of the PSMs. Search results are highly dependent on database choice and suggest potential pitfalls when searching six-frame databases linked to database size and target-decoy error modelling. We believe this leads to over-conservative significance estimates for six-frame translation databases and a reduction in sensitivity – and hence
fewer confident PSMs and peptides. We investigate the source of the errors, biases in six-frame translation databases, and suggest a variety of approaches to address these problems. This can be achieved by modifying the six-frame database or by conceptually translating the EST sequences. The results have significance for any group carrying out proteogenomic searches against genomic or transcriptome-based sequences.

3.2. Materials and Methods

The mass spectrometry data used in this Chapter was also used in Chapter 2 (described in section 2.2.3).

3.2.1. EST dataset

A total of 339,314 ESTs were sequenced from 64 cDNA libraries derived from 21 chicken tissues, and then clustered and assembled using BLASTN and PHRAP to generate 85,486 EST contigs (Boardman et al., 2002). These are available via http://www.chick.umist.ac.uk. Two different sets of protein sequences were predicted from each EST contig using the EORF (described in 3.2.4.) and ESTScan2 (Iseli et al., 1999) algorithms.

3.2.2. Preparation of chicken samples and mass spectrometry

We used a comprehensive dataset of peptide spectra generated for an unrelated DT40 project, kindly provided by colleagues at the University of Cambridge (Kathryn Lilley, personal communication). The MS/MS data were derived from a proteomic analysis of the DT40 chicken cell line which used the LOPIT protocol, and was originally published in 2009 (Hall et al., 2009). Full details are available in the original paper (Hall et al., 2009), but briefly, DT40 cells were fractionated by density gradient centrifugation and 7 fractions were chosen for analysis. The selected fractions were labelled with four-plex iTRAQ reagents and digested with trypsin. Labelled peptides were pooled together and separated using two-dimensional liquid chromatography. LC-MS/MS was performed using
an ultimate-nano-LC system (Dionex) coupled to a QSTAR XL mass spectrometer (Applied Biosystems). The QSTAR XL was operated in information-dependent acquisition mode in which 1 s MS scans were performed (m/z 400–1600) followed by 3 s product ion scans (m/z 100–1580) on the two most intense doubly and triply charged peptides. The LOPIT protocol and iTRAQ labelling were incidental to our study which focuses solely on the relative merits of database composition and attendant statistical treatments to peptide identification.

3.2.3. Databases

Ensembl protein sequences were downloaded in fasta format from the Ensembl FTP server (ftp://ftp.Ensembl.org). Version 56 of the ‘pep.all’ set containing translations from known and novel genes was used. The UniRef90 database (release 15.11) was downloaded from (http://www.uniprot.org/downloads). A custom Perl script was used to generate the six-frame translation database from 85,486 EST contig sequences, generating 3 forward and 3 reverse-frame sequences using the standard genetic code. Protein sequences were also predicted from the EST contigs using the ESTScan2 and EORF programs. Briefly, EORF calculates a score for each reading frame based on its codon usage bias and sequence homology derived from BLAST searches against UniProt, and was used to predict 67,125 ORFs from the EST contigs (a small fraction were rejected as they are unlikely to be coding). ESTScan2 (Iseli et al., 1999) uses a hidden Markov model to predict the correct reading frame and UTRs. ESTScan2 predicted a total of 62,161 protein sequences from the ESTs (again rejecting non-coding ESTs).

3.2.4. EORF description

A Needleman-Wunsch-based (Needlema.Sb and Wunsch, 1970) dynamic programming approach forms the core of the EORF algorithm. The methodology can be conceptualised as the alignment of the EST nucleic acid sequence to the six possible protein translations of that sequence. Every codon position in each of the 6 possible reading frames of the EST nucleotide sequence is assigned a score based on the
synonymous codon bias for the organism in question. This codon position score is calculated according to equation 1.

\[ S_i = M \log_{10} \left( \frac{C}{Z_i} \times (1 - E_1)(1 - E_2)(1 - E_3) \right) - U \]

Where \( S \) = codon position score
\( i \) = codon position in the sequence (for a particular reading frame)
\( M \) = matrix multiplier value (default 100)
\( C \) = Synonymous codon bias probability
\( Z \) = Mean (unbiased) probability of codon usage for the amino acid in question (e.g. if there are 4 possible codons for that amino acid \( Z = 0.25 \))
\( E_1 \) = Error probability associated with codon position 1
\( E_2 \) = Error probability associated with codon position 2
\( E_3 \) = Error probability associated with codon position 3
\( U \) = Stop Codon Penalty (default 60, has non-zero value only if codon specifies ‘STOP’ signal)

The stop codon penalty is implemented as a heuristic measure based on the observation that stop codons are largely absent from the correct reading frame. Incorporation of PHRED/PHRAP sequence quality scores into the dynamic programming matrix allows the codon position scores associated with sequencing errors to be down-weighted according to the calculated probability of sequencing errors in the trinucleotide codon.

EORF can determine the most likely strand for protein translation using the synonymous codon usage across the 6 possible reading frames, from user input, or from BLAST information. A two-dimensional dynamic programming matrix is constructed for a sequence of \( n \) codons with codon position on the X-axis and the reading frame on the Y-axis, for the 3 possible reading frames associated with the selected direction of translation.
(i.e. forward or reverse). Equation 2 is used to calculate the score of a matrix cell $F(x, y)$. The dynamic programming matrix is written out for the reverse strand if a top-scoring BLASTX alignment is observed in the opposite sense to the EST sequence with an expectation value better than the command line specifiable threshold (default 1E-20). Alternatively, if BLAST input data is not available, the direction of translation is determined by calculating dynamic programming matrices for both the forward and reverse strand and selecting the matrix (i.e. direction of translation) according to the highest value of $F(n, y)$ where $n = \max(x) - 3$.

**Equation 2:**

$$F(x, y_a) = \begin{cases} 
(F(x-1, y_a) + S(x, y_a)) + (-\log B) * V \\
(F(x-1, y_b) + S(x, y_a) - Q) + (-\log B) * V \\
(F(x-1, y_c) + S(x, y_a) - Q) + (-\log B) * V 
\end{cases}$$

Where $Q = \text{Frame-shift penalty (default 66)}$

$y_a, y_b, y_c$ refer to the 3 possible reading frames, $y_a$ is the frame of the matrix cell $(F(x, y_a))$ for which a score is calculated

$S(x, y_a) = \text{The codon position score for the frame in which the current matrix cell score is calculated}$

$B = \text{BLASTX expectation value for the alignment corresponding to frame } y$

$V = \text{BLAST Multiplier (default 10)}$

Inclusion of BLASTX (Altschul *et al*., 1990) data effectively increases the codon position score ($S_i$) for the relevant frame(s) in the EST sequence region that matches a protein. The BLASTX weighting is proportional to the expectation value associated with the given sequence alignment. Only alignments with expectation values better than the defined threshold (default 1E-6) are incorporated into the EORF algorithm’s predictive
mechanism. For sequences with an expectation value of 0.0 the weighting value (in equation 2) defaults to 1E-3000, however this default may be altered as a command line argument.

Once the dynamic programming matrix has been constructed, the highest $F(n, y)$ ($n = \max(x) - 1$) is used as a starting point for tracing a path back through the matrix according to the directionality associated with each $F(x, y)$. The path is traced back to an $F(x, y)$ having the smallest value of $x$ and defines the maximum possible predicted coding region for that nucleotide sequence. This path is then translated to amino acid sequence according to the genetic code provided in the synonymous codon usage table. EORF enables estimation of the coding potential for EST contigs, according to two metrics. Firstly, combination of the length-normalised final matrix score, and secondly the interval between the highest and lowest matrix traceback values. EORF also generates a file with sequences where the matrix traceback did not converge to a single finishing point, which is indicative that there is no clear protein coding region in the sequence. Some sequence positions are indicated as “N” (any base) in EST data, which can result from ambiguities in the trace files and sequence masking in the assembly process. There are a finite number of possible codons for tri-nucleotides containing one or two “N” bases. The average synonymous codon bias for these codons is calculated and incorporated into the synonymous codon bias table.

3.2.5. Mass spectrometry database searching

A total of 403,820 centroided spectra were searched against protein sequences derived from the ESTs using Mascot version 2.0, with a precursor MS error tolerance of 0.2 Da and MS/MS error tolerance of 0.8 Da. Fixed modifications were Cysteine beta-methylthiolation and iTRAQ labelling of Lysine residues and N-termini. Variable modifications were iTRAQ labelled Tyrosine and Methionine oxidation. Up to 1 missed cleavage was permitted.

Four different databases were searched: Ensembl version 56, EORF predictions, ESTScan2 predictions, and the six-frame translations. In addition, various combinations of EORF, ESTScan2 and six-frame sequences were searched to find the database which
allows for the most PSMs. Typically, the majority of PSMs for a database search are incorrect, hence it is important to accurately predict which PSMs represent real peptides present in the samples. For statistical evaluation of the data, decoy databases were constructed by reversing each protein sequence in the original ‘target’ database. Two types of target-decoy searches were performed: separate and composite. In the composite search, the decoy database was concatenated onto the target database and then used as a single database for the Mascot searches, whereas the separate search involved independent searches against target and decoy databases.

3.2.6. Statistical methods for validating PSMs

From composite (concatenated) searches, FDRs were calculated using the methods published by Ellias and Gygi (Ellias and Gygi, 2007), and Käll et al. (Käll et al., 2008a) to estimate the proportion of false positive PSMs that have accumulated at a given Mascot score giving $FDR_{EG}$ and $FDR_{Käll}$ respectively. The q-value was calculated as the minimal FDR at which a PSM is accepted. The $FDRScore$ method (Jones et al., 2009) was used to combine PSMs from the EORF and ESTScan2 searches. For each of the two database searches, a custom perl script was used to assign q-values against Mascot scores, to identify step points (where the q-value increases). From this, a linear regression was calculated between each step point. The $FDRScore$ was calculated for each Mascot score between the step points, according to the gradient of the line. PSMs common to both searches were then merged by calculating the geometric mean of their $FDRScores$. The PSMs were then resorted by their Average $FDRScores$ to calculate a new set of q-values, from which a second $FDRScore$ was calculated, termed the combined $FDRScore$.

In parallel to the composite database estimates, the Mascot scores for separate database searches were used to calculate the local FDR, or PEP, using the software tool Qvality (Käll et al., 2008c; Käll et al., 2009). Default parameters were used to generate a set of PEPs and PEP-derived q-values linked to Mascot scores.
3.2.7. Estimating the proportion of correct PSMs

EST contigs and their attendant reading frames were assigned to Ensembl proteins via BLASTX searches against the Ensembl 56 database. Assignments were made for the top scoring hits which passed the following cut-offs: Identity > 95%; Coverage > 50 residues; E-value < 0.001. EST contigs with significant hits were extracted along with the top scoring reading frame. We assumed that the highest scoring reading-frame contained the correct ORF. This information was integrated with the six-frame translation database, to identify the sequences in the ‘correct’ reading frame owing to the BLASTX match. PSMs having a match with the correct reading frame were then assumed to be correct.

The probability distributions of the amino acid frequencies were calculated for each reading frame and compared with the probability distributions for the Ensembl and UniProt90 protein sequences. The amino acid frequency of the entire correct reading frame set was calculated separately from the incorrect reading frames. The Mann-Whitney U test was used to measure the degree of divergence between the amino acid distributions of the correct frames with each of the incorrect frames different distributions.

3.3. Results and Discussion

3.3.1. Searching against Six-frame or redundant databases affects sensitivity.

When searching high-throughput mass spectrometry data against a protein database using a target-decoy strategy, it is usually the case that the target database is composed of genuine protein sequences that could be present in the sample, and random (false) matches to target and decoy database are equally likely. However, when searching against a six-frame database these conditions are not necessarily met. One anomaly is that at most only one out of the six possible reading frames translated from a nucleotide sequence is likely to be coding (presuming that there is only one protein coding ORF at any given locus) and can lead to “true” target PSMs. This is illustrated in Figure 3.1 for a hypothetical EST sequence that codes for a protein, along with its six conceptual translations (forward frames 1, 2, 3 and reverse frames 4, 5, 6) and the attendant six decoy frames. Only the
target frame A2 contains a true coding sequence and is matched by a single PSM in this case. Frame B1 and B3 are in the correct direction (and could contain correct PSMs if there were a frameshift mutation in the nucleotide sequence) whilst frames C4-C6 are not. Although the target and decoy databases contain the same number of sequences, amino acid composition and tryptic peptides, the five “wrong” target frames are not likely to be protein-like. This expansion of the target database is in principle similar to simply adding more proteins (perhaps from another species) to the database, but in this case the additional targets are clearly “wrong” and not likely to be protein-like in composition. Moreover, the expansion is particularly big, adding five extra sequences for every original target. This could confound the assumptions normally held for target-decoy FDR calculations and lead to incorrect statistical modelling (Granholm and Kall, 2011).

![EST contig](image)

**Figure 3.1.** Schematic of EST translation for target:decoy database generation.

Translation of transcriptome data such as ESTs in all six reading frames increases the proportion of ‘junk’ sequence. In this simplified model, only one of the six reading frames is correct (sequence A in frame 2). Sequences denoted by “B” are in the correct direction and therefore in some circumstances could constitute part of the correct ORF as a result of pre-mRNA splicing or frame-shift errors. Sequences denoted by “C” are in the wrong direction and are therefore incorrect. Decoy sequences are created by reversing the six corresponding target six sequences, so that decoy1 is the reverse of B1, decoy 2 the reverse of A2, and so on.
To test this, we searched tandem mass spectra derived from a chicken DT40 cell line against different protein databases generated from assembled EST contigs, applying a variety of FDR and PEP-based confidence measures to generate significant PSMs. For FDR estimation, two widely-used approaches using concatenated target-decoy databases were used to generate FDR$_{\text{Käll}}$ (Kall et al., 2008a) and FDR$_{\text{EG}}$ (Elias and Gygi, 2007) estimates with attendant q-values. Finally, the PEP and PEP-derived q-values were also estimated from separate database searches using Qvality (Kall et al., 2009). As described in the methods, five different protein sequences databases were generated from the EST contigs and assessed by the number of PSMs accepted at a consistent threshold (q-value/PEP<0.01) for the different confidence measures (Table 3.1). Decoy sequences were created by reversing the target sequences (see also Methods).

Table 3.1. Unique peptide identifications at q-value/PEP < 0.01 for different database searches.

<table>
<thead>
<tr>
<th>Database</th>
<th>FDR$_{\text{EG}}$ q-value</th>
<th>FDR$_{\text{Käll}}$ q-value</th>
<th>PEP-derived q-value</th>
<th>Qvality PEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESTScan2</td>
<td>10519</td>
<td>12291</td>
<td>10721</td>
<td>7323</td>
</tr>
<tr>
<td>EORF</td>
<td>9606</td>
<td>11502</td>
<td>9290</td>
<td>7143</td>
</tr>
<tr>
<td>six-frame translations</td>
<td>6730</td>
<td>8620</td>
<td>7328</td>
<td>4951</td>
</tr>
<tr>
<td>EORF + ESTScan2</td>
<td>9702</td>
<td>11466</td>
<td>9854</td>
<td>6962</td>
</tr>
<tr>
<td>EORF + ESTScan2 + six-frame</td>
<td>6778</td>
<td>8616</td>
<td>7405</td>
<td>5020</td>
</tr>
<tr>
<td>EORF + ESTScan2 (FDR$_{\text{Score}}$)</td>
<td>11532</td>
<td>13813</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.1 shows that considerably fewer PSMs were accepted when searching the six-frame database at a q-value/PEP cut-off of 0.01, in comparison to the other databases searched. This observation is independent of the chosen measure of statistical significance, suggesting it is a problem specific to the six-frame database searches (whether concatenated or separate). This reduced sensitivity can be attributed to several factors. Firstly, the six-frame database is more than seven times greater in size than the other
databases, and it is known that large databases can lead to more conservative statistical estimates (Fitzgibbon et al., 2008; Granholm and Kall, 2011) (or conversely, small database size leads to over-estimates of significance and likely false positives). Secondly, the unusual nature of the six-frame database could be confounding and inflating the FDR. In all likelihood, only one of the six forward frames is “correct” and can be matched by genuine peptide spectra. This could lead to an effective imbalance in the true “target” sequences and false “decoy” sequences, since at least five of the target frames are also likely to be wrong, which in turn compromises FDR estimates.

A further possible explanation for the poor PSM sensitivity observed for the six-frame searches could be derived from the ESTScan2/EORF translations correcting errors such as frameshifts, leading to peptides that are absent from the six-frame database. To test this, using the FDR\textsubscript{Käll} metric, we compared unique peptide sequences from ESTScan2 PSMs to those derived from six-frame PSMs, in Figure 3.2a. Only 389 peptides (~3% of ESTScan2 peptides derived from PSMs with q-values less than the 0.01 threshold) are unique to the ESTScan2 database searches. Most of these are expected to come from the translational corrections applied by ESTScan2. Whilst these peptides do contribute to the higher sensitivity of the ESTScan2 search, their overall contribution is minimal. In contrast, 11902 peptide sequences derive from PSMs found in both databases and 7917 of these have PSM q-values less than the 0.01 significance threshold. Hence, 3985 peptides are “lost” in the six-frame search despite having PSMs in the Mascot output with the same score but with q-values > 0.01. Thus the majority (91%) of the additional ESTScan2 peptides have arisen as a result of the inflated q-values associated with the six-frame target-decoy error modelling as opposed to correction of frame shift errors. It should also be noted that some 703 peptides are exclusive to the six-frame search as ESTScan2 did not predict a coding sequence, and thus could not be matched by any spectra. These general trends are matched when considering data at the PSM level instead of the peptide level (see Appendix 5).
Figure 3.2. Overlap of peptides identified in pairwise database searches.

Overlap of unique peptide sequences derived from PSMs in the searches against: a) the ESTScan2 and six-frame databases, b) ESTScan2 and EORF databases. In both cases, \( FDR_{\text{Kall}} \) q-value cut-offs of 0.01 for the various searches are indicated by dotted lines, black for ESTScan2 and white for the six-frame or EORF searches. PSMs are sorted by Mascot score from low scores (bottom) to high scores (top). The majority of the unique accepted peptides identified in ESTScan2 but missed by the six-frame database were present in both databases but have q-values that exceed the six-frame q-value threshold.
In contrast, fewer peptide identifications are “lost” when comparing ESTScan2 and EORF search results, shown in Figure 3.2b. In this case, ESTScan2 generates more accepted PSMs and attendant unique peptides, but only 1040 additional peptides are “lost” to EORF owing to different FDR modelling whilst 10810 are shared. The same trend is observed at the PSM level, where only 2% of accepted ESTScan2 PSMs were missed by EORF due to FDR differences, and 87% of accepted ESTScan2 PSMs were also accepted by EORF (see Appendix 5).

We also note that combining the EORF and ESTScan2 databases leads to fewer PSMs at an equivalent q-value/PEP threshold, as shown in Table 3.1. This could be due to increased size as well as redundancy when combining similar databases (i.e. EORF and ESTScan2) where target protein sequences are more likely to be shared. Redundancy of target sequences is already known to degrade search quality; this can be addressed by merging highly similar sequences into a single entry and appending sequence variants in a compatible fashion for MS database searches (De Souza et al., 2010). One potential downside of this approach, however, is the potential loss of subtle variations that represent different biological entities that could be important for proteogenomics.

To circumnavigate this redundancy issue and integrate PSMs from separate ESTScan2 and EORF searches, we used the FDRScore method (Jones et al., 2009) that was initially designed to merge results from multiple search engines (Jones et al., 2009; Kwon et al., 2011; Alves et al., 2008). This approach removes redundancy at the PSM level by combining FDRKäll q-values from matched PSMs across two database searches and models database-specific PSMs independently, to derive an integrated FDR estimate. This resulted in 1522 additional PSMs compared to ESTScan2 alone and 5193 compared to the six-frame database. For our spectral dataset this results in the largest number of significant PSMs from all the approaches considered at the same nominal significance threshold of 0.01. However, this is only applicable to the FDRScore method when used to calculate q-values for concatenated target-decoy databases.

The relationship between q-value/PEP and Mascot score was further investigated in order to explain the poor sensitivity of the six-frame translation searches. The PEP, PEP-derived q-values, and empirical q-values (FDRKäll) were plotted against Mascot scores for each database search (Figure 3.3). Figure 3.3a confirms that the Mascot ion score for
equivalent PSMs is independent of the database searched, a relationship that holds for all the paired searches we ran and supports comparisons across the different database searches. Figure 3.3 confirms that the PEP is clearly the more conservative approach, and that, independent of the significance measure used, the six-frame translation PSMs have higher q-values/PEPs compared to the other database searches.

Biases in the target-decoy database construction methods for FDR calculations have been noted before, which can lead to effective decoy database sizes larger than the target (Wang et al., 2009). We contend that standard six-frame translation databases may also suffer from similar problems leading to inflated q-values and PEPs, as observed in Figure 3.3 The extent of this inflation can be gauged by comparing the maximum q-values of the different database searches. For example, Figure 3.3d shows that the maximum PEP-derived q-value (the y-intercept) for the six-frame translation PSMs is much greater than the equivalent q-values for the EORF and ESTScan2 PSMs; approximately 0.8 compared to 0.53. Similarly, the maximum FDR_{Käll} q-value for the six-frame database search is much larger than the equivalent q-values from other searches (Figure 3.3b), and the same trend is also observed for the PEP (Figure 3.3c). The differences in these profiles show the effect that the choice of error model and database has on search results for high-throughput proteomics.

Collectively, these results point to the statistical modelling and not the search engine scoring that lead to a sensitivity reduction in accepted PSMs at a fixed statistical threshold for six-frame searches, and highlight the importance of redundancy removal and careful database design for estimating PSM statistics in proteogenomics. However, although we observe a broad range of FDR and PEP estimates for the same PSMs we still do not know which ones are closest to the truth; although one might presume the six-frame values are overly conservative. We try to address this in the next section.
Figure 3.3. Variation of search statistics with Mascot score.

Plots show the calculated q-values and PEPs for PSMs from different proteogenomic database searches and their dependence on Mascot ion score. In a) Mascot Scores of equivalent PSMs from two independent database searches are plotted, in this case ESTScan2 vs. six-frame, although identical plots were obtained for all pairwise comparisons. In b) q-values calculated using FDR_{Kall} are plotted against Mascot ion score, c) PEPs calculated using Qvality, and d) q-values calculated from Qvality, for different database search combinations. In the key, 6F denotes the six-frame searches.
3.3.2. The target-decoy approach over-estimates the q-value/PEP for the six-frame database search.

In order to further study the most appropriate database and error model for the six-frame search results, we investigated an alternative measure of truth for PSMs passing a given statistical threshold. We reasoned that for all EST contigs with significant BLASTX hits to the Chicken Ensembl56 set of proteins, only one of the six translated frames is “correct” based on the top-scoring BLASTX hit to Ensembl. Corresponding PSMs to this “correct” frame were also classified as “correct”, all others as “incorrect”.

We therefore expect that close to 100% of all PSMs are “correct” at the lowest q-values/PEPs and, correspondingly, target PSMs with very high q-values/PEPs should be false. Accordingly, PSMs with q-values and PEPs close to zero should be in the “correct” reading frame almost 100% of the time, whereas manifestly incorrect PSMs should be in the “correct” reading frame approximately 17% (one sixth) of the time. Figure 3.4 shows that the proportion of correct reading frame PSMs is in fact close to 92.5% at the lowest q-values/PEPs. This small discrepancy can be explained by the fact that the reading frame with the most significant BLASTX hit does not always correspond with (and contain) all the true protein sequence. This will be mostly due to sequencing and mis-assembly errors from the ESTs generating “frameshift” errors that push the true coding sequence into multiple frames. Therefore, we likely under-estimate the proportion of correct PSMs by approximately 7% at the lowest (most significant) q-value/PEP thresholds due to these instances. This represents the maximum error for the estimated percentage of true positive PSMs. Figure 3.4a shows the percentage of correct PSMs from the total number of PSMs accepted for each q-value calculated using both the FDR\textsubscript{Käll} and FDR\textsubscript{EG} q-values, as well as from separate database searches using Qvality. The dashed line represents the percentage of PSMs falling in the “correct” reading frame that would be expected at a given q-value threshold, presuming the error modelling is accurate (see Appendix 6 for derivation). For example, approximately 58.3% of PSMs should be in the “correct” frame at a q-value of 0.5, which falls slightly further still to 54% when we factor in the 7% correction described above (Figure 3.4a). However, the observed percentages are significantly higher than would be expected by chance, at close to 66%, 73% and 75% of PSMs assigned to the correct reading frame for the FDR\textsubscript{Käll}, Qvality and FDR\textsubscript{EG} based q-values respectively. This highlights the overly conservative nature of the q-value estimates for the six-frame
database search. Interestingly, the FDR$_{K\text{äll}}$ q-value yields percentages that are closest to expectation suggesting that for this case, at least, it is a more accurate FDR calculation.

Figure 3.4. Estimating the proportion of true positive PSMs identified in the six-frame database search. PSMs were considered to be ‘correct’ if the reading frame contained the top-scoring match to an Ensembl56 protein through a BLASTX search. Plots show: a) The percentage of ‘correct’ reading frame PSMs that fall below each of the three types of q-values and PEP, and b) the same percentage but plotted for local Qvality PEP bins of 0.01.
The proportion of true positive PSMs at different PEPs was also estimated, again assigning “correct” reading frames via BLASTX searches (Figure 3.4b). In this case the percentage “correct” was calculated within 0.01 bins because the PEP, unlike the q-value, is a local measure of significance which can become overly conservative when thresholds are applied to a list of PSMs (Choi and Nesvizhskii, 2008a). The extent of this conservativeness can be seen in Figure 4a, where the percentage of correct PSMs remains above 80% for PEPs above 0.8. The binned PEP data in Figure 4b is noisy compared to the cumulative data in Figure 4a, but nevertheless shows a similar trend to the q-value data. Again, the PEP overestimates the expected proportion of true positive PSMs and both q-value and PEP suffer from poor accuracy when deployed without correction in six-frame database searches.

Nevertheless, the PEP is a highly informative statistic in proteogenomics since a novel gene or splice-variant might only be identified via a single PSM, and we therefore need to know the likelihood of this being a correct match. Hence, it has assisted the high-throughput identification of genes (Borchert et al., 2010) and is an alternative to heuristics such as the somewhat arbitrary two-peptide rule (Gupta and Pevzner, 2009). However, most tools require a decoy database in order to generate a null model, which allows tools such as Qvality to estimate PEPs more accurately (Kall et al., 2009). Likewise, PeptideProphet utilises decoy databases to allow PEP calculations that are free from parametric assumptions to provide accurate PEPs for different search engines and data sets (Choi and Nesvizhskii, 2008b; Shteynberg et al., 2011).

3.3.3. Six-frame databases confound the target-decoy assumption

The source of the atypical statistical estimates generated from the six-frame searches is not immediately obvious. Although differences in target and decoy databases sizes can alter the number of accepted PSMs at a given FDR threshold (Fitzgibbon et al., 2008), thereby breaking the target-decoy assumption, this does not apply here. Our target and decoy database sizes are matched in all searches, including the six-frame searches. Indeed, any difference between the number of unique target and decoy peptides can be
factored into the FDR calculation using a normalisation step (Wang et al., 2009; Elias and Gygi, 2007); this is implemented in the \textit{Qvality} software used here for separate database searches. Regardless of the metric used to estimate significance, the fewest accepted PSMs are always observed in the six-frame database.

The unique feature of the six frame database is that only one in six target sequences is likely to contain the “true” target and most of the target database is inflated by sequences in the ‘wrong frame’. This could imbalance the distributions of target and decoy PSM scores, which we investigate here.

Normally, a probability distribution of PSM scores would have a tail to the right, corresponding to the correct target PSMs (Kall et al., 2008a). Figure 3.5 shows the distributions of target and decoy Mascot scores for the top ranked PSMs identified in the EORF, ESTScan2 and six-frame databases for separate target and decoy database searches. The EORF and ESTScan2 searches (Figures 3.5b and 3.5c) show a clear difference between the target and decoy score distributions, with the expected tail on the right of high-scoring PSMs that are likely to be correct (i.e. a low PEP or FDR). However, this clear distinction is absent for the six-frame database searches in Figure 3.5a, with median scores of 14.98 and 13.42 respectively for target and decoy PSMs. The ESTScan2 median scores for target and decoy PSMs are considerably different and better separated, at 19.33 and 10.77 respectively, with similar values for EORF. For the six-frame searches the increase in the median decoy PSM score stems from the increased number of candidate sequences against which each spectrum can match. This in turn would lead to more decoy PSMs out-competing their target equivalents in a concatenated database search, potentially generating more false negatives for the six-frame search than there are for the other searches. This is indeed the case when the cumulative frequency distributions of the target and decoy PSM Mascot scores are considered for the six-frame searches in comparison to ESTScan2 or EORF (see Appendix 7), where decoy PSMs are assigned higher scores in the six frame database searches for both concatenated and separate search strategies.
Figure 3.5. Mascot ion score distributions for target and decoy PSMs.

Plots show target and decoy PSMs ion score distributions, for all rank 1 PSMs, when target and decoy databases were searched separately. Density plots were generated for: a) standard six-frame database search, b) ESTScan2 search, and c) EORF search.
This inflation of the target database with ‘wrong frame’ sequences has consequences for FDR estimates. We illustrate this in Table 3.2 where we consider a search against a hypothetical standard database that produces 1000 true positive target PSMs and 10 decoy PSMs. We then consider an inflated database that contains the same target sequences plus extra sequences (i.e. alternatively translated ‘wrong’ frames) that are unlikely to be matched as true positives. We presume the same 1000 true PSMs will be returned from a search against this database, but false positive matches (as estimated from decoy PSMs) will likely increase (say to 20), particularly at modest search engine scores. This is because the decoy database is increased whilst the effective target database (i.e. “hit-able” sequences) remains the same. The resulting FDR estimate is then increased, leading to fewer accepted PSMs at the same fixed cut-off.

<table>
<thead>
<tr>
<th>Standard database</th>
<th>Target PSMs</th>
<th>Decoy PSMs</th>
<th>FDR at fixed score threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000 TPs</td>
<td>10 FPs</td>
<td>= 10/1000 = 0.01</td>
</tr>
<tr>
<td>Inflated database (e.g. six-frame)</td>
<td>1000 TPs</td>
<td>20 FPs</td>
<td>= 20/1000 = 0.02</td>
</tr>
</tbody>
</table>

This hypothesis is consistent with the data shown in Figure 3.5a where additional higher scoring decoy PSMs are identified owing to the inflated nature of the six-frame database. Conversely, the presence of the additional five ‘wrong frame’ target sequences increases the chances of assigning spectra to a false positive target, thereby lowering the median target score. Ultimately, this means that the PEP would be over-estimated because it depends on the relative heights of the target and decoy distributions. Likewise, the six-frame FDR will be over-estimated because the decoy PSMs, with relatively higher scores, would be encountered sooner when the sorted list of scores is traversed from high to low during the FDR calculation.
This is illustrated when comparing the ratio of target:decoy PSMs in a concatenated search at a fixed Mascot score between six-frame and ESTScan2 searches (Figure 3.6). Both distributions converge towards the expected 0.50 for PSMs with low Mascot scores, although a small bias towards target PSMs remains for the six-frame database even at scores below 5. However, generally, the six-frame ratio is lower, particularly at ion scores above 15. This leads to increased numbers of decoy PSMs, which in turn leads to an increased estimate in false target PSMs and an increased FDR estimate. The same effect is observed when considering the absolute numbers of target and decoy PSMs in the six-frame database compared to a subset of it composed of just the three forward frames (Figure 3.7A). Although this database is half the size, it contains most of the true target PSMs (since most correct translations are in one of the forward frames). The numbers of reported target PSMs are almost identical between the two searches at all Mascot scores (with an average ratio of 1.05:1); whilst there are 1.4× more decoy PSMs on average in the six-frame search results. This shows that the “true” target PSM set is essentially invariant in the two searches and that few additional target PSMs are accepted in the six frame searches (despite the doubling in database size) while many more decoy PSMs are accepted. This is confirmed in Figure 3.7B where the ratios of accepted 6frame:3frame PSMs are shown for target and decoy PSM sets. The target PSM ratio (in red) stays close to unity whilst the decoy ratio grows steadily from 1 to an average of 1.4. This is a major contributor to the increased FDR estimates observed for the six-frame searches.

The observations detailed above support the hypothesis outlined in Table 3.2, that the inflated nature of the six-frame translation database is the underlying reason for the overestimation of the FDR for the six-frame searches. However, could this simply be a consequence of increased database size? Since database size can affect FDR and PEP estimates, it has been suggested that larger database sizes can reduce the variance in estimates of the number of decoys (false) PSMs, leading to better (more precise) estimates of significance. However, we note that in general this leads to more conservative estimates (Figure 3.8a), consistent with previous observations (Fitzgibbon et al., 2008). To confirm this, we recalculated the q-value for PSMs drawn from a subset of the six-frame database of identical size to ESTScan2, repeating this 1000 times to generate an average FDR profile, shown in Figure 3.8b. This demonstrates how the atypical database composition must also be influencing the FDR calculations, since the six-frame subset q-value profile is
more conservative than the ESTScan2 search despite the fact the database sizes are identical. The overestimated FDR for the six-frame searches is therefore due both to database size and atypical composition. This highlights the need to exercise caution when generating/modifying proteogenomic databases in order to provide more accurate q-value/PEP estimates.

Figure 3.6. Percentage of target PSMs of all PSMs reported at a given Mascot score, comparing 6-frame vs. ESTScan2 searches.

The 6-frame searches have a smaller bias towards the target PSMs at a given fixed Mascot score, which in turn leads to a larger number of decoy PSMs that are observed at a given Mascot score. This value is then used in turn to estimate a correspondingly larger number (and fraction) of the target PSMs to be estimated as “false”. This produces a larger FDR estimate (and in turn q-value) and hence fewer PSMs are ultimately accepted.
Figure 3.7. The six-frame database leads to a greater number of decoy hits, while the target hits remains almost unchanged.

A) Cumulative frequency plots of the total target and decoy PSMs from six-frame and three-frame database searches accepted at fixed Mascot scores. This illustrates the difference between results from databases of differing size (six-frame is twice the size of three-frame). The number of accepted target PSMs at a given fixed Mascot threshold is broadly identical, whereas the number of accepted decoy PSMs remains noticeably higher at all Mascot scores for the six-frame database search compared with the three-frame search. B) Ratios of accepted 6frame:3frame PSMs at different Mascot scores.
Figure 3.8. Effect of database size on FDR of the six-frame PSMs.

a) Larger databases generally report fewer accepted unique peptides at the same PEP threshold. b) Subsets of sizes equal to the ESTScan2 database were randomly sampled (1000 times) from six-frame database. The mean q-values were calculated from the samples to give an FDR profile. This size-normalised six-frame database resulted in FDRs greater than the ESTScan2, but lower than the six-frame PSMs, showing that database composition is also responsible for the inflated FDR of the six-frame search.
3.3.4. Equalising the target and decoy databases improves the sensitivity for the six-frame searches.

In order to estimate accurate FDRs/PEPs the ratio of target to decoy sequences needs to be as close to 1:1 as possible, or at least properly understood and quantified. We reasoned that reducing the six-frame database to a single target and decoy sequence for each EST contig would provide a more accurate FDR estimate. A naïve but simple way to achieve this is by randomly selecting only one frame from each of the six target reading frames for each EST contig to generate a random-frame database. A second approach is to select the single frame containing the most PSMs to generate the top-hit-frame database. In both cases, a single target sequence is selected from the EST contigs along with its reversed (decoy) sequence. The associated PEPs were then recalculated for these modified databases and compared to the equivalent values from the standard six-frame searches, shown in Figure 3.9 and Table 3.3. As expected, this has a marked effect on PEP estimates for equivalent PSMs between the paired searches. Both methods uniformly lower the PEP compared to equivalent six-frame PSMs leading to additional unique peptide matches. Indeed, although the random-frame database has 83% of the original sequences removed, the number of peptides matched is 657 in excess of expectation at a PEP cutoff of 0.01, presuming that only one-sixth of the original PSMs should remain. This is an interesting finding as it shows that choosing a single frame, effectively at random for many EST contigs, improves sensitivity even though some “correct” frames will have been removed by chance. This would not, of course, be a viable strategy in practice but serves to illustrate the point that less conservative statistical estimates can be achieved by filtering the database.
Figure 3.9. Comparison of equivalent PEPs from standard six-frame searches against alternate database searches. PEPs derived from several search strategies are plotted against the six-frame equivalents, with the same sequence-spectra-Mascot score. In a), PEPs derived from simple filtering approaches based on selection of a single frame by: random (random-frame), the most PSMs (top-hit PSM), or the three forward frames, are plotted against the six-frame PEP values. In b), PEPs derived from searches against the six-frame-predicted, ESTScan2 and EORF databases are plotted against the six-frame equivalents. In both plots, direct equivalence of PEP values against the standard six-frame database searches is shown as a dashed line. In all cases, selection of single frames, three forward frames, frame prediction and/or translation by EORF or ESTScan2 reduces the estimated PEP.

Table 3.3. Unique peptide identifications at different Quality PEP cut-offs derived from searches over different databases. Six-frame standard refers to the standard, unfiltered six-frame translation databases. Random-frame to the subset of the six-frame database where a single frame is selected randomly for each contig. Three-frames-forward refers to just the forward frames only. The six-frame-predicted database is produced by retaining only the frame that has the most significant BLASTX hit. Top-hit-frame refers to the subset of the six-frame database where the frame with the most PSMs is selected.

<table>
<thead>
<tr>
<th></th>
<th>PEP cut-off</th>
<th>Number of target sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Six-frame standard</td>
<td>4951</td>
<td>5887</td>
</tr>
<tr>
<td>Random-frame</td>
<td>1483</td>
<td>1739</td>
</tr>
<tr>
<td>Three-frames-forward</td>
<td>5654</td>
<td>7050</td>
</tr>
<tr>
<td>Top-hit-frame</td>
<td>6624</td>
<td>8330</td>
</tr>
<tr>
<td>Six-frame-predicted</td>
<td>5601</td>
<td>7251</td>
</tr>
<tr>
<td>ESTScan2</td>
<td>7325</td>
<td>8907</td>
</tr>
<tr>
<td>EORF</td>
<td>7146</td>
<td>8782</td>
</tr>
</tbody>
</table>
We tested other approaches to improve sensitivity, exploiting *a priori* knowledge on the EST contigs. Given that the direction of sequencing is usually known in transcript libraries, we observe that simply retaining just the forward three frames has an advantage (*three-frames-forward*), producing over a thousand more unique peptides with PEPs of 0.02 or less (Table 3.3). Similarly, considering homology to predict the mostly likely coding frame improves performance. For the EST contigs with BLASTX hits to Ensembl56, the single frame with the most significant E-value was retained, to generate a *six-frame predicted* database. Figure 3.9 shows how this approach significantly reduces the PEPs relative to the standard six-frame searches, as do EORF and ESTScan2. It should be noted though, that although the *six-frame-predicted* database led to the lowest PEP estimates, this approach yields fewer significant PSMs compared to EORF and ESTScan2, shown in Table 3.3. This highlights a downside to the BLASTX-filtering approach in this instance; only about 22% of the EST contigs have significant BLAST hits whereas EORF and ESTScan2 predict protein sequences for the majority of them. The additional peptide matches from EORF and ESTScan2 searches may come from novel genes or isoforms that are not yet annotated in the Ensembl56 known gene set.

Collectively, these results demonstrate a variety of approaches to apply when considering proteogenomic searches against nucleotide databases such as EST/cDNA or genomic six-frame translations. Minimally, selecting one single candidate reading frame from the six possible leads to superior FDR and PEP estimates, and using empirical evidence to select one of the six frames via BLASTX searches against existing protein databases is better still. Nucleotide sequences that have no significant BLASTX hits still need to be dealt with, but simple strategies here can also be applied. For example, even weak BLASTX hits can be used to suggest the most likely frame that contains some level of coding features. Similarly, statistics such as codon usage can suggest “protein-like” features that point to the mostly likely frame. We note here that the “correct frame” sequences share similar amino acid composition statistics with Ensembl proteins (Figure 3.10) and their amino acid frequency distribution is much closer to that of Ensembl proteins than the “incorrect frame” sequences.
Figure 3.10. Amino acid composition of the translated EST contig reading frames. The ‘correct’ reading frame, as determined by BLASTX searches, more closely resembles amino acid composition in Ensembl genes.

<table>
<thead>
<tr>
<th>Correct-frame</th>
<th>Frame 1</th>
<th>Frame 2</th>
<th>Frame 3</th>
<th>Frame 4</th>
<th>Frame 5</th>
<th>Frame 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl</td>
<td>0.010</td>
<td>0.063</td>
<td>0.063</td>
<td>0.064</td>
<td>0.069</td>
<td>0.069</td>
</tr>
<tr>
<td>UniRef90</td>
<td>0.026</td>
<td>0.096</td>
<td>0.096</td>
<td>0.097</td>
<td>0.101</td>
<td>0.101</td>
</tr>
</tbody>
</table>

Table 3.4. Kullback-Leibler divergence shows amino acid distribution for the correct reading frame is similar to annotated proteins. ‘Correct-frame’ refers to the reading frames with best blast hit to Ensembl 56, and frames 1-6 are those which do not have the best blast hit.

Other good practice is evident in the literature, such as the prot4EST pipeline which includes a rule-based approach that identifies the longest ORF from the six-frame translations (Wasmuth and Blaxter, 2004). Similarly, (Adamidi et al.) (2011) selected the
three longest ORFs from RNA-Seq transcripts for database searching. Another approach, which does not require sequence homology, utilises information about the experimental protocol used for cloning and sequencing the mRNAs. One small-scale proteomics study identified 51 novel seminal fluid proteins by searching 2D-LC-MS data against only the forward frames of translated UniGenes (Walters and Harrison, 2010). When applied to shotgun proteomics data this improves sensitivity relative to standard six-frame searches, a trend we also note here. Our EST data is generated from directional cloning and subsequent 5’-end sequencing of chicken cDNAs, leading to a bias where 97% of the “correct” frames from BLASTX searching are in frames one, two, or three (Figure 3.11). Searching against this three-frame database leads to over 1000 additional peptide identifications filtered at a PEP<0.02 (Figure 3.7, Table 3.3). This improved performance is likely the result of the reduction in erroneous targets, which is reduced to two out of three in the 3-frame search from five out of six. Although the performance is inferior to the BLAST-filtered single-frame and ESTScan2/EORF searches, the three-frame approach can be used for all EST contigs, and not just those with BLASTX matches. Moreover, this approach can be easily implemented for newly sequenced genomes when there are no homologs annotated within the clade, something which would create problems for EORF/ESTScan2 as these tools require codon usage information.

Ultimately, however, our most sensitive searches were derived from the EST translation tools, ESTScan2 and EORF, which generate a single translated sequence across the most likely reading frames. The most effective search strategy overall in terms of sensitivity involved the two database searches using the FDRScore algorithm for combining multiple search results (Jones et al., 2009).
Figure 3.11. EST contig directionality.

The majority of significant BLASTX hits are in frames 1-3. EST contigs were matched with Ensembl56 proteins and the number of times each frame formed the high-scoring segment pair was determined.

3.4. Conclusion

This study has highlighted some of the pitfalls when searching against nucleotide databases via six-frame translations and how inflating the target database with incorrect sequences perturbs FDR and PEP estimates. Although some methods exist for reducing the peptide search space in proteogenomic studies (Edwards, 2007; Robinson et al., 2009) most do not formally consider the inherent biases that can arise through the improper use of the target-decoy approach. We show here that naïve six-frame searching leads to over-conservative statistics and potential loss of high-quality peptide evidence for genomic annotation.

We recommend proteogenomic practitioners consider the following redundancy removal guidelines when searching against nucleotide databases such as those generated from cDNAs/ESTs, RNA-seq, gene prediction software or raw genomic sequence.
• **Removal of incorrect reading frames.** Even random removal of five out of six translated frames did not degrade relative search performance in our tests, providing better statistical estimates at the expense of some true PSMs. We would not, however, recommend random removal of reading frames as a practical strategy.

• **Selection of most likely frame based on PSMs.** Retention of the single frame (and its decoy) with most PSMs prior to calculation of FDR/PEP provides less conservative statistics, although multi-pass approaches have to be performed appropriately (Bern and Kil, 2011; Everett et al., 2010; Gupta et al., 2011).

• **Selection of most likely frame based on homology.** BLASTX searches against a protein database, to identify the single frame mostly likely to be coding removes redundancy and improves sensitivity. If close homologues are not available, even weak matches to known proteins should enrich for coding sequence.

• **Selection of frame based on coding potential.** When no homologues are available, similarity in amino acid composition or codon usage can help select the most likely coding frame.

• **Translation software.** Gene prediction (e.g. Augustus) or EST translation software (such as ESTScan or EORF) can overcome introns and frameshifts, as well as remove redundancy. EST translators can be combined with BLASTX data to improve accuracy (Wasmuth and Blaxter, 2004).

Ideally, we suggest that gene prediction or EST translation software are likely to be most effective, since six possible frames are reduced to one single frame, whilst frameshift mutations and introns can be overcome. Moreover, presence of a transcript is *bona fide* evidence that a genomic region is at least transcribed and in most cDNA libraries the sequence is normally largely free of introns and the clonal direction is also known; so only three forward frames need be considered.

One should not lose sight of the *raison d’être* of proteogenomics, to identify novel genes and novel gene structure that may not exist in extant genome annotations or transcriptome data sets. This necessitates searching against translated genome sequence, and in such cases, use of a gene prediction tool such as Augustus (Stanke et al., 2006; Brosch et al., 2011), perhaps used with reduced stringency to capture some atypical genes/gene structure, is preferable to raw searches against six-frames. Alternately, if
searching against raw genomic sequence we suggest selecting one candidate translation (out of six) with the closest homology with Ensembl or UniRef proteins (i.e. similar to the six-frame-predicted database), in addition to applying compositional filters to limit regions unlikely to be coding. This should improve the overall sensitivity of the searches whilst improving the statistical modelling of incorrect PSMs.

Despite its limitations, there are some advantages of using the six-frame translation approach. EORF and ESTScan2 predict a single protein sequence from a transcript because in most cases the mRNA contains encodes a single ORF. However, overlapping reading frames have been observed in eukaryotes. This phenomenon was thought to be almost exclusive to viruses, bacteriophages, mitochondria and plasmids, but now alternative reading frames (ARFs) are thought to be more prevalent in eukaryotes than previously thought and the term ‘ARFome’ has been coined as a result (Chung et al., 2007). The INK4A tumour suppressor gene is one well documented case, which encodes two functionally related proteins involved in cell cycle progression (Quelle et al., 1995). A six-frame translation database contains all possible frames for a given transcript, and therefore multiple reading frames of the same transcript could potentially be identified by PSMs. Indeed, this was observed for some ESTs in this study and further work is needed to investigate whether such cases represent ARFs or false positive PSMs.

Finally, it should be stressed that care needs to be taken when applying target-decoy strategies to shotgun proteomics data and proteogenomics data in particular. As several authors have pointed out, if the underlying database or search engine is not compliant with the assumptions of the target-decoy approach, significance values can be under-estimated (Gupta et al., 2011; Nesvizhskii, 2010; Bern and Kil, 2011; Everett et al., 2010). This is particularly acute for multi-pass search strategies that change the target-decoy structure between search phases. Here, we ensure the numbers of candidate target and decoy sequences in the database are the same, though we do favour target sequences with certain properties (i.e. those likely to be coding, have the most PSMs, etc), which could introduce some biases. Nevertheless, the evidence that standard six-frame searches considerably over-estimate FDRs/PEPs seems overwhelming, which highlights the need for databases to be checked for compliance with the target-decoy approach (in addition to the search engines). It is clear that more work is needed to produce better error models and databases to develop more reliable statistical confidence estimates, which would greatly benefit
proteogenomics by helping to achieve a more comprehensive representation of the proteome.

3.5. References


of the yeast proteome by single-shot UHPLC runs on a bench-top Orbitrap. *Molecular & cellular proteomics, 11*(3),M111.013722.


Chapter 4: Improving Genome Annotation using Shotgun Proteomics and De Novo-Assembled Transcripts

Genome annotation pipelines such as Ensembl do not normally consider high-throughput transcriptome sequencing data, at least directly, when constructing gene models. Therefore, the integration of proteomics and transcriptomics data in a statistically robust manner can provide a more comprehensive annotation of protein-coding gene structure on the genome. Chapter 3 showed that the conceptual translation of transcript sequences using the algorithms EORF and ESTScan2, prior to MS-database searching, enables a greater number of accepted peptide identifications compared with other approaches whilst retaining data quality. Chapter 4 builds on this work by using these translation tools to generate high quality PSMs to improve the annotation of the chicken genome. In this chapter, we assess the utility of our proteogenomics pipeline for discovering and improving gene models in a vertebrate model system on a genome-wide basis. In addition, we highlight the advantages of using de-novo EST assembly by identifying protein-coding transcripts that are missing or only partially present in the existing genome assembly.

The author of this thesis was responsible generating the results and figures and formulating the main ideas. The EORF algorithm and EORF protein database used in for this chapter were produced by Ian Overton. Simon Hubbard directed the work.
Abstract

Proteogenomics aims to bridge the gaps between the different ‘omics’ and to discover missing cellular components required for a systems-wide view of biology, including a comprehensive set of genes and their encoded proteins. Mass spectrometry (MS) based proteomics is a powerful approach to gain evidence for protein-coding genes and splice variants. However, protein and peptide identification is critically dependant on searching the MS data against an appropriate proteome database. Alternative strategies in proteogenomics include searching against genome sequences, computational gene predictions, and transcriptome data. Searching against the in-silico translated transcriptome offers a number of advantages compared with other methods, and we demonstrate the utility of this approach by improving the annotation of the chicken genome. Three different protein sequence databases were generated from 85,486 EST contigs (assembled from 339,314 ESTs), which were interrogated with 403,820 tandem mass spectra. This allowed the confirmation of 8,932 tryptic peptides present in Ensembl protein sequences, and the discovery of 1,383 novel peptides unique to the ESTs. The novel peptides are shown to correspond to novel genes, splice variants, extensions to existing Ensembl gene models, and polymorphisms. In addition, 826 EST contigs with only low-quality alignments to the genome are proposed to contain genes that are absent from the genome assembly. Selected examples of novel peptides are considered in detail, including those which map to the hitherto unannotated genes: Aldolase C and LMAN2. The EST and proteomics data also led to the detection of both the transcription and translation of the ARF6 pseudogene and reassignment as a protein-coding gene. Integration of proteomics and EST data can contribute substantially to genome annotation, even for organisms that are at an advanced stage of the annotation process.
4.1. Introduction

In the context of genome annotation, it can be argued that determining the correct gene structure is as challenging as finding the genes themselves. Although computational predictions allied to homology and transcript data have helped discover the majority of human protein-coding genes, the accurate prediction of complete exon-intron structures has proven to be more problematic. For instance, it has been estimated that around 50% of gene annotations have at least some incorrect genomic structure including gene boundary, reading frame and exon-intron boundary errors (Guigo et al., 2006). Indeed, the stochastic nature of the RNA transcriptional and splicing machinery highlights how difficult it is to accurately predict comprehensive gene models from a genome sequence alone (Pan et al., 2008). In particular, predicting the combinations of splice sites that are used in the pre-mRNA for a given tissue has proven very difficult. Therefore, identification of the gene structure requires experimental data to complement de novo genome annotation. Advances in transcriptome sequencing and mass spectrometry-based proteomics technologies are being used for this purpose (Pan et al., 2008; Merrihew et al., 2008). Indeed, several recent studies have integrated different forms of experimental evidence into their annotation pipelines and report improvements in sensitivity and selectivity (Robinson et al., 2009; Allen et al., 2004).

As primary transcripts, the high-throughput sequencing of cDNAs and/or Expressed Sequence Tags (ESTs)\(^3\) remains a powerful method for identifying genes in both sequenced and unsequenced genomes. The dbEST database currently contains more than 73 million entries (Boguski et al., 1993) and since the advent of next-generation sequencing transcriptome resources such as this, which now include the results from RNA-seq experiments, are set to grow rapidly. Although ESTs are generally short (\textit{circa} 500 bases) and relatively error prone sequences, assembly into contigs reduces ambiguous genome mappings, mitigates sequencing errors and can improve coverage of splice variants. This is done either by first aligning the ESTs to the genome to guide the assembly process, and/or by de novo assembly into contigs without a reference genome. The latter approach has the advantage of generating EST contigs when the genome sequence is

\(^3\) ESTs are referred to here, though similar conceptual challenges are presented with RNA-seq data, namely short sequences and some relatively high error rates, both of which can be partly mitigated by the depth and coverage now achievable with the latest technology.
incomplete. The EST contigs can then be aligned to the genome sequence using programs such as BLAT (Kent, 2002) or EXONERATE (Slater and Birney, 2005), which can reveal new splicing patterns and other post-transcriptional processing events (Edwards, 2007). In addition, cDNA libraries derived from different individuals can be compared to identify polymorphisms (SNPs, INDELS) (Garg et al., 1999; Wang et al., 2010).

Mass spectrometry-based proteomics has emerged as a promising resource for genome annotation and complements transcriptome sequencing efforts in the quest to comprehensively study the full repertoire of cellular components (reviewed in (Armengaud, 2010)). Several proteogenomics studies have used shotgun proteomics as a high-throughput means of annotating protein-coding genes (Merrihew et al., 2008; Kalume et al., 2005; Tanner et al., 2007; Gupta et al., 2008; Sevinsky et al., 2008; Savidor et al., 2006). These methods are advantageous for functional annotation because peptide identifications are usually derived from the mature protein and are therefore directly relevant to the post-translational level (which cannot be observed by transcriptomics). Proteogenomics can confirm and improve upon existing gene annotations (Kuster et al., 2001; Schrimpf et al., 2009; Tress et al., 2008) as well as providing direct evidence for novel protein-coding genes (Merrihew et al., 2008; De Groot et al., 2009; Findlay et al., 2009), thus verifying that both transcription and translation occur at specific loci. Indeed, since ncRNA and pseudogenes can resemble protein coding genes in terms of their sequence and chemical modifications (Oshiro et al., 2002; Kondo et al., 2010) they can be overlooked by annotators. Furthermore, many small ORFs may be translated and functional, as has been suggested in yeast (Oshiro et al., 2002) and Drosophila where small bioactive peptides can function in embryogenesis (Kondo et al., 2010).

The choice of database for searching with mass spectra is a crucial factor in proteogenomics, as demonstrated in Chapter 3. The two most common methods search spectra against either a raw/six-frame translated genome or against existing annotations such as Ensembl (Flicek et al., 2010). However, whole genome-based searches generally suffer from poor sensitivity (especially in complex vertebrate genomes) where huge databases will be generated that contain relatively low percentages of true protein-coding sequence. This in turn leads to poor sensitivity from statistical estimates as shown in Chapter 3, or worse, poor precision (specificity) if statistical approaches are not applied and high false positive rates will be observed. In addition to these challenges, searching
raw genomes leads to an inability to identify intron-spanning peptides, whereas searches based on existing annotations are restricted to known genes. Although the specificity for peptide identifications can be increased, such as by algorithms which utilise peptide isoelectric focusing (Sevinsky et al., 2008), ultimately it is the quality of the sequence database which allows accurate protein identifications. A recent proteogenomics study achieved success in improving the annotation of the Arabidopsis genome by searching spectra against gene predictions made by AUGUSTUS (Castellana et al., 2008; Stanke et al., 2006), thus restricting the search space to candidate genes.

A more empirical approach is to search spectra against transcriptome data such as EST sequence data, and this approach has several advantages. ESTs are prima facie evidence of gene expression, most usually from protein coding regions. ESTs are also typically mature, spliced transcripts, which span intronic regions. Equally, with the advent of modern sequencing technology via RNA-seq they offer very good coverage of the expressed genes in a cell or tissue. Indeed, ESTs offer high sensitivity for MS/MS database searching owing to the enrichment of true protein-coding sequences and can be used when the genome sequence is unavailable, incomplete or unassembled. ESTs have been used in a variety of different ways in proteogenomic studies, including lending support for candidate gene annotations (c.f. (Borchert et al., 2010; Edwards, 2007; Wright et al., 2009)) and discovery of novel protein sequences (Robinson et al., 2009).

While some search engines can match spectra directly to nucleotide sequence (e.g. Mascot (Perkins et al., 1999), which translates nucleotide sequences “on-the-fly”) it is usual to perform an in silico translation prior to the database search to give flexibility when generating decoy databases. EST databases can be very large and contain errors and non-coding sequence. For example, frameshift errors and premature stop codons will confound proteomic searches and non-coding sequence such as 3’UTRs are overrepresented in ESTs. Therefore, the recognition of coding from non-coding sequences is crucially important. One solution is to conservatively filter EST sequences to reduce redundancy and produce smaller and more reliable databases, for example by requiring matches to annotated genes or by clustering ESTs into contigs (c.f. (Boardman et al., 2002)). An alternative (or additional), less constrained, strategy involves conceptual translation of ESTs (or EST contigs) using protein prediction tools such as ESTScan2 (Iseli et al., 1999) and OrfPredictor (Min et al., 2005). These algorithms limit the size of EST databases by
predicting the most likely protein sequence from the six possible reading frames, correcting frame-shift errors and excluding predicted UTRs. For example, the ESTScan2 program was instrumental in a recent study which took advantage of the EST2Secretome pipeline to generate a database of putative secretory protein sequences (Robinson et al., 2009).

In the present study, assembled EST contigs are translated into amino acid sequence prior to MS database searching. Once identified by PSMs, the EST contigs are then mapped to the genome. The efficacy of this approach is demonstrated on a genome-wide scale in a vertebrate system, chicken (*Gallus gallus*). Despite being sequenced in 2004 (Hillier et al., 2004), the annotation of the chicken genome lags behind many other mammalian genomes due to the lack of closely related genomes and difficulties encountered during the assembly of microchromosomes (Douaud et al., 2008). Sequencing of the chicken transcriptome (Hubbard et al., 2005; Wang et al., 2011), and the genome sequences of other birds (Dalloul et al., 2010; Warren et al., 2010) has improved matters, but no comprehensive proteogenomics data has yet been generated for chicken.

To address this, a total of 403,820 MS/MS spectra were used to identify proteins predicted by EORF, ESTScan2 (Iseli et al., 1999) and six-frame translations of chicken EST contigs, and the ability of each translation method to enable proteogenomic discoveries was tested. The proteogenomics pipeline was able to validate existing genes, identify polymorphic variations and uncover candidate novel chicken genes, splice-variants and even genes missing from the genome assembly. This is illustrated with examples to show how using EST’s, refined by proteomic data, can improve the genome annotation of a model organism.

### 4.2. Materials and Methods

The EST dataset used in this chapter was described in section 3.2.1, and details of the conceptual translation methods were outlined in sections 3.2.3 and 3.2.4. Mass spectrometry analysis of the chicken proteome was described previously in sections 2.2.3 and 3.2.2.
4.2.1. EST dataset

A total of 339,314 ESTs were sequenced from 64 cDNA libraries derived from 21 chicken tissues (Boardman et al., 2002). Clustering and assembly of these using BLASTN (Altschul et al., 1990) and PHRAP (Gordon et al., 1998) was used to generate 85,486 EST contigs. Two different sets of protein sequences were predicted from each EST contig using the EORF and ESTScan2 (Iseli et al., 1999) algorithms. EORF was described in Chapter 3. ESTScan2 uses a hidden Markov model to detect and reconstruct coding sequences from ESTs. In addition, a custom Perl script was used to generate six-frame translations of each EST contig nucleotide sequence.

4.2.2. Conceptual translations

The standard method for creating a protein sequence database from a set of nucleotide sequences is to perform a six-frame translation. A custom Perl script was used to translate each EST contig in all six possible reading frames (3 forward, 3 reverse) using the standard genetic code. The translated EST contig sequence was split into its constituent ORFs, generating a set of potential ORFs for each EST contig. This resulted in a large number of short sequences (i.e. less than 100 amino acids in length). Although it is unlikely that these represent full length proteins, they were retained in case they represent partial sequences or bioactive peptides.

4.2.3. Preparation of chicken samples and mass spectrometry

The MS/MS data were derived from a proteomic analysis of the DT40 chicken cell line which used the LOPIT protocol (Hall et al., 2009), kindly provided by collaborators (Kathryn Lilley, University of Cambridge, personal communication). DT40 cells were fractionated by density gradient centrifugation and 7 fractions were chosen for analysis. The selected fractions were labelled with four-plex iTRAQ reagents and digested with trypsin. Labelled peptides were pooled together and separated using two-dimensional
liquid chromatography. LC-MS/MS was performed using an ultimate-nano-LC system (Dionex) coupled to a QSTAR XL mass spectrometer (Applied Biosystems).

4.2.4. Databases

All chicken chromosomes (May 2006 assembly) and Ensembl protein sequences (version 56) were downloaded in fasta format from the Ensembl FTP server (ftp://ftp.Ensembl.org). The ‘pep.all’ set containing translations from known and novel genes was used. Each chicken chromosome contained masked low complexity regions. The chicken IPI database (version 3.28) was downloaded from (ftp://ftp.ebi.ac.uk). The UniRef90 database (release 15.11) was downloaded from (http://www.uniprot.org/downloads).

4.2.5. Mass spectrometry database searching

A total of 403,820 spectra from the iTRAQ experiments on DT40 cells, were searched against five different protein sequence databases using Mascot version 2.0. A peptide error tolerance of 2 Da, and MSMS error tolerance was 0.8 Da. Modifications were fixed as cysteine beta-methylthiolation, and iTRAQ labelling of lysine residues and N-termini. Variable modifications were iTRAQ labelled tyrosine and methionine oxidation. Up to 1 missed cleavage was allowed.

Five different databases were searched: chicken Ensembl protein sequences, chicken IPI protein sequences, proteins predicted from EST contigs by EORF (described above), proteins predicted from EST contigs by ESTScan2, and proteins derived from six-frame translations of the EST contigs. A decoy database was concatenated to the end of each of the databases which contained a reverse of each sequence in the original database. The false discovery rate (FDR) for each search was calculated using the notation by Elias and Gygi (Elias and Gygi, 2007), and converted into a q-value. Peptides were retained if they had a q-value lower than 0.01.
4.2.6. Coverage of proteins predicted from ESTs

The proteome coverage of the two predictive methods, EORF and ESTScan2, and of the six-frame nucleotide translations were analysed through a BLAST-based comparison with the Ensembl v56 and UniRef90 chicken protein sequence databases. BLASTX searches were performed for each EST contig nucleotide sequence. The best BLAST match for each EST contig was used and counted only if the following criteria were met; 95% sequence identity, alignment length greater than 50 residues and expectation value better than 0.001. The genome coverage for each of the three translation methods was analysed further by searching the EST contigs against the chicken genome using the EXONERATE pairwise sequence alignment algorithm (Slater and Birney, 2005). The est2genome model was used and only top scoring matches were retained for each EST contig.

4.2.7. Identification of mismatches in peptide sequences

The EST contigs identified by Mascot were searched against Ensembl proteins sequences using BLASTX. Each identified peptide sequence was aligned, using BLASTP, to the corresponding Ensembl protein sequence that matched the parent EST contig via the BLASTX search. The peptides were categorised by the number of amino acid mismatches that appeared in the BLASTP alignment. For peptides matching multiple proteins, only the best alignment was retained. The PAM30 substitution matrix was used to enable the detection of short sequence matches. The analysis was performed using the three different conceptually translated EST databases described earlier.

4.2.8. Validation of high-confidence SNPs

The BLASTP analysis used to identify mismatches in peptide sequences resulted in a number of peptides which contained a single amino acid mismatch with respect to the genome (i.e. a SAAP). A high-confidence set of 733 SNPs contained in the EST contigs was downloaded from the BBSRC ChickEST ftp site.
A custom perl script was used to extract each SNP together with fifteen nucleotides flanking both sides of the SNP. Each peptide sequence containing a SAAP was aligned (using tBLASTn) with the EST contig sequences of the high-confidence set of SNPs. Peptide matches were retained if the parent EST contig was matched with 100% sequence identity and query coverage, and the aligned region covered the 31-mer SNP sequence.

4.2.9. Finding novel protein-coding regions

Two methods were used to identify peptides and EST contigs that represent novel genes: direct and indirect. In the direct approach, each EST contig was mapped to the chicken genome using EXONERATE (Slater and Birney, 2005). A custom Perl script was used to extract the exon coordinates of the EST contigs in GTF format based on the EXONERATE alignment output. Peptide sequences identified from the ESTscan2 / EORF translations were then assigned coordinates based on their position within the EST contig-genome alignment. This procedure also allowed for the mapping of intron-spanning peptides. To identify those peptides and EST contigs which may represent novel genes or transcripts, we compared the GTF coordinates of the peptides and EST contigs with those of Ensembl v56 genes. Peptides were matched to Ensembl genes if they were on the same strand and their coordinates overlapped. EST contigs with introns greater than 18kb in length were not considered here because they exceed the maximum chicken intron length, and therefore most likely represent artefacts (e.g. chimeric EST contigs). This allowed the identification of peptides and EST contigs that map to genomic loci where there are no annotated genes present.

The indirect approach involved using BLASTX to search the EST contigs against both UniRef90 entries and the chicken Ensembl protein sequences to give additional evidence for the authenticity of the novel genes discovered by the EXONERATE mappings. Firstly, EST contigs with significant BLASTX hits (percentage identity > 80%; coverage >=50) to Ensembl proteins were ignored because these most likely represented genes that had already been annotated. These parameters may represent a rather conservative cut-off but were designed to stringently reduce false positives at the expense
of missed cases. Secondly, BLASTX was used to search the EST contigs against UniRef90. Matches with an expectation-value lower than 1e-5 were retained, providing evidence of protein-coding potential via homology to another known gene. EST contigs with no Ensembl match, but which match to a UniRef90 sequence, were considered to represent novel genes models or splice variants. The indirect mapping approach is independent from the genome assembly and allows the identification of genes that are absent from the genome assembly.

4.2.10. Finding partial and missing genes

The set of PSM-identified EST contigs that were mapped with an EXONERATE score of less than 2000 represent low-quality alignments that contain chimeric EST contigs, partial genes, and genes absent from the genome assembly. Partial genes were identified by EST contigs having a significant BLASTX hit to UniRef90, and a near perfect alignment block of more than 150 bp, followed by a misaligned region that was not present anywhere in the genome sequence. To rule out the presence of a chimeric EST contig from the fusion of two unrelated ESTs, the misaligned region was searched against the genome using BLAT (Kent, 2002) to ensure that the EST contig was not derived from two separate genes. Alignments with a score of lower than 1000 and a significant UniRef90 hit, were considered to be protein-coding genes missing from genome assembly.

4.2.11. Functional annotation of the novel LMAN2 gene

The ESTScan2 prediction of chicken LMAN2 protein was searched against the NCBI non-redundant database using BLASTP. A total of 17 LMAN2 orthologs were identified. Protein sequences of the 17 orthologs were extracted from the blast output (E-value < 1e-50). ClustalW2 (Larkin et al., 2007) was used to perform a multiple sequence alignment of the putative chicken LMAN2 protein with its orthologs. Functional regions of the LMAN2 orthologs were extracted from Genbank flat-files and mapped onto the predicted chicken LMAN2 protein sequence.
4.3. Results and Discussion

4.3.1. EST translation approaches for genome annotation

Searches with BLAST against the Ensembl chicken proteome enabled estimation of
gene coverage for 85,486 EST contigs (Boardman et al., 2002), as well as the EORF and
ESTScan2 conceptual translations. The EST contigs matched to 52% of the Ensembl
annotated chicken proteome providing coverage for 64% of protein-coding genes
(BLASTX, ≤95% sequence identity; expect<0.001; alignment length ≥50 amino acids).
Table 4.1 shows that a substantial proportion of the BLASTX-matched Ensembl proteins
were identified by EORF (96%) and ESTScan2 (83%) (BLASTP, thresholds as above).
This demonstrates the fidelity of the EST translation tools as relatively few proteins were
“lost” owing to incorrect conceptual translation or filtering due to perceived lack of coding
potential. Additionally, EORF had greater sensitivity than ESTScan2, matching a
significantly (based on Poisson confidence intervals) larger number of Ensembl proteins
(p<10^{-11}). As expected, BLASTX searches of EST contigs against the Ensembl proteome
yielded significantly more matches than separate BLASTP searches of EORF and
ESTScan2 translations (respectively p<0.05, p<10^{-16}). Interestingly, combining the output
of EORF and ESTScan2 matches significantly more Ensembl genes via BLASTP than
BLASTX does from the untranslated contigs (p<0.05). This suggests that in tandem, the
two translation tools EORF/ESTScan2 are able to overcome EST sequencing errors (e.g.
frameshifts), which adversely affect the BLASTX alignment statistics and lead to poor
quality matches. This is perhaps not totally surprising since BLAST itself has no intrinsic
model of frameshifts, introns or splicing but that translation tools can potentially overcome
these issues. Furthermore, Table 4.1 shows that EORF and ESTScan2 combined match to
2376 more Ensembl proteins (2300 more genes) than ESTScan2 alone, underlining the
advantage of integrating both algorithms for proteogenomic protein identification
purposes.
Table 4.1. BLAST comparisons of EST contig data sets to Ensembl genes. EORF predicts protein sequences from a greater number of EST contigs. Predicted proteins from EORF and ESTScan2 were combined to form the ‘union’ database.

<table>
<thead>
<tr>
<th>EST contig dataset</th>
<th>Search algorithm</th>
<th>Total EST contigs</th>
<th>EST contigs with blast hit</th>
<th>Ensembl proteins</th>
<th>Ensembl genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembled EST contigs</td>
<td>BLASTX</td>
<td>85486</td>
<td>19082</td>
<td>11562</td>
<td>10690</td>
</tr>
<tr>
<td>EORF predicted proteins</td>
<td>BLASTP</td>
<td>67137</td>
<td>17391</td>
<td>11067</td>
<td>10246</td>
</tr>
<tr>
<td>ESTScan2 predicted proteins</td>
<td>BLASTP</td>
<td>62161</td>
<td>14628</td>
<td>9631</td>
<td>8949</td>
</tr>
<tr>
<td>EORF + ESTScan2 proteins (union)</td>
<td>BLASTP</td>
<td>75051</td>
<td>19285</td>
<td>12007</td>
<td>10734</td>
</tr>
<tr>
<td>EORF + ESTScan2 proteins (overlap)</td>
<td>BLASTP</td>
<td>54247</td>
<td>12910</td>
<td>8691</td>
<td>8461</td>
</tr>
<tr>
<td>Unique EORF protein matches</td>
<td>BLASTP</td>
<td>12890</td>
<td>4590</td>
<td>2376</td>
<td>1785</td>
</tr>
<tr>
<td>Unique ESTScan2 protein matches</td>
<td>BLASTP</td>
<td>7914</td>
<td>1785</td>
<td>940</td>
<td>488</td>
</tr>
</tbody>
</table>

4.3.2. Identification rates using the different databases

The proteogenomics pipeline used to annotate the chicken genome is outlined in Figure 4.1. This pipeline was applied to the chicken genome as a case study, but is generic and suitable for analysing any genome where ESTs are also available. In total, five different sequence databases were searched using Mascot (Perkins et al., 1999). The MS/MS spectra were matched against the IPI (Kersey et al., 2004) and Ensembl (Flicek et al., 2010) protein databases to estimate coverage of two extant proteome annotations.
Spectra were also searched against three sets of protein predictions generated by EORF, ESTScan2 and six-frame translation of the EST contigs.

![Figure 4.1. Proteogenomics workflow summary.](image)

The EST contigs were used as a platform for proteomics identifications. Three databases derived from the conceptual translation of the EST contigs were compared for their ability to enable peptide identifications and improve genome annotation. Two different routes were taken to match identified peptides with the genome: direct and indirect. The direct route mapped peptides based on the EXONERATE alignment of their parent EST contigs whereas the indirect route is based on BLASTX searches against Ensembl and UniRef90 protein sequences.

The search databases produced by EORF and ESTScan2 both outperformed six-frame translations of the EST contigs in terms of generating high quality peptide spectrum matches (PSMs) from the MS/MS data, equating to a greater number of candidate protein and gene identifications. As discussed in chapter 3, the greater number of peptides identified after searching the EORF and ESTScan2 databases is due to at least two factors:
the correction of frameshift errors by EORF/ESTScan2, and the statistical implications of searching spectra against a much larger six-frame translation database protein sequence database. Figure 4.2 shows an example of an ESTScan2 PSM that is not present in the six-frame translations. This is likely due to a sequencing error which has resulted in a frameshift in the region of the EST containing the ESTScan2 PSM. Chapter 3 suggests the largest gain is from the better FDR modelling estimates achieved from ESTScan2/EORF searches compared to six-frame searches, but clearly the translation tools are also able to correct some sequencing errors too. Overall, the EORF and ESTScan2 translation algorithms enable the identification of significantly more PSMs from an EST database (respectively $p<10^{-6}$, $p<10^{-8}$) when compared with six-frame translation at an equivalent FDR.

Figure 4.2. EXONERATE alignment of the signal recognition particle (54kDa) protein sequence predicted by ESTScan2.

The PSM highlighted in blue was matched exclusively to the ESTScan2 sequence. The ESTScan2 sequence (top) contains an INDEL with respect to the EST nucleotide sequence, creating a frameshift, which prevents an equivalent PSM in the six-frame translation database.
4.3.3. Mapping ESTs and their PSMs to the chicken genome

Searching against the EORF and ESTScan2 databases allowed the confirmation of existing Ensembl gene models. Two mapping approaches were employed to match PSMs to the genome: a direct approach involving the alignment of PSM-identified EST contigs to the genome sequence; and an indirect approach using BLASTX to match PSM-identified EST contigs to the Ensembl protein sequences. The direct approach would be expected to give the most accurate peptide-to-gene matches, whereas the indirect approach is useful when there are missing genes due to gaps and errors present in the genome sequence. Table 4.2 shows the number of Ensembl genes and proteins verified with concurrent EST and PSMs \((q<0.01)\) for the indirect mapping approach. The ESTScan2 and EORF databases enabled the confirmation of over 2000 Ensembl genes using stringent EST-mapping criteria and high confidence PSMs.

**Table 4.2.** Gene models confirmed using BLASTX to map PSM-confirmed-EST contigs to Ensembl protein sequences. PSMs were obtained from three different database searches using \(q\)-value cut-off of 0.01.

<table>
<thead>
<tr>
<th>Database</th>
<th>Total peptides</th>
<th>EST contigs</th>
<th>Ensembl proteins</th>
<th>Ensembl genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESTScan2</td>
<td>10513</td>
<td>5474</td>
<td>2355</td>
<td>2161</td>
</tr>
<tr>
<td>EORF</td>
<td>9606</td>
<td>5265</td>
<td>2285</td>
<td>2105</td>
</tr>
<tr>
<td>Six-frame</td>
<td>6637</td>
<td>4534</td>
<td>1907</td>
<td>1756</td>
</tr>
</tbody>
</table>

The direct mapping approach involved the alignment of PSM-identified EST contigs to the genome sequence using EXONERATE, to obtain exact peptide and EST coordinates. The EST contigs were mapped to gene-dense regions of the genome, and this is illustrated using chromosome 1 as an example (Figure 4.3). The genome sequence was split into 150 bins and the Ensembl protein-coding genes and EST contigs were counted for each bin. The Pearsons’ correlation between the number of EST contigs and Ensembl genes in each binned region was calculated. This shows that the number of mapped EST
contigs across the genomic loci broadly corresponds with that of the Ensembl genes. However, the correlation between Ensembl genes and EST contigs is stronger ($r=0.53$) for the set of EST contigs identified via EORF and ESTScan2 PSMs, than for the six-frame translation PSMs. This follows the trend observed in Table 4.2, and confirms that the PSMs identified via the EORF and ESTScan2 translations contain a greater proportion of true positive identifications corresponding to a greater number of identified genes.

![Figure 4.3](image)

**Figure 4.3. Distribution of PSM-identified EST contigs compared with known Ensembl genes located on chicken chromosome 1.**

The genome sequence of chromosome 1 was split into 150 bins of equal size. The ‘transcript’ track shows the number EST contigs within each bin. Two sets of contigs were mapped: a) EST contigs identified by Mascot searches against the EORF and ESTScan2 databases, and b) EST contigs identified by Mascot searches against the six-frame translation database. Pearson’s correlation coefficients ($r$-values) are shown at the bottom.
Despite the majority of EST contigs corresponding to existing gene models, a number of novel peptides were identified. Figure 4.4 presents the overlap of the combined set of peptides from the EST translations and the peptides identified from the IPI and Ensembl databases, comparing peptides taken from all PSMs filtered at an FDR of < 0.01 for all searches. The majority of peptides are observed in all three datasets. Approximately 70% of the IPI-matched peptides and 61% of the Ensembl-matched peptides were also identified using both the EORF and ESTScan2 databases. The high coverage observed reflects the quality and transcript diversity of the EST data and the sensitivity of the two algorithms, especially since only 64% of the Ensembl genes are represented by strong matches in the EST contig set. Despite the good agreement and high overlap, a significant number of peptide identifications were unique to all the database sources. Encouragingly, from a proteogenomics gene discovery perspective, 1383 peptides matched to the EORF and ESTScan2 databases were not found in the Ensembl proteome. These novel peptides may represent novel splice variants, novel genes, and polymorphisms that were missed by the Ensembl annotation pipeline. Indeed, a large number of splice-variants and exons are thought to be missing from the Ensembl annotation set (Tang et al., 2007). Finally, novel peptides could also be derived from genes that are missing from the current genome sequence, but present in the EST contigs.

4.3.4. Identification of single amino acid polymorphisms (SAAPs)

The proteomics and transcriptomics data used in this study was derived mostly from two strains of domestic chicken (White Leghorn and Rhode Island Reds), whereas the genome sequence itself obtained from Red Junglefowl and considerable polymorphism has been observed at the nucleotide level, both within the genome and in the ESTs themselves (Wong et al., 2004; Hillier et al., 2004). Since our experimental proteomic data is from the B cell-derived DT40 cell line, searched against ESTs from a variety of tissues and developmental stages, which are subsequently cross-referenced to the Red Junglefowl genome, we would expect to observe additional variation. Using proteomics, the prevalence of single amino acid polymorphisms (SAAPs) was examined across these genetically diverse datasets. For each of the EST contigs with PSMs and a strong hit to Ensembl proteins (BLASTX E-value < 1e-30; percent identity >= 90; coverage >= 50
amino acids), the matched peptides were compared against the Ensembl protein (see Methods for detail) in order to detect amino acid mismatches with respect to the genome. The vast majority of the peptides have an exact match to the Ensembl protein sequence (Table 4.3). However, a further 135 peptides containing single amino acid mismatches were identified from searches of the EORF and ESTScan2 translations. These constitute ~2% of the total peptide identifications and represent candidate SAAPs. Although some of the SAAPs may be artefacts resulting from EST sequencing errors, the generation of a consensus sequence during EST assembly will ensure that these cases are rare because polymorphisms are often supported by multiple ESTs. Many of the SAAPs are expected to reflect genetic changes due to adaptive selection during domestication (Wong et al., 2004; Rubin et al., 2010) which may therefore inform vertebrate biology and selective breeding.

EORF and ESTScan2 peptides also identified in Ensembl and IPI. The majority of PSMs are shared between the two databases, but the EORF/ESTScan2 databases contain a substantial number of unique peptides representing possible novel sequences.

Figure 4.4. Overlap between peptides identified in the different database searches (q<0.01).
Table 4.3. Amino acid polymorphisms identified by mapping PSMs identified from the ESTs, to Ensembl protein sequences. EST contigs with a BLASTX match to an Ensembl protein were retained and the identified peptides for each contig were mapped to the Ensembl protein sequence using BLASTP to determine the number amino acid matches relative to the Ensembl annotation.

<table>
<thead>
<tr>
<th>Database</th>
<th>Exact match</th>
<th>Single mismatch</th>
<th>No match</th>
</tr>
</thead>
<tbody>
<tr>
<td>EORF</td>
<td>7647</td>
<td>135</td>
<td>1834</td>
</tr>
<tr>
<td>ESTScan2</td>
<td>8366</td>
<td>138</td>
<td>2009</td>
</tr>
<tr>
<td>Six-frame</td>
<td>5184</td>
<td>90</td>
<td>1363</td>
</tr>
</tbody>
</table>

The SAAPs identified by PSMs were compared with 733 high-confidence non-synonymous SNPs (rSNPs) previously identified from the 339,314 ESTs (Wong et al., 2004). Four of the 135 candidate SAAPs confirmed the presence of rSNPs at the protein level (Table 4.4). The four SAAPs were also present in dbSNP (Smigielski et al., 2000). One of these SAAPs had already been confirmed experimentally (MSH6, ENSGALG00000008957), while three had not yet been validated. These three SAAPs were in the genes STIP1 (ENSGALG00000001032), HCCS (ENSGALG00000005571) and a hypothetical gene similar to Ovotransferrin (ENSGALG00000006453). This strongly suggests that the set of SAAPs shown in Table 4.4 contains bona fide polymorphisms. Although this represents only a small fraction of the total of 135 candidates, there is no a priori reason to suppose that a large overlap should be found with existing SNP sets, but nevertheless it is reassuring to see some have already been observed independently.
Table 4.4. Confirmation of SAAPs. Peptides containing a single amino acid mismatch with respect to Ensembl annotations were verified by matches to SNPs present in a high-confidence dataset and the dbSNP.

<table>
<thead>
<tr>
<th>Peptide Identification</th>
<th>(gene &gt; EST)</th>
<th>Ensembl protein</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VVDGATHDANAVGK</td>
<td>gtg &gt; gcg, Val &gt; Ala</td>
<td>ENSGALP00000038835</td>
<td>MSH6</td>
</tr>
<tr>
<td>DDSKVEDIWSFLSK</td>
<td>aac &gt; agc, Asn &gt; Ser</td>
<td>ENSGALP00000040343</td>
<td>Similar to Ovotransferrin</td>
</tr>
<tr>
<td>KAQQEENMDESР</td>
<td>gtg &gt; atg, Val &gt; Met</td>
<td>ENSGALP0000001527</td>
<td>STIP1</td>
</tr>
<tr>
<td>WEALHAMECPCGPSLMR</td>
<td>gtg &gt; atg, Val &gt; Met</td>
<td>ENSGALP0000008923</td>
<td>Cytochrome c-type heme lyase</td>
</tr>
</tbody>
</table>

The EORF and ESTScan2 translations also had significant numbers of PSMs with multiple mismatches to the corresponding Ensembl gene. Additionally, there were a relatively large number of EORF and ESTScan2 PSMs that did not match to Ensembl (via BLASTP) despite a good alignment for the parent EST contig with an Ensembl protein (BLASTX). The set of PSMs without Ensembl matches may represent hitherto unannotated gene structure, novel genes, or artefacts. These possibilities are considered further in the following sections.

4.3.5. Identification of novel genes

Although the chicken genome is relatively mature (Rubin et al., 2010), the annotation process is not complete and significant improvements can still be made, for example by coupling de novo predictors and RT-PCR validation (Eyras et al., 2005). However, a significant fraction of eukaryotic genomes are transcribed by RNAPol II and polyadenylated (Kapranov et al., 2007). Therefore, transcript evidence alone may be insufficient for robust protein-coding gene identification because there are likely many ncRNAs that resemble protein-coding mRNAs. Proteomics is therefore well placed to help distinguish these possibilities. The use of two independent sources of experimental data (EST and proteomics) as well as appropriate protein sequence prediction algorithms,
resulted in the discovery of novel genes that are missing from the Ensembl (v. 56) annotations, and the correction of a number of gene models.

Figure 4.5 shows the number of potential novel genes and gene model improvements made using the direct mapping approach. The stringent EXONERATE score of 2000 ensured that only high quality alignments were retained. Novel genes were identified by EST contigs having one or more significant PSM which does not overlap with Ensembl genes. A total of forty-nine novel genes were identified this way. There were also forty introns that were identified by PSMs that likely represent novel exons. Such cases most likely represent alternative splicing events which result in novel protein isoforms. Indeed, chicken is thought to undergo high levels of alternative splicing (Tang et al., 2007; Chacko and Ranganathan, 2009). The difference between the various strains of Chicken from which the EST data was derived, and the Red Junglefowl genome sequence could potentially allow the detection of strain-specific alternative splicing, which has been observed previously in planarian worms (Resch et al., 2012).

Interestingly, there were forty-one cases of EST contigs that map to Ensembl genes, but extend beyond the gene boundary (Figure 4.5). PSMs were identified within these extended regions, therefore confirming that these Ensembl gene models are incomplete. The EST-based proteogeomic approach employed here was able to identify connections between the discrete peptide and gene loci, something which would not be possible using standard six-frame-translated genome databases. Indeed, it is plausible that many novel genes identified by searching six-frame genome translations could instead represent an extension of a neighbouring gene. This highlights one of several advantages of using ESTs as a platform for peptide identifications. Another interesting feature of Figure 4.5 is that novel peptides were discovered at 12 distinct loci on chromosome ‘Un_Random’. ‘Un_Random’ is in fact not part of the assembled genome, and consists of genes (or gene fragments) that reside on orphan genomic contigs which were not assigned to a chromosome during the assembly process. However, more than 10% of the PSM-identified-ESTs map to these unassembled genomic contigs. The fact that a substantial proportion of the sequenced transcriptome cannot be mapped to the assembled chromosomes shows that the chicken genome assembly is far from complete or perfect.
Figure 4.5. Novel peptide identifications improve genome annotation in chicken.

High-confidence EST-to-genome mappings were generated by EXONERATE (score > 2000) and compared to Ensembl gene models. The novel peptide sequences identified from each EST contig were assigned genomic coordinates based on the location of the parent EST contig with respect to the genome sequence. Peptides were categorised according to their genomic loci into those that map within introns (intron), those that map to loci where Ensembl genes are absent (novel gene), and those representing gene extensions. The unassembled virtual chromosome (Un_random) is labelled as ‘Unk’ and the Mitochondrial chromosome is labelled ‘MT’. The six microchromosomes are absent from the genome assembly and are therefore not shown.

The novel peptides identified after searching EORF/ESTScan2 were visualised using a custom track on the Ensembl genome browser. Figure 4.6 shows specific examples to demonstrate the three types of novel peptides. Figure 4.6a shows a single-exon gene, discovered by an EST contig alignment. The absence of an intron in the EST-to-genome alignment would normally make this a low confidence gene prediction because it could represent an unspliced mRNAs (pre-mRNA), retrotransposed pseudogene, or ncRNA. However, the EST contig was confirmed by four PSM’s and therefore it is highly-likely that this represents a novel protein-coding gene. Moreover, a gene similar to the mammalian (mouse) SDOS gene [ENSMUSG00000022516] was predicted at this locus by the Genscan (Burge and Karlin, 1997) gene prediction algorithm (but not yet validated by experimental data), providing further support for this novel gene discovery.
Figure 4.6b shows a gene extension in the chicken PKN1 gene [ENSGALG00000006237]. In this case, Genscan predicted a much larger gene than Ensembl. Two EST contigs confirm both the 5’- and 3’- extensions to the Ensembl gene predicted by Genscan. Multiple peptide identifications show that the extended regions are protein-coding rather than UTRs. Figure 4.6c shows an example of a gene extension in the HSPA14 gene [ENSGALG00000006643], which encodes a heat shock protein. Although only a single peptide was identified in the extended region, this peptide is highly-likely to represent a genuine PSM (PEP=0.00476). The connectivity between the peptide and the adjacent gene was established from the EST-to-genome alignment, highlighting the benefit of this EST-based approach. It would be difficult to identify connectivity between peptide and Ensembl gene model through searching the genome sequence directly, and such peptides could be incorrectly interpreted as representing a novel gene.

The final class of novel peptide map within Ensembl introns and an example of this is shown in Figure 4.6d. The EST-to-genome alignment shows a number of novel exons with respect to the Ensembl gene, two of which were confirmed by the proteomics data. In addition, the EST-to-genome alignment identified a multiple exon skipping event with respect to both the Ensembl annotation and Genscan prediction. The two exon skipping events confirmed by PSMs are also compatible with the Genscan prediction, which shows that the Genscan prediction represents a novel splice-variant. As discussed in Chapter two, this extra layer of proteomics support is important to rule out the possibility of degradation of the mRNA induced by the alternative exon.

The examples in Figure 4.6. show that Genscan predicts the correct gene model. The Ensembl annotation pipeline requires experimental support for exons, and is therefore more stringent than Genscan. This will inevitably result in some genuine exons being missed by Ensembl and detected by Genscan. Although Genscan tends to make more false positive gene predictions than Ensembl, it is nevertheless a useful tool used when proteomics data is used to confirm dubious exon predictions that differ from the Ensembl annotation.
Figure 4.6. Different classes of novel peptides improve genome annotation

A) EST contig and associated PSMs map to an unannotated locus. B) PSMs map to a region upstream of the PKN1 gene. C) A high-confidence PSM confirms an extension of the HSPA14 gene. D) Exon skipping events identified in the NAA25 gene.
4.3.6. Identification of a translated pseudogene

Interestingly, we found one PSM-identified EST contig which mapped to a pseudogene [ENSGALG00000012267]. Although the transcription of pseudogenes has been observed and they have been shown to interact with protein-coding genes, in regulating the expression of their paralogous gene for example, pseudogenes by definition, do not encode functional proteins. However, in this case there is unequivocal evidence that this pseudogene is protein-coding because it is confirmed by both a mapped EST contig and two high-confidence unique PSMs. Both PSMs were identified at \( q < 0.01 \), with one having a PEP of 0.003. Moreover, the EORF protein sequence predicted from the pseudogene contains a significant match (1e-63) to UniRef90.

The term ‘pseudogene’ appears to be a misnomer in this case, because it implies that the sequence of DNA no longer encodes functional products. Indeed, a chicken ARF6 protein sequence is present in UniProt, suggesting that the pseudogene is actually a functional gene. This finding also highlights a discrepancy between Ensembl and UniProt. The results indicate that it is the UniProt annotation which is correct, and that the pseudogene is in-fact a protein-coding gene that has been misannotated by the Ensembl pipeline.

Further analysis of the EORF translation showed that it is similar to the ADP-ribosylation factor 6 protein encoded by the mammalian ARF6 gene. The ARF6 gene is also present in birds, and comparison with the equivalent Zebra Finch gene [ENSTGUT00000013743] shows that the EORF protein sequence predicted from the EST contig is almost identical to the Zebra Finch ARF6 protein, with only a single INDEL present (Appendix 8). This conservation at the protein-level would not be expected to be observed in genuine pseudogenes which no longer encode functional proteins, and highlights the importance of using sequence homology information, as well as experimental data, to annotate genomes.
4.3.7. Partial and ‘missing’ genes

The conservative EXONERATE score threshold used for identifying novel genomic features ensured the accurate mapping of peptides against the genome to reassess and improve gene models. Manual inspection of the alignments revealed that EST contigs with an EXONERATE score lower than 2000 contain a large percentage of mismatches. There are several reasons for these low-quality alignments which include both EST sequencing/assembly errors and genome assembly errors (i.e. the gene is absent from the genome assembly), as well as contaminant sequences. Figure 4.7 shows the number of low quality alignments mapped to each chromosome. These lower quality mappings typically constitute approximately 20% of the total mapping for each of the autosomes. However, the unassembled virtual chromosome Un_random or ‘Unk’ contains over 60% low-quality mappings, suggesting that it is enriched for partial or fragmented genes.

![Figure 4.7. EST contig-to-genome mappings by chromosome.](image)

EST contigs with PSMs were mapped to the assembled chicken genome and the unassembled fraction (Unk). Alignments generated by EXONERATE were separated into high-quality (Score ≥ 2000)) and low-quality (Score < 2000) alignments based on the EXONERATE score.
A substantial fraction (43%) of the lower-confidence mappings in Figure 4.7 were also matched to UniRef90 protein sequences by BLAST, and are also supported by PSMs. These low-confidence mappings are therefore likely to include genuine protein-coding genes that are missing from the current genome assembly. Indeed, genome sequences are not error-free because repetitive DNA sequences can result in the exclusion of large blocks (up to 10kbp) of DNA sequence from the assembly (Salzberg and Yorke, 2005), and this would at least partly explain the large numbers of mammalian genes that are absent from the chicken genome assembly (Hillier et al., 2004). The ability to identify protein-coding genes that are missing from the genome assembly is a useful feature of EST-based proteogenomics.

In addition to containing ESTs with BLASTX matches to UniRef90, the low-confidence mappings also include ESTs contigs that have multiple PSM’s. Of the 826 low-confidence mappings, 566 EST contigs were identified by more than two unique peptides. This set are likely to be enriched for EST contigs representing protein-coding genes that are absent from the genome sequence. Manual inspection of the low-confidence mappings reveals two types of gene: those which are partially present in the genome sequence, and those that are missing entirely. A total of 102 genes were confirmed to be completely missing from the genome assembly. There is strong evidence to suggest that these represent protein-coding genes based on their associated PSMs and similarity with UniRef90 sequences. Table 4.5 shows examples of these missing genes that were identified by multiple PSMs. Manual inspection of their alignments confirm that only spurious matches are present. The high prevalence of genes missing from genome assemblies has been reported by others (Ye et al., 2011; Alkan et al., 2011). However, the chicken genome has an additional complication due to the presence of twenty-eight microchromosomes, nine of which are completely absent from the genome assembly. Remarkably, it has been estimated that 75% of chicken genes are located on the microchromosomes (Mcqueen et al., 1998). Therefore, the missing genes identified by the EST contigs are likely to include these microchromosome genes, and their proteogenomic annotation is made feasible by the de novo EST assembly approach employed here.
Table 4.5. Examples of genes missing from the chicken genome assembly. EST contigs having an EXONERATE score of less than 1000 and a significant UniRef90 hit were considered as potential ‘missing’ genes. Five examples having multiple attendant high-quality peptide identifications are shown.

<table>
<thead>
<tr>
<th>EST contig</th>
<th>Peptide identifications</th>
<th>UniRef90 match</th>
<th>UniRef90 Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>39159.1</td>
<td>13</td>
<td>Q4R5P7</td>
<td>similar to human HLA-B associated transcript 1 (BAT1)</td>
</tr>
<tr>
<td>55163.1</td>
<td>10</td>
<td>Q7ZY48</td>
<td>Rpl13a-prov protein</td>
</tr>
<tr>
<td>357875.1</td>
<td>11</td>
<td>P52810</td>
<td>40S ribosomal protein S9</td>
</tr>
<tr>
<td>24206.1</td>
<td>9</td>
<td>Q2TBV3</td>
<td>Electron transfer flavoprotein subunit beta</td>
</tr>
<tr>
<td>343967.1</td>
<td>7</td>
<td>Q7YR37</td>
<td>ATP-binding cassette sub-family F member 1</td>
</tr>
</tbody>
</table>

The low-quality mappings also reveal cases in which genes are only partially present in the genome assembly. A common problem witnessed in genome assembly is the fragmentation of genes so that they appear on multiple scaffolds. Table 4.6 shows example cases that were identified by eight or more unique peptides. However, in these cases one segment of the EST contig maps to the genome, whereas the rest of the EST contig does not apparently match anywhere in the current assembled genome sequence. The unmapped regions of the contigs were directly identified by PSMs, thus ruling out obvious artefacts such as contamination from vector sequence or other forms of exogenous non-coding DNA present in the EST contig. This therefore suggests that these EST contigs do indeed represent genes that are only partially present in the genome assembly. Genome assembly errors such as these are likely to increase with the use of next generation sequencing (NGS) platforms, which can miss thousands of exons compared with a Sanger-sequenced reference genome (Alkan et al., 2011). Naturally, the latter is more expensive and time-consuming and arguably also benefits from a properly mapped scaffold. As an alternative, transcriptome sequencing combined with proteomics could complement these NGS genome assemblies by identifying protein-coding genes that are missing from the genome assembly.
Table 4.6. Examples of genes only partially present in the chicken genome assembly. EST contigs having an EXONERATE score of less than 2000 and a significant UniRef90 hit were considered as potential partial genes. Five examples having multiple attendant high quality PSMs are shown.

<table>
<thead>
<tr>
<th>EST contig</th>
<th>Chromosome*</th>
<th>UniRef90</th>
<th>Score</th>
<th>Peptide Identifications</th>
<th>UniRef90 Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>335791.3</td>
<td>19</td>
<td>P09972</td>
<td>1008</td>
<td>20</td>
<td>Fructose-bisphosphate aldolase C</td>
</tr>
<tr>
<td>41573.3</td>
<td>Un_random</td>
<td>Q93636</td>
<td>1378</td>
<td>8</td>
<td>Protein F29G6.3b</td>
</tr>
<tr>
<td>333047.6</td>
<td>Un_random</td>
<td>Q1HGK7</td>
<td>1272</td>
<td>9</td>
<td>Katanin p60 ATPase-containing subunit A1</td>
</tr>
<tr>
<td>345198.1</td>
<td>Un_random</td>
<td>UPI0000D9A924</td>
<td>1286</td>
<td>8</td>
<td>PREDICTED: staphylococcal nuclease domain containing 1</td>
</tr>
<tr>
<td>354701.2</td>
<td>Un_random</td>
<td>Q6Q122</td>
<td>1353</td>
<td>4</td>
<td>NADH dehydrogenase B14.5b chain</td>
</tr>
</tbody>
</table>

*Un_random is not a true chromosome; it is a collection of clone EST contigs that cannot be assigned to a specific chromosome.

4.3.8. Discovery of chicken Aldolase C

The Aldolase C fructose-bisphosphate gene is one notable example of a partial gene, which was therefore not annotated by Ensembl, and only present as fragment in IPI. Table 4.6 shows that this gene was identified by twenty peptides. Aldolase genes encode ubiquitous enzymes involved in glycolysis. In most vertebrates, there are three main aldolase isoforms (A, B and C) encoded by different genes and differentially expressed during development (Shiokawa et al., 2002). However, only Aldolase B had been annotated in chicken. A recent proteomic analysis in zebra finch identified aldolase C, raising the possibility that the chicken gene set is missing the aldolase isozyme (Sloley et al., 2007). Both Ensembl and IPI contained Aldolase B [IPI: IPI00601855] and a fragment of Aldolase C [IPI: IPI00683436] from chicken. This fragment matched completely with a
region of both the EORF and ESTScan2 translations of EST contig ‘335791.3’ and also with the translations of two similar but truncated EST contigs (335791.1 and 335791.2).

The putative Aldolase C EST contig was identified via twenty Mascot PSMs, in strong support that this is expressed as translated protein. Ten of these peptides were identified within the region of the EST contig that was absent from the genome sequence. Figure 4.8 shows an alignment of the chicken Aldolase C protein sequence predicted by our pipeline, with its Aldolase B paralog [Swiss-Prot: P07341] and two human homologs: Aldolase C [Swiss-Prot: P09972] and Aldolase B [Swiss-Prot: P05062]. The substantial majority of aligned residues were identical in all four sequences, which is perhaps unsurprising considering their essential roles in glycolysis. The putatively orthologous human and chicken aldolases had significantly greater sequence identity than the paralogs. Nevertheless, all but two of the matched peptides were unique to the chicken Aldolase C. In addition, the EORF and ESTScan2 predicted protein sequences contain significant BLASTP hits to Aldolase C in multiple species. Together, these findings suggest that the EST contig identified here represents the complete chicken Aldolase C gene.

It is perhaps surprising that such a well-studied enzyme and key component of the glycolytic pathway would have evaded annotation. Therefore the most plausible explanation for its absence is that it was missing from the genome sequence to begin with and this conclusion is supported by its low EXONERATE score. Inspection of the EXONERATE alignment in (appendix 9) reveals that only 196bp out of 780bp was matched to the genome sequence, and this region contains the Aldolase C fragment present in the IPI database. Moreover, BLAT analysis confirmed that remainder unmatched region of the EST contig did not map to any other part of the genome, ruling out the possibility of a chimeric sequence resulting from the fusion of ESTs from two or more distinct loci. This indicates that only a fragment of the Aldolase C gene is present in the genome sequence, and this would account for the fragment present in the IPI database.
Figure 4.8. Alignment of the novel Aldolase C protein sequence with chicken aldolase B and Human aldolase B and C.

A total of 20 different peptides (shaded) were all identified by Mascot searches against the EORF and ESTScan2 databases. Peptides were coloured according their level of redundancy, where yellow peptides are unique to the putative chicken aldolase C; blue peptides are also found in human aldolase C; red peptides are found in all four aldolase protein sequences.
4.3.9. Functional annotation of novel genes

The protein sequences encoded by the novel genes were investigated further to provide functional annotation. For example, one of the EST contigs was mapped, via EXONERATE (alignment score 3764), to a gene-absent locus on chromosome 22. Although no gene was annotated, the surrounding region of chromosome 22 has a relatively high gene density with approximately ~10 genes per Mb. The EST contig contained a strong BLASTX hit (expect = 1E^-125) to the lectin mannose-binding 2 (LMAN2) [UPI0000F2B5FB] protein present in UniRef90, which is a lectin involved in the transport of glycoproteins. Nine PSMs were identified from the EST contig all of which matched as exactly to the genome sequence (Table 4.5). Sequence analysis of the ESTScan2 protein prediction shows that the EST contig shares homology with LMAN2 genes in other organisms (Appendix 10). Interestingly, the monotreme Platypus LMAN2 gene shares the highest sequence identity with the putative avian chicken gene. Key regions involved in carbohydrate recognition, intracellular transport and protein-protein interactions are conserved in the chicken LMAN2 protein (Appendix 10). There were subtle differences at various sites in the chicken protein which may provide an insight into its binding specificity. For example, mammalian LMAN2 proteins contain isoleucine at the 5th carbohydrate binding site, whereas chicken contains a methionine, which was confirmed by a PSM spanning this region.

The majority of functionally important regions of LMAN2 are located in the much larger luminal carbohydrate recognition domain (Appendix 10). The amino acid residues that participate in carbohydrate binding are highly conserved amongst the orthologs, including LMAN2. Appendix 10 also shows the location of the ‘RKR’ ER retention signal which is known to be present in human LMAN2, and in the majority of LMAN2 orthologs. However, chicken and the two Xenopus species contain sequence ‘RKH’ in place of this, and zebra finch contains the sequence ‘NKR’. Therefore it appears that either chicken LMAN2, along with frog and zebra finch, contain a unique ER retention signal, or alternatively, the chicken LMAN is not localised in the ER at all.

Human LMAN2 forms homodimers, which interact via their flat beta sheet regions to form homotetramers (Loris et al., 2000). A pair of residues important for homotetramer formation differs between some of the orthologs (Appendix 10). In contrast to its
orthologs, the chicken homotetramer interaction pair consists of a leucine and isoleucine at the first and second residue respectively. This was confirmed by a peptide identification which spans the two positions and rules out the possibility that the substitutions are artefacts caused by EST sequencing errors. These amino acid substitutions, confirmed by proteomics data, reflect differences in the structure of the chicken LMAN2 compared with its orthologs, and provide valuable functional annotation for this novel gene.

4.4. Conclusion

In this chapter, the conceptual translation of ESTs via protein prediction algorithms has been investigated as a tool to improve proteogenomic annotations. The results show the potential of this process on a large scale for a vertebrate model organism, by mapping high quality PSMs to the genome. This approach is more sensitive than searching against the six-frame translation, and allows the identification of more PSMs, peptides and attendant protein-coding genes. In addition, novel peptides not present in extant annotations, from IPI or Ensembl, were identified and found to correspond to novel genes, gene model amendments, polymorphisms and pseudogenes. The de novo assembly of the ESTs allowed the identification of protein sequences that were missing or only partially present in the genome sequence, which would otherwise be undetectable using standard proteogenomic methods. The results indicate that the current gene models and the genome sequence are both incomplete, and that a combined proteomics and transcriptomics approach can refine gene models in a model organism.

Despite the methodological differences between sequencing by synthesis and chain-termination sequencing, the end products (i.e. transcript sequences) are the same. Therefore, the proteogenomics methods and bioinformatics analysis employed here are transferable to RNA-seq data. Using the latest transcriptome sequencing technologies, we would expect to attain a higher level of coverage and sequencing depth. Recent work carried out in the parasitic nematode *T. muris* by myself and colleagues, has shown that searching spectra against assembled RNA-seq data can be used to analyse the proteome of newly-sequenced organisms. We therefore propose that the integration of RNA-seq and proteomics data should play a fundamental role in future genome research.
4.5. References


Chapter 5: Conclusions

The work presented in this thesis has addressed key issues regarding the integration of high-throughput proteomics and transcriptomics data with genome sequences. Specifically, we have mapped mass spectrometry-based proteomics data with the genome sequences of several eukaryotes to make simultaneous advances to the fields of proteomics and genomics. The exponential increase in the volume and quality of mass spectrometry datasets has meant that advances in downstream bioinformatics analysis are required to make biological findings from large datasets, which are often complex and noisy. This is especially relevant to proteogenomics because errors and artefacts can manifest in at least three different types of data including mass spectrometry, transcriptome and genome sequencing. The work presented here shows how deficiencies in genome annotation and proteome peptide identifications can be overcome by integrating and combining these datasets. There were four main aspects to this work:

1) Assessing the current level of proteomics support for gene models (Chapters 2 and 3).
2) Estimating the extent of proteome coverage that is theoretically achievable using gene-centric proteomics (Chapter 2).
3) Benchmarking different methods for mapping proteomics data against the genome using EST sequences as a proxy to the genome sequence (Chapters 3 and 4).
4) Demonstrating the utility of EST-based proteogenomics for improving annotation of the chicken genome (Chapter 4).

Chapter 2 starts by looking at the level of alternative splicing annotated in current gene models for several eukaryotic model organisms and shows that the proportion of exons that are unique to a particular splice variant is low, falling to less than 10% in the C. elegans genome. This corresponds to a relatively small number of tryptic peptides that can be used to identify specific protein isoforms. Given the stochastic nature of current shotgun proteomics experiments, typically only a limited subset of peptides are identified for a given protein. Furthermore, there is a bias towards identifying peptides from proteins that are highly abundant in the sample, making the identification of “constitutive” peptides
more likely than those unique to a particular isoform. Indeed, we found that a relatively small proportion of experimentally identified peptides were mapped to unique exon regions. We also show that shotgun proteomics experiments have been ineffective at identifying splice-variants when compared with the level of alternative splicing that is theoretically detectable based on current proteomics technologies. We conclude that the missing protein isoforms can only be identified using alternative proteomics strategies, which utilise SRM and/or additional endopeptidases, and we have presented initial work directed at this. The set of isoform-specific peptides that are present in PeptideAtlas, and predicted by tools such as CONSeQuence (Eyers et al., 2011), PeptideSieve (Mallick et al., 2007), can provide a catalogue of peptides to monitor in SRM experiments. Indeed, there is also now a growing resource of SRMs available for several organisms via SRMAtlas (www.srmatlas.org) to facilitate this as well as tools to help predict, design and validate them (c.f. Skyline (Bereman et al., 2012) and MRMaid (Mead et al., 2009)). Such approaches are also amenable to the application of quantitative proteomics techniques that will be needed to help distinguish protein isoforms. By obtaining sufficient quantitative data, coupled to the knowledge of underlying isoformal gene structure, it may be possible to deconvolute protein level quantitation even when largely “semi-constitutive” peptide data is available, so long as sufficient peptides are unique to a subset of isoforms.

There are several possible reasons for the immense discrepancy between the number of identified splice-variants and the number that are theoretically detectable. One possibility is that the existing peptide identifications in PeptideAtlas are generally from “shotgun” experiments that are not optimised for identifying splice variants, or any lowly-expressed protein for that matter; most non-standard protein isoforms are expected to be expressed largely in distinct tissues or developmental stages and not ubiquitously. The bulk of high-throughput “discovery” proteomics is not targeted at such biological sophistications and is frequently “data-directed” acquisition determined by the most abundant ions in the first mass spectrometer. Alternatively, on top of the tissue- or stage-specific expression patterns, a large fraction of predicted splice variants are not translated into protein, as proposed previously based on the observation that splice-variants are predicted to disrupt protein domain architecture (Melamud and Moult, 2009). It is likely that all factors contribute to the low rate of isoform detection, and hence their relative level of contribution to the overall protein volume in cells and tissues under differing conditions.
can only be determined using targeted approaches (e.g. SRM-based experiments) which ideally are quantitative.

As discussed, there is a paucity of discriminating peptides present in current shotgun datasets. However, this should not be viewed too negatively since our results show categorically that in principle the large majority of the isoform-specific tryptic peptides ought to be experimentally tractable. This is encouraging, but dictates a targeted approach for studies aiming to characterise more than a handful of proteins and their isoforms.

The mapping of peptide identifications to existing gene models has limited discovery potential because annotation pipelines fail to correctly identify all genes and splice variants. There is increasing evidence which shows that the transcriptome is much more diverse and complex than previously thought (Ju et al., 2011; Pan et al., 2008). The current view is that transcription is pervasive (Clark et al., 2011) and dynamic (Mercer et al., 2012) resulting in thousands of uncharacterised transcripts. This has been highlighted recently by the ENCODE project which has assigned functional elements to the majority of the human genome. Moreover, there is significant variation in the types of RNA expressed between different cell-lines. Therefore, proteomics should consider newly-sequenced transcripts, in a cell type-specific manner, because new RNA sequences are emerging all the time. To address this, Chapters 3 and 4 explore the use of EST-derived databases for mass spectrometric database searches, which increases the scope of proteogenomics by allowing the discovery of proteins which are absent from publically available sequence databases. Additionally, this provides the reassurance that they are at least transcribed if not actually translated. A benchmarking study was carried out in Chapter 3 to compare different in-silico translation approaches for the identification of peptide sequences in ESTs. The EORF and ESTScan2 tools generated higher quality and more concise protein sequence databases from ESTs compared with the standard six-frame translation approach and enabled a greater number of significant peptide identifications. Primarily, this stems from improved statistical modelling to increase search sensitivity as well as the inclusion of novel and alternate genes/isoforms which support new discovery. Initially, and much to our surprise and disappointment, we found that the concatenation of the ESTScan2 and EORF sequences into a single database had a detrimental effect on sensitivity; however, this was later overcome by integrating the two sets of PSMs using the FDRScore approach (Jones et al., 2009). Coupled to the even poorer results (fewer accepted PSMs) with the
six-frame translation database we were motivated to investigate the underlying causes of these sensitivity differences. Results show that searches against six-frame translation databases lead to an over-estimate of the FDR and PEP that is primarily responsible for the sensitivity loss and fewer peptide identifications. Modifications to the six-frame database were shown to improve the sensitivity and we have shown and discussed the relative merits of each approach. The six-frame translation approach is widely used in the proteogenomic literature, but our findings suggest that such studies may be discarding many genuine peptide identifications and generating a large number of false negatives.

The work in Chapter 3 demonstrated the utility of EORF and ESTScan2 for the identification of peptides from ESTs. In Chapter 4, we integrated these translation tools into a proteogenomics pipeline for the annotation of the chicken genome. This led to identification of over a thousand novel peptides that were absent from both extant protein annotation datasets from Ensembl and IPI, over one hundred of which represent examples of novel gene models and improvements to existing gene models. The EST-based approach rectified gene models which would have been impossible or very difficult if the genome sequence was directly searched with mass spectra. The discovery of a large number of novel peptides in a model organism is perhaps surprising, but similar findings have also been reported by proteogenomics studies in worm (Merrihew et al., 2008) and mouse (Brosch et al., 2011). The work presented here goes further by discovering not only genes that were missed by annotation pipelines, but also genes that are completely missing from, or only partially present in the chicken genome assembly. It has been shown previously that the chicken genome assembly is incomplete (Ye et al., 2011), partly due to the avian-specific microchromosomes (Douaud et al., 2008), and we have demonstrated the proteomics discovery of a substantial number of these missing genes. The chicken genome assembly has undergone other revisions during its history, in part due to the misassignment of sequence between the sex chromosomes Z and W, and also due to the absence of some of the microchromosomes in the draft genome assembly. The immunological community has also noted the absence of many key MHC-linked genes in the genome which are known to be present. These issues highlight the important role proteogenomics can play in adding value to genome annotations.

The comparison of transcriptomic data from domesticated chicken with the genome sequence of the ancestral Red Junglefowl provided a unique opportunity to identify
polymorphisms and we found novel peptides that confirm non-synonymous SNPs present in the EST contigs. The extra experimental support provided by the proteomics data provides convincing evidence that effectively rules out the possibility that these SNPs have arisen as artefacts from base changes resulting from EST sequencing errors. A recent proteogenomics study has successfully used a similar approach to identify polymorphisms in multiple *M. tuberculosis* strains (De Souza *et al.*, 2011). This approach represents a novel application for proteogenomics which is complementary to traditional SNP identification methods, owing to the relative ease at which proteomics data can be generated. Again, a targeted strategy could be envisioned here, where putative SNPs are translated into the protein domain and mapped to candidate tryptic peptides. Design of appropriate SRMs and subsequent sensitive mass spectrometric analysis would support their validation, and indeed open up the possibility of directly quantifying the effects of these SNPs at the protein level.

Although proteogenomics can identify and correct many gene models from a single MS experiment, it has perhaps not been utilised as fully by the genome researchers as it could. Understandably, there is some scepticism regarding the protein inference problem because identified peptide sequences are often present in multiple protein sequences in the database searched. The so-called “one-hit wonder” problem adds further to the challenge, when a putative protein identification rests on a single peptide (Veenstra *et al.*, 2004). This is especially relevant for proteogenomics studies because the databases searched are generally much larger and a peptide sequence can be present in multiple gene models or splice-variants of the same gene. Various attempts have been made to overcome this using an Occam’s razor approach (Meyer-Arendt *et al.*, 2011; Qeli and Ahrens, 2010; Nesvizhskii *et al.*, 2003), but there are many instances where the unambiguous identification of a single protein is not possible. However we have generated a catalogue of proteotypic peptides which can be used to unambiguously identify protein isoforms (*Chapter 2*), which can potentially direct future targeted proteomics experiments.

Overall, this proteogenomics study has improved genome annotation in chicken, and several other eukaryotic organisms. But perhaps more importantly, we have developed a generic, best-practice approach for mapping shotgun proteomics data to transcript sequences, something which is highly relevant considering the rapid expansion of both types of high-throughput datasets. This is especially useful in organisms which lack a high-
quality, reference genome or in newly sequenced genomes that have not yet been comprehensively annotated. For organisms at a more advanced stage of genome annotation, we recommend using targeted proteomics experiments to confirm splice isoforms and to resolve conflicting gene models.

5.1. References


Appendixes

Appendix 1: Pairwise sequence alignment showing the differences between 2 protein isoforms encoded by the TPM3 gene.

The top sequence belongs to the ENSGALT0000002204 protein (encoded by ENSGALT00000022043), and the bottom sequence belongs to the ENSGALT00000039898 protein (encoded by ENSGALT00000040695). Tryptic peptides with associated experimental spectra are highlighted in bold text coloured background. Intron spanning and internal peptides are coloured in red and blue respectively. Broken vertical arrows represent exon boundaries that preserve codons, whereas unbroken vertical arrows represent exon boundaries that are internal to codons.
Appendix 2: Conflicting annotations of the human TGFBI gene.

The TGFBI gene was annotated as a single isoform gene in Swiss-Prot, whereas the gene is predicted to encode multiple isoforms in Ensembl (v. 48). Peptides (bold, coloured regions) were uniquely mapped to ENSP00000306306 and ENSP00000381794 suggesting two isoforms that can be independently tracked via these two peptides. However, the mapped regions are also present in the single Swiss-Prot protein, hence there could be either 3 different isoforms or a single isoform which is derived from a combination of the 2 different Ensembl splicing patterns. In this instance, the simplest explanation is that there is only 1 isoform and the Ensembl annotations are not correct although this is not proven. Interestingly, in the latest Ensembl annotation (v.55) there is a third isoform that mirrors the Swiss-Prot protein.
Appendix 3: QconCAT designed to quantify DmNav isoform expression using a trypsin digest.

TM = peptide covers a transmembrane region
Cons = peptide covers a constitutive exon
MC = missed cleavage propensity
PA = present in PeptideAtlas

<table>
<thead>
<tr>
<th>Order</th>
<th>Feature</th>
<th>Location</th>
<th>Peptide sequence</th>
<th>Mass</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>exon H</td>
<td>External</td>
<td>YLNNA(TM)MIGNSINH(QD)NR</td>
<td>2073.965</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Constitutive</td>
<td>-</td>
<td>IA(EE)H(KE)</td>
<td>796.408</td>
<td>0.450</td>
</tr>
<tr>
<td>3</td>
<td>RP49 (PA)</td>
<td>-</td>
<td>VYG(GEIAH)GVSSK</td>
<td>1348.645</td>
<td>0.437</td>
</tr>
<tr>
<td>4</td>
<td>exon D</td>
<td>TM</td>
<td>NY(H)DH(K)</td>
<td>812.357</td>
<td>0.402</td>
</tr>
<tr>
<td>5</td>
<td>Constitutive</td>
<td>-</td>
<td>LQGSFPPHEL(AS)TLEDPP(Y)YSNV(L)TLFVVVS(K)</td>
<td>3524.797</td>
<td>0.445</td>
</tr>
<tr>
<td>6</td>
<td>exon F</td>
<td>External</td>
<td>LT(Q)ID(Q)DSPGK</td>
<td>1286.647</td>
<td>0.412</td>
</tr>
<tr>
<td>7</td>
<td>exon K</td>
<td>TM/External</td>
<td>VYFTNA(W)CWLDFVIVMLSLNLA AVWSGADD(DF)PAFR</td>
<td>4102.052</td>
<td>0.403</td>
</tr>
<tr>
<td>8</td>
<td>exon I</td>
<td>External</td>
<td>SVE(VE)SESVSIQR</td>
<td>1546.784</td>
<td>0.417</td>
</tr>
<tr>
<td>9</td>
<td>exon I</td>
<td>External</td>
<td>QPA(PTTAHQ)AT(K)</td>
<td>1249.641</td>
<td>0.476</td>
</tr>
<tr>
<td>10</td>
<td>exon D</td>
<td>External</td>
<td>SW(PTLNL)LISIMGR</td>
<td>1599.881</td>
<td>0.388</td>
</tr>
<tr>
<td>11</td>
<td>exon C</td>
<td>TM</td>
<td>N(Y)TDHK</td>
<td>776.345</td>
<td>0.445</td>
</tr>
<tr>
<td>12</td>
<td>exon E / exon F</td>
<td>External</td>
<td>TNQISW(I)WSEGK</td>
<td>1447.710</td>
<td>0.445</td>
</tr>
<tr>
<td>13</td>
<td>RP49</td>
<td>-</td>
<td>HMLPTGF(K)</td>
<td>929.479</td>
<td>0.436</td>
</tr>
<tr>
<td>14</td>
<td>exon L</td>
<td>TM</td>
<td>VYFTNA(W)CWLDFVIV(M)SLNLA AVS(G)ADDVPAFR</td>
<td>3978.062</td>
<td>0.410</td>
</tr>
<tr>
<td>15</td>
<td>exon L</td>
<td>External</td>
<td>VVVNALVQAIPSIFNV(L)LLVCLIFW(L)IAIMGVQLFAGK</td>
<td>4158.391</td>
<td>0.431</td>
</tr>
<tr>
<td>16</td>
<td>exon C</td>
<td>TM/External</td>
<td>TVGALGNLTFVLCIIIIFIA(VMG)MQLFGK</td>
<td>3115.704</td>
<td>0.398</td>
</tr>
<tr>
<td>17</td>
<td>RP49 (PA)</td>
<td>-</td>
<td>GQYLM(P)NIYG(S)NK</td>
<td>1540.734</td>
<td>0.417</td>
</tr>
<tr>
<td>18</td>
<td>RP49 (PA)</td>
<td>-</td>
<td>ELEV(LLMQ)NR</td>
<td>1243.659</td>
<td>0.497</td>
</tr>
<tr>
<td>19</td>
<td>Constitutive</td>
<td>-</td>
<td>YD(DED)EGPQPDP(TEQ)GVIPVR</td>
<td>2809.257</td>
<td>0.464</td>
</tr>
<tr>
<td>20</td>
<td>exon L</td>
<td>TM</td>
<td>MQGM(R)</td>
<td>621.273</td>
<td>0.438</td>
</tr>
<tr>
<td>21</td>
<td>exon H</td>
<td>External</td>
<td>LEHELNHR</td>
<td>1046.526</td>
<td>0.492</td>
</tr>
<tr>
<td>22</td>
<td>RP49</td>
<td>-</td>
<td>FLYVHNVR</td>
<td>883.503</td>
<td>0.421</td>
</tr>
<tr>
<td>23</td>
<td>exon D</td>
<td>TM/External</td>
<td>TMGALGNLTFVLCIIIIFIA(VMG)MQLFGK</td>
<td>3147.676</td>
<td>0.414</td>
</tr>
<tr>
<td>24</td>
<td>exon E</td>
<td>External</td>
<td>LT(Q)ISDQPGER</td>
<td>1443.695</td>
<td>0.423</td>
</tr>
<tr>
<td>25</td>
<td>exon K</td>
<td>External</td>
<td>WEGMK</td>
<td>649.289</td>
<td>0.498</td>
</tr>
<tr>
<td>26</td>
<td>Constitutive</td>
<td>-</td>
<td>SLFRPFT(R)</td>
<td>1022.566</td>
<td>0.416</td>
</tr>
<tr>
<td>27</td>
<td>exon B</td>
<td>External</td>
<td>GVSVYYFPTEDDD(DED)GPTFK</td>
<td>2279.975</td>
<td>0.464</td>
</tr>
<tr>
<td>28</td>
<td>exon H</td>
<td>External</td>
<td>GLSQD(D)DTASINSYGSH(K)</td>
<td>2006.918</td>
<td>0.391</td>
</tr>
<tr>
<td>29</td>
<td>Constitutive</td>
<td>-</td>
<td>ELSVQIEQ(R)</td>
<td>1100.583</td>
<td>0.479</td>
</tr>
</tbody>
</table>
Appendix 4: QconCAT designed to quantify DmNav isoform expression using an endoproteinase Lys-C digest.

TM = peptide covers a transmembrane region
Cons = peptide covers a constitutive exon

<table>
<thead>
<tr>
<th>Order</th>
<th>Feature</th>
<th>Location</th>
<th>Peptide sequence</th>
<th>Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>exon C</td>
<td>TM</td>
<td>SWPTLNLISIMGRTVGALGNLTFLCLIHIFAVMGNOILFGK</td>
<td>4697.57</td>
</tr>
<tr>
<td>2</td>
<td>Cons</td>
<td>-</td>
<td>ELERK</td>
<td>673.38</td>
</tr>
<tr>
<td>3</td>
<td>exon D</td>
<td>TM / External</td>
<td>YYFQEGWNIFDFIVALSLLLEGVQGLS VLRSLRLRVFK</td>
<td>5047.77</td>
</tr>
<tr>
<td>4</td>
<td>RP49</td>
<td>-</td>
<td>HFIRHQSDRYAK</td>
<td>1556.80</td>
</tr>
<tr>
<td>5</td>
<td>exon K</td>
<td>TM / External</td>
<td>VVVALNVQAPSIFNVLVCLIFWLAIMGVQGLFAGK</td>
<td>4158.39</td>
</tr>
<tr>
<td>6</td>
<td>exon L</td>
<td>TM</td>
<td>VYFTNAVCWLVFIVMVSLINFVSLVGAGQIQAFLK</td>
<td>3978.06</td>
</tr>
<tr>
<td>7</td>
<td>exon D</td>
<td>TM / External</td>
<td>SWPTLNLISIMGRTMGALGNLTFLCLIHIFAVMGNOILFGK</td>
<td>4729.55</td>
</tr>
<tr>
<td>8</td>
<td>exon E / exon F</td>
<td>External</td>
<td>LNQISDQPSGERTNQISWIWSEGK</td>
<td>2873.39</td>
</tr>
<tr>
<td>9</td>
<td>exon I / exon A</td>
<td>External</td>
<td>VSTTSLSLGPSFNNIRGRSGSSHK</td>
<td>2570.37</td>
</tr>
<tr>
<td>10</td>
<td>exon F</td>
<td>External</td>
<td>LTNQISDQPSGK</td>
<td>1286.65</td>
</tr>
<tr>
<td>11</td>
<td>exon D</td>
<td>TM / External</td>
<td>NYHDHK</td>
<td>812.357</td>
</tr>
<tr>
<td>12</td>
<td>RP49</td>
<td>-</td>
<td>GQYLMNPNGYGSNK</td>
<td>1540.73</td>
</tr>
<tr>
<td>13</td>
<td>Cons</td>
<td>-</td>
<td>DIFRFSASK</td>
<td>1069.56</td>
</tr>
<tr>
<td>14</td>
<td>exon I</td>
<td>External</td>
<td>MSIRSVSESES ויש QPAPTTAHQATK</td>
<td>3265.67</td>
</tr>
<tr>
<td>15</td>
<td>RP49</td>
<td>-</td>
<td>EIVERAK</td>
<td>843.481</td>
</tr>
<tr>
<td>16</td>
<td>exon C</td>
<td>TM</td>
<td>NYTDHK</td>
<td>776.35</td>
</tr>
<tr>
<td>17</td>
<td>exon B</td>
<td>External</td>
<td>DVMLNDIIEQAAGRHSASDRGVSVYYFP TEDDEDGPTFK</td>
<td>4714.19</td>
</tr>
<tr>
<td>18</td>
<td>Cons</td>
<td>-</td>
<td>MTEEDSDISEEESLFRPFRESLVQIEEQRIAEHEK</td>
<td>4392.13</td>
</tr>
<tr>
<td>19</td>
<td>RP49</td>
<td>-</td>
<td>MTIRPAYPK</td>
<td>1231.69</td>
</tr>
<tr>
<td>20</td>
<td>exon A</td>
<td>External</td>
<td>MSIRSVSEVESLSPGSPFNNIRGRSGSSHK</td>
<td>3616.84</td>
</tr>
<tr>
<td>21</td>
<td>RP49</td>
<td>-</td>
<td>GIDNRRRRFK</td>
<td>1415.82</td>
</tr>
<tr>
<td>22</td>
<td>exon H</td>
<td>External</td>
<td>YLNNATOMIQSHNQRDRLEHELNHRLSL QDDDTASINSYGSHK</td>
<td>5091.39</td>
</tr>
</tbody>
</table>
Appendix 5: Overlap of peptides identified in pairwise database searches.

Overlap of PSMs in the searches against: a) the ESTScan2 and six-frame databases, b) ESTScan2 and EORF databases. In both cases, \( FDR_{Kaall} \) q-value cut-offs of 0.01 for the various searches are indicated by dotted lines, black for ESTScan2 and white for the six-frame or EORF searches.
Appendix 6: FDR frame modelling for PSMs accepted in the six-frame database searches.

Assuming there are a set of 12000 PSMs from a search of a target:decoy database of six-frame translations:

Proportion of true positive and false positive PSMs at different FDRs.

<table>
<thead>
<tr>
<th>Total PSMs</th>
<th>12000</th>
<th>12000</th>
<th>12000</th>
<th>12000</th>
<th>12000</th>
<th>12000</th>
<th>12000</th>
<th>12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total target</td>
<td>6000</td>
<td>6300</td>
<td>6650</td>
<td>7040</td>
<td>7500</td>
<td>8000</td>
<td>8600</td>
<td>9200</td>
</tr>
<tr>
<td>TP target</td>
<td>0</td>
<td>600</td>
<td>1300</td>
<td>2080</td>
<td>3000</td>
<td>4000</td>
<td>5200</td>
<td>6400</td>
</tr>
<tr>
<td>Total decoy</td>
<td>6000</td>
<td>5700</td>
<td>5350</td>
<td>4960</td>
<td>4500</td>
<td>4000</td>
<td>3400</td>
<td>2800</td>
</tr>
<tr>
<td>FP target</td>
<td>6000</td>
<td>5700</td>
<td>5350</td>
<td>4960</td>
<td>4500</td>
<td>4000</td>
<td>3400</td>
<td>2800</td>
</tr>
<tr>
<td>FDR</td>
<td>1</td>
<td>0.90</td>
<td>0.80</td>
<td>0.70</td>
<td>0.60</td>
<td>0.50</td>
<td>0.40</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Expected number of PSMs in each reading frame at different FDR's (assuming Frame 1 is correct)

<table>
<thead>
<tr>
<th>FDR</th>
<th>Frame 1</th>
<th>Frame 2</th>
<th>Frame 3</th>
<th>Frame 4</th>
<th>Frame 5</th>
<th>Frame 6</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR = 1</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>6000</td>
</tr>
<tr>
<td>FDR = 0.9</td>
<td>1550</td>
<td>950</td>
<td>950</td>
<td>950</td>
<td>950</td>
<td>950</td>
<td>6300</td>
</tr>
<tr>
<td>FDR = 0.8</td>
<td>2191.67</td>
<td>891.667</td>
<td>891.667</td>
<td>891.667</td>
<td>891.667</td>
<td>891.667</td>
<td>6650</td>
</tr>
<tr>
<td>FDR = 0.7</td>
<td>2906.67</td>
<td>826.667</td>
<td>826.667</td>
<td>826.667</td>
<td>826.667</td>
<td>826.667</td>
<td>7040</td>
</tr>
<tr>
<td>FDR = 0.6</td>
<td>3750</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>7500</td>
</tr>
<tr>
<td>FDR = 0.5</td>
<td>4666.67</td>
<td>666.667</td>
<td>666.667</td>
<td>666.667</td>
<td>666.667</td>
<td>666.667</td>
<td>8000</td>
</tr>
<tr>
<td>FDR = 0.4</td>
<td>5766.67</td>
<td>566.667</td>
<td>566.667</td>
<td>566.667</td>
<td>566.667</td>
<td>566.667</td>
<td>8600</td>
</tr>
<tr>
<td>FDR = 0.3</td>
<td>6866.67</td>
<td>466.667</td>
<td>466.667</td>
<td>466.667</td>
<td>466.667</td>
<td>466.667</td>
<td>9200</td>
</tr>
<tr>
<td>FDR = 0.2</td>
<td>8333.33</td>
<td>333.333</td>
<td>333.333</td>
<td>333.333</td>
<td>333.333</td>
<td>333.333</td>
<td>10000</td>
</tr>
<tr>
<td>FDR = 0.1</td>
<td>9983.33</td>
<td>183.333</td>
<td>183.333</td>
<td>183.333</td>
<td>183.333</td>
<td>183.333</td>
<td>10900</td>
</tr>
</tbody>
</table>
Expected percentage of PSM's in the correct reading frame, taking into account the error associated with correct frame assignments

<table>
<thead>
<tr>
<th>FDR</th>
<th>Correct frame (%)</th>
<th>Approximate error (%)</th>
<th>Lower bound (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>100.0</td>
<td>7.8</td>
<td>92.2</td>
</tr>
<tr>
<td>0.10</td>
<td>91.6</td>
<td>7.02</td>
<td>84.6</td>
</tr>
<tr>
<td>0.20</td>
<td>83.3</td>
<td>6.24</td>
<td>77.1</td>
</tr>
<tr>
<td>0.30</td>
<td>74.6</td>
<td>5.46</td>
<td>69.2</td>
</tr>
<tr>
<td>0.40</td>
<td>67.1</td>
<td>4.68</td>
<td>62.4</td>
</tr>
<tr>
<td>0.50</td>
<td>58.3</td>
<td>3.9</td>
<td>54.4</td>
</tr>
<tr>
<td>0.60</td>
<td>50.0</td>
<td>3.12</td>
<td>46.9</td>
</tr>
<tr>
<td>0.70</td>
<td>41.3</td>
<td>2.34</td>
<td>38.9</td>
</tr>
<tr>
<td>0.80</td>
<td>33.0</td>
<td>1.56</td>
<td>31.4</td>
</tr>
<tr>
<td>0.90</td>
<td>24.6</td>
<td>0.78</td>
<td>23.8</td>
</tr>
<tr>
<td>1.00</td>
<td>16.7</td>
<td>0</td>
<td>16.7</td>
</tr>
</tbody>
</table>

This data analysis models the expected number of PSMs in the "correct" reading frame, presuming perfectly accurate FDR modelling. It also factors in the observed 7% of PSMs that are in the "wrong" frame owing to other reasons, such as frame shift errors or mistakes propagated from BLAST.

The two columns in red shown are the values used to plot the dotted line in the figure 4.
Appendix 7: Cumulative frequency distributions of Mascot scores, for target and decoy PSMs from two database searches.

A larger fraction of decoy PSMs have higher scores in the six-frame searches in comparison to the ESTScan2 searches, and *vice versa* for target PSMs. This leads to higher estimate of false positive target PSMs at an equivalent Mascot score, and a more conservative FDR estimate for the six frame searches.
Appendix 8: EORF prediction of a protein sequence encoded by the chicken ARF6 gene which was mis-annotated as a pseudogene in Ensembl.

The EORF predicted sequence was aligned with the Zebra Finch ARF6 protein using EMBOSS pairwise alignment tool.
Appendix 9: Exonerate alignment of the EST contig containing the Aldolase C gene.

Query: 335791.3
Target: 19 dna_rm:chromosome chromosome:WASHUC2:19:1:9939723:1:[revcomp]
Model: est2genome
Raw score: 1046
Query range: 345 -> 780
Target range: 5906247 -> 5706120

```
346 : GGGCATCGTTGAGCATCAAGGTGACCAAAAGTGTCGTC-TCATTTTCTGTGAGCTGAC : 399
5906247 : GGGCCGCTTCCTGGGAGAGCTGTCGTGCTGATGGG---CTGT---CCGAGCAGCTCGG : 5906199
400 : AGATGGGAGGACACACCACGCTGAGGCTGAGTGATGCCGAGAAGG : 449
5906198 : GCCTGGGAGGAGCA---G-AGGGTTTGG---G-GGCTGTCTGTGAGGAGCTGTCG : 5906150
450 : CTCAGTACAAGAAAGATGGGGCTGAC <<< Target Intron 1 <<< T : 476
5906149 : CCCAGGA-GGCTGGGAGGAGGAGCCAGCgt..................acA : 5706519
477 : TGGCAAGTGCGGTGCGTGTCGCATGAGCTCTACAGGCTCTATGACAGG : 5706518
706518 : TGGCAAGTGCGGTGCGTGTCGCATGACAGGCTCTATGACAGG : 5706464
525 : GCCCTGCGCC-ATCATG-GAAAAGCGGGCCAGTGTGCGCCAGACGGCCACACAGCAGTCTA : 575
5706463 : GCCATGGCTATTTGAAGGGCTCCCACAATTGAGGGCCAAAATGCTGAGG : 5706409
576 : --TGCCACGAGAAGGTATAGTCGCCACATGTTGGAAGCCAGAGATCCTGTTCTGTGTCG : 628
5706408 : GGTGTATGCAGAAGGCTGATGCCTTGAAGCCAGAGATCCTGTT GTG : 5706353
629 : GACCCAGCAGCTCAAGAGGGAGGTCACAGATGTGACAGAGAAAG <<< Target Int : 668
5706352 : GACCCAGCAGCTCAAGAGGGAGGTCACAGATGTGACAGAGAAAG : 5706311
669 : ron 2 <<< GTGCTGGGGCTGCTCAACAGGCGCTCAGCAGACCAGCAGTCTA : 711
5706310 : ron 2 <<< GTGCTGGGGCTGCTCAACAGGCGCTCAGCAGACCAGCAGTCTA : 711
712 : C TCGAGGGGACCTGCTTCGAAAGACACATGTTGACACCCAGGAGGATCTTTCGTGCCCA : 767
5706189 : C AgTGCAGAGCAGCTGCTGATGAGCCACACAGCACGCAGTCTA : 5706134
768 : CCAAGTACAGCCC : 780
5706133 : CCAAGTACAGCCC : 5706121
```
Appendix 10: Multiple sequence alignment of the chicken LMAN2 novel protein and 16 putative orthologs.

PSMs are highlighted in alternating colours for clarity. Blue highlighted regions include: (I) carbohydrate binding sites; (II) Homotetramer interaction sites; (III) Transmembrane regions; (IV) ER retention signal.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_110432.1</td>
<td>Human</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>XP_001150690.1</td>
<td>Chimpanzee</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>XP_001100192.1</td>
<td>Macaque</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>XP_002710070.1</td>
<td>Rabbit</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>XP_001492091.1</td>
<td>Horse</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>Q2HJD1.1</td>
<td>Cow</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>XP_863934.1</td>
<td>Dog</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>NP_001013392.1</td>
<td>Mouse</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>XP_001379939.1</td>
<td>Opossum</td>
<td>IRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>034124_3_ESTScan_Chicken</td>
<td></td>
<td>---------------</td>
</tr>
<tr>
<td>XP_001508603.1</td>
<td>Platypus</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>NP_001136372.1</td>
<td>WC Frog</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>AA125510.1</td>
<td>AC Frog</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>NP_991282.1</td>
<td>Zebra Fish</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>NP_0019213623.1</td>
<td>Zebra Finch</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>NP_508151.1</td>
<td>Nematode</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
<tr>
<td>NP_651224.1</td>
<td>Fruit Fly</td>
<td>VRLTPDMQSKQGALNWRVPFCFLRDDELQVHFKIHQQKKNLHGGLAIYW 134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accession</th>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_110432.1</td>
<td>Human</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>XP_001150690.1</td>
<td>Chimpanzee</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>XP_001100192.1</td>
<td>Macaque</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>XP_002710070.1</td>
<td>Rabbit</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>XP_001492091.1</td>
<td>Horse</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>Q2HJD1.1</td>
<td>Cow</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>XP_863934.1</td>
<td>Dog</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_001013392.1</td>
<td>Mouse</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_001379939.1</td>
<td>Opossum</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>034124_3_ESTScan_Chicken</td>
<td></td>
<td>---------------</td>
</tr>
<tr>
<td>XP_001508603.1</td>
<td>Platypus</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_001136372.1</td>
<td>WC Frog</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>AA125510.1</td>
<td>AC Frog</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_991282.1</td>
<td>Zebrafish</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_0019213623.1</td>
<td>Zebra Finch</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_508151.1</td>
<td>Nematode</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
<tr>
<td>NP_651224.1</td>
<td>Fruit Fly</td>
<td>TKDRMQPGPGVFGLNMDKFGVGVFLGVPDYPNEKEQKQΕ---------------RVFP 173</td>
</tr>
</tbody>
</table>

*:..:*:::*::*::*::*::*:**
NP_110432.1 | Human
XP_001150690.1 | Chimpanzee
XP_001100192.1 | Macaque
XP_002710070.1 | Rabbit
XP_001492091.1 | Horse
QNZJD1.1 | Cow
XP_868344.1 | Dog
NP_001013392.1 | Mouse
XP_001379939.1 | Opossum
034124_3_ESTScan | Chicken
XP_001508603.1 | Platypus
NP_001136372.1 | WC Frog
AAI29541.1 | AC Frog
NP_991288.2 | Zebrafish
XP_002191623.1 | Zebra Finch
NP_508151.1 | Nematode
NP_651224.1 | Fruit fly

IV

ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348
ALFLIVFSLVFSVFAIVIGIIYLNYKQEQSRRKRY 348

III

*:*:*