Chapter 5

Refining annotation of the mouse genome using mass spectrometry

5.1 Introduction

This chapter applies the work of the previous chapters in the form of a pilot study to validate and extend genome annotation for *Mus musculus* as available through Ensembl and Vega. In section 1.2 the current strategies of genome annotation, including Ensembl and Vega, were discussed in detail and a brief introduction to the field of proteogenomics was provided in section 1.3.

In this work I build upon these efforts and apply a two stage search strategy with the aim of validating and refining mouse (Waterston *et al.*, 2002) genome annotation for the first time. MS data, obtained from the Peptide Atlas project (Desiere *et al.*, 2006) and generated in-house, was first searched against a peptide centric non-redundant superset of Ensembl, Vega and IPI (Kersey *et al.*, 2004) that was generated with GenoMS-DB (see chapter 4). IPI, commonly used as a standard protein database for MS proteomics, was included for completeness. It is expected that these databases comprise most of the proteome and due to the limited search space, peptide identification sensitivity is maintained at a high level. In a second stage, I have incorporated protein predictions from Augustus that significantly inflate search space, but enable refinement of existing gene annotations. These data were then used to validate existing Ensembl and Vega gene models at the gene, exon and splice-boundary level. Interestingly, I show evidence of alternatively translated protein variants and discuss the implications of not detecting any translational evidence for transcripts that are tagged with nonsense mediated decay (NMD) (Maquat, 2005), which are discussed in section 5.3.2.5. Furthermore, I highlight the value of proteogenomics to refine gene structures: significant peptide identifications were made outside annotated coding regions as well as within annotated pseudogenes. Novel exons or exon boundaries, as well as a set of novel genes that are not annotated in existing databases, were also identified. Lastly, the pre-computation of genome mapping for all peptides, as provided through GenoMS-DB, enabled me to assess for the first time not only the value of proteogenomics for observed peptides, but also offer a perspective of what could theoretically be achieved with this approach.

Parts of this chapter will be published by the author of this thesis (Markus Brosch), my supervisors (Tim Hubbard, Jyoti Choudhary), Lu Yu and Mark Collins who run the mass spectrometry experiments and Jennifer Harrow and co-workers who will further investigate the results in collaboration with the HAVANA team at the Wellcome Trust Sanger Institute.

5.2 Methods

5.2.1 Tandem mass spectrometry data

This pilot study is based on 10,465,149 tandem MS spectra, where 729,583 spectra were obtained from in-house experiments on nuclear protein extracts of murine embryonic stem cells and murine brain membrane fractions. These experiments were performed by Lu Yu and Mark Collins and the experimental procedures follow the methods described in section 2.2 (sample 1).

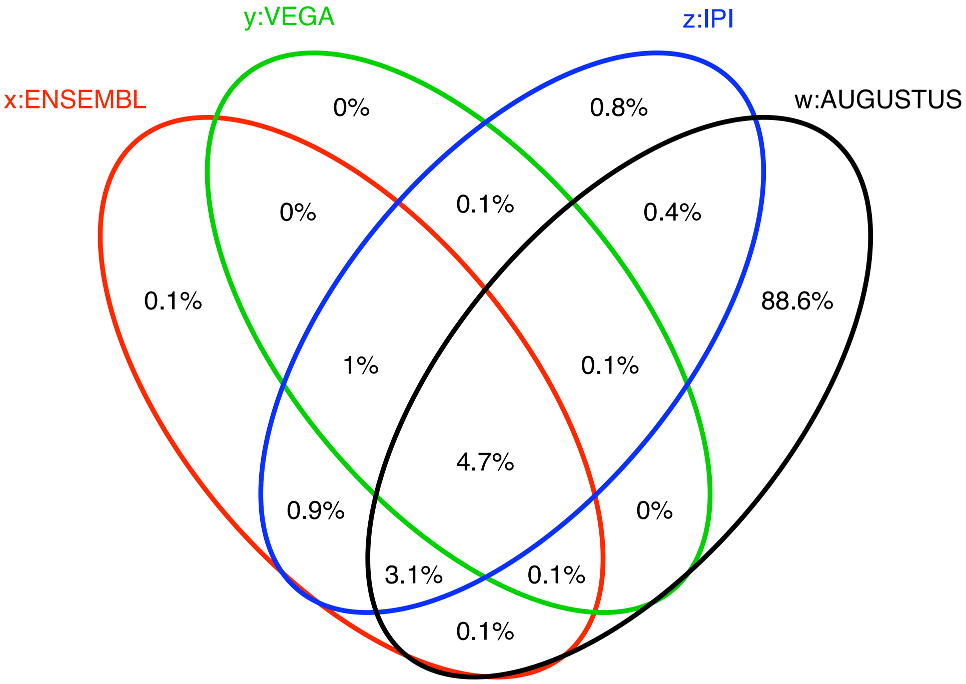
9,735,566 spectra were provided by Eric Deutsch and Zhi Sun (Institute for Systems Biology, Seattle, US) as Mascot mgf peaklist files. These data were selected from the *Mus Musculus* Peptide Atlas data repository (unpublished, Feb. 2009 data snapshot, `http://www.peptideatlas.org/repository/`). Data were not associated with any publication records, but short descriptions suggested sampling across various tissues of mouse such as brain, liver, lung, heart, kidney, testes and placenta.

5.2.2 Search database construction

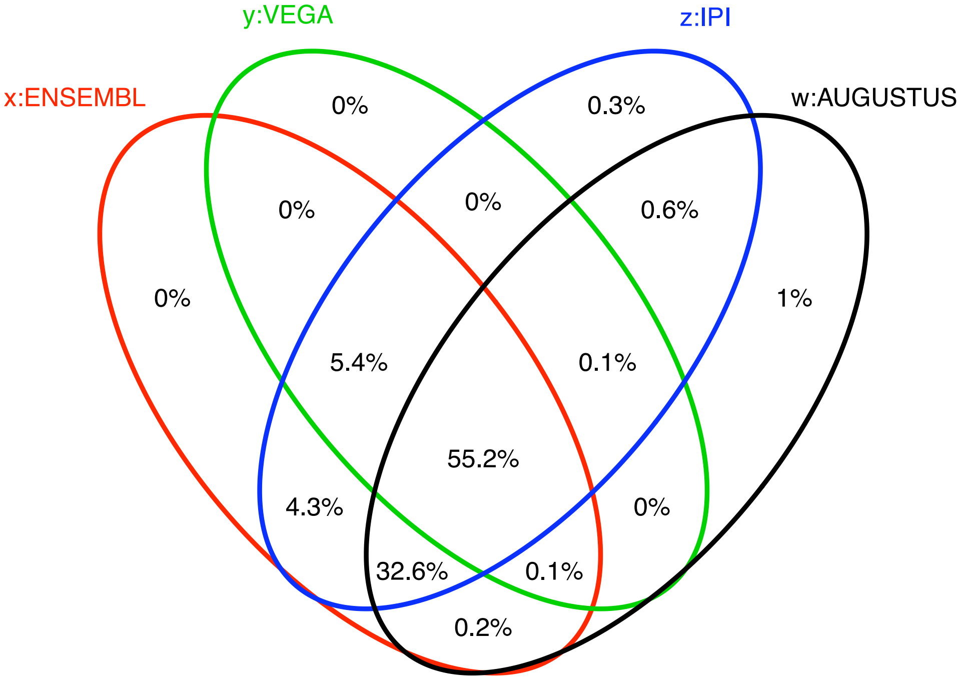
All gene products from Ensembl (mouse, release 54) and Vega (mouse, release 35, December 2008) as well as all protein entries from the IPI database (mouse, v3.55) were tryptically digested *in silico* (cutting after arginine and lysine), allowing up to two missed tryptic cleavages. Protein N-terminal methionine excision by Methionine aminopeptidase (Frottin *et al.*, 2006) was considered and therefore the N-terminus peptide of a protein was staggered. In addition, all potential 2,690 mouse NMD products (internal data release, February 2009) and common external contaminants from cRAP (a maintained list of contaminants and laboratory proteins provided through the Global Proteome Machine Organisation, `http://www.thegpm.org/crap/index.html`) were appended and processed in the same way. In total, 3,276,592 distinct tryptic peptides where generated and integrated into GenoMS-DB together with the corresponding genomic context (see chapter 4). Figure 5.1a illustrates the peptide distribution between the different data sources.

The search database (FASTA flatfile) for Mascot was built by concatenating all tryptic peptides in a non-redundant manner, thereby eliminating multiple occurrences of a peptide in alternative gene products as described in section 4.2.3. The artificial residue “J” was introduced as a spacer element to separate individual peptides, similar to the method described by Schandorff *et al.* (2007).

A second search database was constructed that extends the former database by *ab initio* Augustus (version 2.0.3) gene predictions, resulting in 28,742,036 distinct pep-



(a) All *in silico* digested peptides



(b) All identified peptides ($PEP \leqslant 0.01$, filtered)

Figure 5.1: Four-way Venn diagram of all distinct fully tryptic peptides from Ensembl, Vega, IPI and Augustus.

tides. For this, DNA sequences (*Mus Musculus*, NCBI37) for each chromosome were downloaded from the ensembl data resource (`ftp://ftp.ensembl.org/pub/current_fasta/mus_musculus/dna/`) and Augustus was run on all chromosome sequences, each of which was chopped into 50 Mb slices, overlapping by 2.5 Mb. The Augustus release provided a script (`join_aug_pred.pl`) to re-assemble predictions from individual slices and those that spanned the slice boundaries. The resulting file in GFF format was processed and converted into tryptic peptides in the same manner as described above and imported into GenoMS-DB.

In total, three individual Augustus runs were performed: (a) standard mode, (b) over-prediction mode and (c) single exon gene mode. The standard mode (a) used the recommended default parameters that provide similar performance as other gene prediction tools (Guigo and Reese, 2005; Stanke *et al.*, 2006). In mode (b) Augustus was run with parameter settings that provide maximum sensitivity and also allowing for shorter gene predictions. When Augustus is used directly for genome annotation purposes without any subsequent validation, false positive predictions are generally unwanted and a trade-off between sensitivity and accuracy needs to be made. However, here the aim was to minimize false negatives and thereby maximize sensitivity. The false positive Augustus gene predictions are controlled in the MS peptide-spectra correlation stage with stringent and robust scoring, essentially acting as a validator for this large set of potential genes. Lastly, in mode (c) Augustus was optimized to predict single exon containing genes, which are known to be difficult to annotate. The detailed parameters for these customised runs (b) and (c) were as follows (provided by the author of Augustus, Mario Stanke, personal communication, November 2008):

b) The parameter `/Constant/min_coding_len` in the configuration file `config/-species/human/human_parameters.cfg` was set to 50. The Augustus program parameters were set to: `--sample=1000 --maxtracks=10 --minexonintron-prob=0 --minmeanexonintronprob=0 --alternativesfromsampling=true`.

c) Same parameters as in (b) were used, but additionally the following changes were made in the configuration file `config/model/trans_shadow_partial.pbl`: single exon (final intergenic region) was set to: `1 0 10` and reverse single exon (intergenic region) was set to `24 0 10`.

For both FASTA databases, corresponding decoy databases were constructed for significance assessment (see section 1.1.2.3). However, the default Mascot decoy method was not sufficient: Mascot randomizes each protein sequence (FASTA entry), while retaining the average amino acid composition and length. This does not suffice for the FASTA entries that were artificially constructed in this work, where fully tryptic peptides are concatenated with a spacer residue "J". The chance of obtaining an arginine (R) or lysine (K) residue immediately before "J" when the decoy script is applied, would be approximately 10% (2 in 20 residues), meaning that the decoy database would be significantly depleted in "real" potential decoy matches. I therefore implemented a Perl script that shuffles each unique peptide entry individually by maintaining the tryptic cleavage site, instead of shuffling a whole FASTA entry. After each randomization round, it is tested whether the peptide was either produced before or exists as a natural peptide in the target database. In both of these cases, the randomization process is continued until a new random fully tryptic peptide was determined. Overall this process maintains the trypticity, the amino acid composition, the peptide length distribution as well as the number of peptides in an entry.

5.2.3 Data processing and database searching with Mascot

In-house LTQ-FT and LTQ-FT Ultra (Thermo Fisher Scientific) generated MS raw data files were processed to peak lists with BioWorks (version 3.2 and 3.3, Thermo Fisher Scientific). Processing parameters were identical to those used in section 2.2.3.

All MS peaklist data (in-house and PeptideAltas) were searched with Mascot and post processed with Mascot Percolator. For this, each peaklist file was searched against both target and decoy databases using an enzyme setting that is compatible with the custom made peptide centric search databases; therefore the artificial amino acid “J” was introduced under the mascot config file that defines the amino acid masses. “J” was set to a mass that does not correspond to a naturally occurring amino acid (300 Da). The enzyme was set to cut at the N- and C-terminal of the peptide, thereby only fully tryptic peptides that were separated by “J” were searched with Mascot. For in-house data, parameters were identical to 2.2.4 with the parent mass tolerance set to 20 ppm. Peptide Atlas data was searched with the parameters supplemented with the data file.

5.2.4 Post processing with Mascot Percolator and results integration

Mascot search results were post-processed with Mascot Percolator (1.09, default settings) using Percolator version 1.12. Each peptide-spectrum match was assigned a q-value and a posterior error probability. All sequences that either had a posterior error probability greater than 0.05, or matched any entry of the contaminants protein list, were filtered out. The remaining peptide identifications were integrated into GenoMS-DB (see section 4.2.4).

The Distributed Annotation Server (Dowell *et al.*, 2001) (DAS), which is build on top of GenoMS-DB (see section 4.2.5), provides access to the results of this pilot study. Meta-information for each peptide is provided in the form of scoring statistics (q-value, $-10 \times log_{10}$ transformed posterior error probability), genomic uniqueness of the peptide within the Ensembl/Vega and Ensembl/Vega/Augustus annotation, Mascot search log ID and spectrum ID. The DAS data source can be accessed at `http://das.sanger.ac.uk/das/ms_das/` and can be readily integrated into genome browsers that allow embedding external DAS sources.

5.3 Results and Discussion

5.3.1 Peptide identification and genome mapping

It is expected that the superset of the Ensembl, Vega and IPI database represents most of the mouse proteome and was therefore used for the first pass Mascot search. After post-processing with Mascot Percolator, in total 1,491,410 and 1,772,159 peptides were identified at a q-value (a more advanced notion of the false discovery rate, see section 1.1.2.2) of 1% and 5%, respectively. Applying a maximum allowed probability of 1% and 5% of an individual peptide match to be incorrect (posterior error probability, PEP), 1,124,724 and 1,358,323 peptides were identified, corresponding to a q-value of less than 0.14% and 0.59%, respectively.

When data was searched against the database that was supplemented with the Augustus predictions (see methods), 16% fewer identifications (1,253,074 and 1,490,020 at a q-value of 1% and 5%) were made due to the search space inflation of almost one order of magnitude (figure 5.1a), which increases the chance of incorrectly identifying peptides and hence more restrictive scoring was required in order to maintain the q-value (this is discussed in detail in chapter 2). 967,131 and 1,171,060 peptides were identified with a maximum PEP of 1% and 5%, corresponding to q-values of 0.12% and 0.57%, respectively. Interestingly, 88.1% of the distinct peptide identifications ($PEP \leqslant 1\%$) overlapped between Ensembl and the Augustus predictions (figure 5.1b), suggesting good sensitivity for the chosen Augustus configuration.

For subsequent analyses, only the best PEP and q-value score for each peptide sequence was considered (PEP significance threshold $\leqslant$ 5%), resulting in 95,606 distinct peptide identifications, 3,260 of which matched common contaminants. Since all fragment ion (MS/MS) data were generated with collision induced fragmentation (Biemann, 1988; Roepstorff and Fohlman, 1984) and analyzed with a low resolution instrument, Leucine/Isoleucine as well as Lysine/Glutamine sequence isoforms could not be differentiated due to identical or similar residue mass. Therefore, sequences

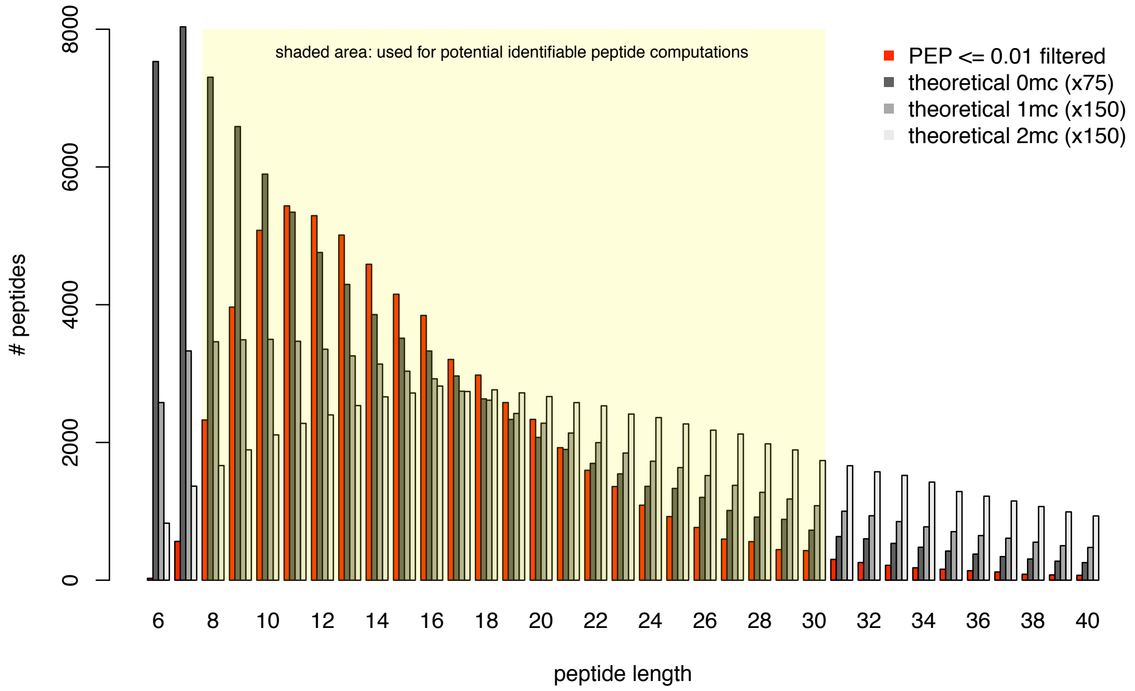


Figure 5.2: Peptide length distribution of identified peptides that passed the filtering criteria (red) and of the potential identifiable peptides as derived from the protein digest (black/grey). The number of peptides for the latter are scaled down by a factor of 75 for peptides without a missed cleavages (0mc) site (black) and by 150 for peptides with one (1mc, mid-grey) or two (2mc, light-grey) missed cleavages respectively. When theoretical genome annotation coverage was computed in this work, only peptides ranging from 8 to 30 residues were considered (shaded area).

that have an isoform in any of these residues were filtered out (1,159 cases). In total, 83% (76,029) of the remaining peptides mapped unambiguously to one genomic locus. Since only fully tryptic peptides were considered, it was further tested whether a semi-tryptic form of the peptide sequence mapped elsewhere in the genome (758 cases). As a last measure, it was evaluated whether peptides with one residue substitution or indel could be identified elsewhere in the genome (6,685 cases, preferentially short peptide identifications), since coding SNPs were not considered in this study. Therefore, a total of 68,586 distinct peptides built the basis for subsequent genome annotation. However, peptide-spectrum matches with a PEP between 1-5% were exclusively used as supplementing peptides and only peptide identifications with a PEP of 1% or better (58,574 cases) were used as a primary annotation data source.

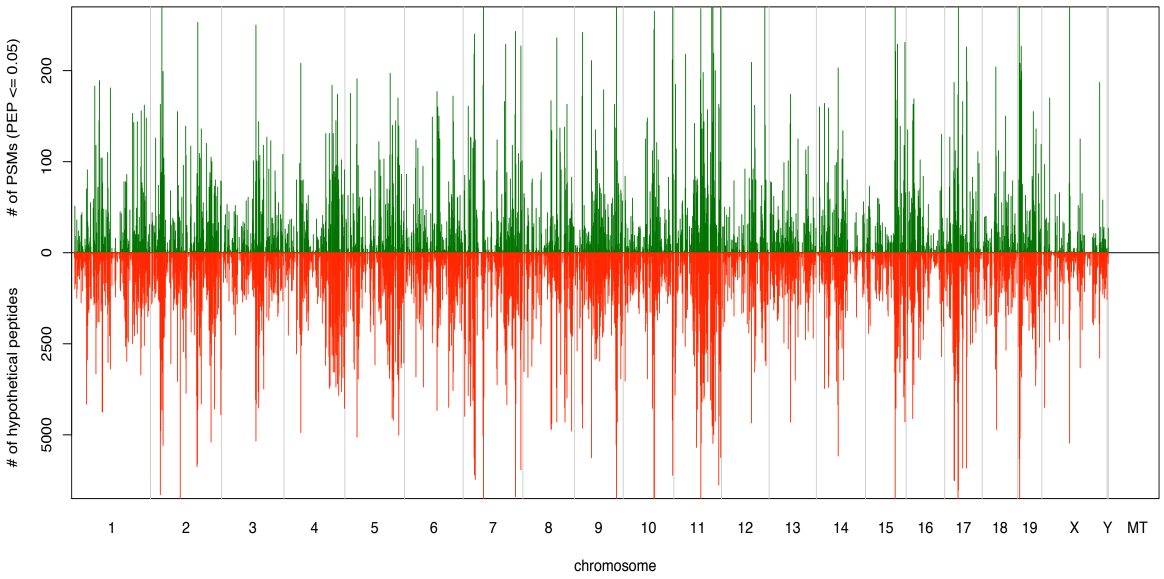
This meant that the chance of a wrong peptide identification would be 1% in the worst-case scenario, corresponding to a false discovery rate of less than 0.14%. Most proteogenomics research studies to date have used a false discovery rate of 1% to 5%, but I have taken a conservative approach to avoid the propagation of erroneous identifications into genome annotation pipelines.

5.3.2 Ensembl/Vega annotation validation

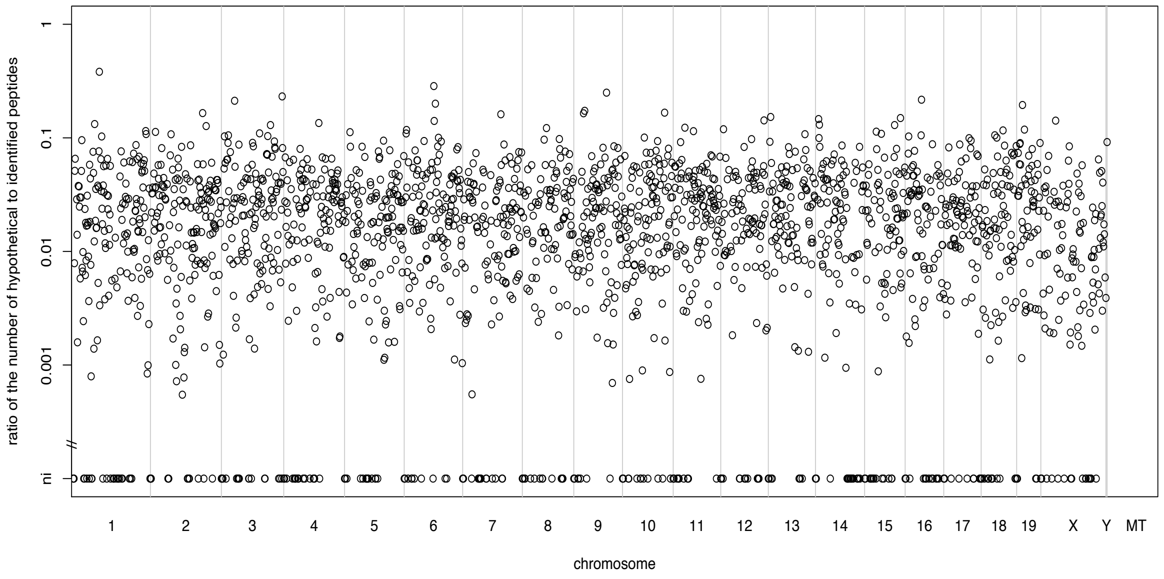
98.1% of all identified peptides ($PEP \leqslant 1\%$) matched the Ensembl/Vega database with only 1.9% attributed solely to IPI and Augustus (figure 5.1b). Therefore I focus first on confirming Ensembl/Vega annotation at the level of gene translation and structure.

5.3.2.1 Genome coverage

Figure 5.3a shows the distribution of fully tryptic peptides across the genome. Each chromosome was binned into 1Mb blocks and the number of potentially identifiable (all *in silico* digested peptides), as well as the number of identified peptides were calculated to evaluate genome coverage. Gene density varies across mouse chromosomes (Waterston *et al.*, 2002) and each gene contains peptides over a range of three orders of magnitude (see next section), the number of identifiable peptides per 1Mb block is also highly variable (nil to 12,910 peptides, median 715). The number of identified peptides (median 10) per 1Mb block is not only dependent on the number of identifiable peptides, but most notably on the expression level of the gene products, which determines the number of peptides that can be sampled by the MS instrument (Ishihama *et al.*, 2005; Lu *et al.*, 2007). The ratio between identified and identifiable peptides varied by more than two orders of magnitude (figure 5.3b). These results indicate that there was no mistake or bias in the data processing towards certain chromosomes and a more in-depth analysis can be conducted.



(a) Peptide counts of all potential identifiable peptides (red) and all peptides that have been identified (green) are plotted for each chromosome. Note the different y-axis scale.



(b) Relative peptide identification rate as defined by all peptides that have been identified versus all potential identifiable peptides.

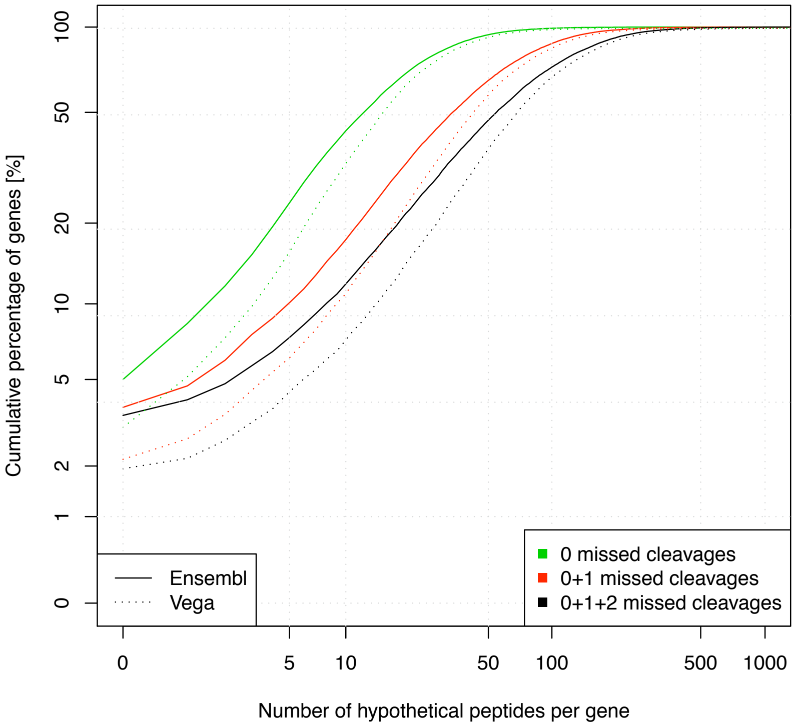
Figure 5.3: Chromosomes were binned into 1Mb blocks and absolute (a) and relative (b) peptide counts were evaluated, allowing the evaluation of peptide coverage at a genome scale.

5.3.2.2 Verification of gene translation

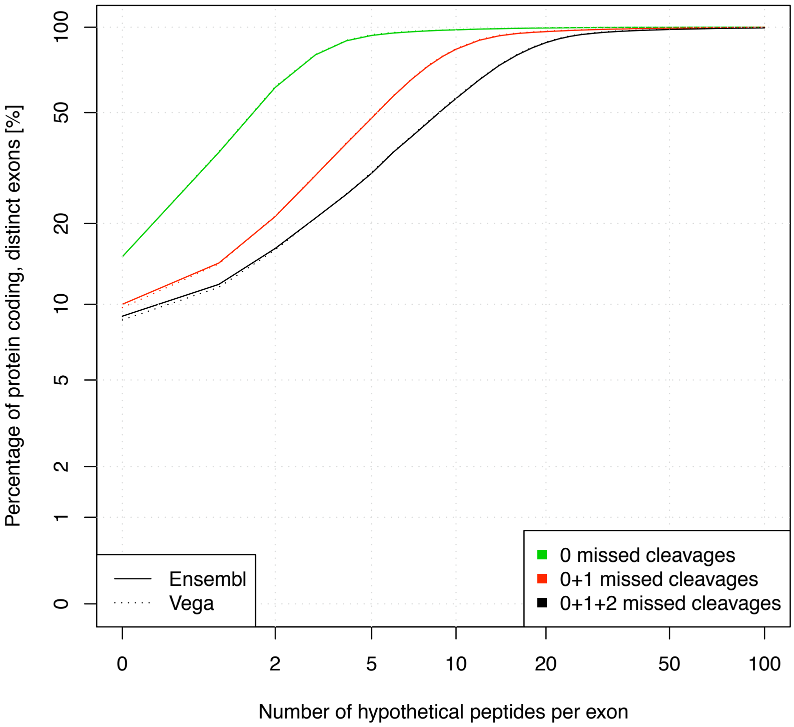
Figure 5.4a shows the cumulative percentage of genes that could be validated theoretically by tryptic peptides that map uniquely to a genomic locus and comprise between eight to 30 amino acids. These are the default peptide parameters for all theoretical considerations in the remainder of this work (peptide length distribution is illustrated in figure 5.2) and no predictions were made about proteotypic peptides (Fusaro *et al.*, 2009; Mallick *et al.*, 2007). Interestingly, when nil, one and two missed cleavages were allowed, 5.0%, 3.8% and 3.5% protein coding Ensembl gene products contain no tryptic peptides and 43.0%, 17.4% and 11.9% contain only ten or fewer peptides respectively. This could potentially limit the chances of gene validation, given that not all peptides are amenable for MS analysis (Fusaro *et al.*, 2009).

Nevertheless, a significant number of 7,221 (4,463) protein coding Ensembl (Vega) genes could be validated with peptide identifications that mapped uniquely to one gene, corresponding to 31.6% (36.7%) of all protein coding genes. However, peptide coverage was limited, with only 7.9% (9.0%) of the genes being validated by more than ten peptides and 0.08% (0.09%) by more than 100 peptides (figure 5.6a).

In order to further study the relationship between identified and potentially identifiable peptides, it was tested whether a linear model could be fitted (figure 5.5). A perfect fit would mean that the MS instrument would sample more peptides from gene products with more potential peptides. However, it was found that there is no correlation ($R^2 = 0.10$) and this is consistent with the above statement that peptide sampling is mainly determined by relative protein abundance. Moreover, genes that are only expressed in specific tissues would not be identified if the tissue of interest was not analyzed. For example, *Obscurin* (ENSMUSG00000061462) is a muscle protein and is amongst the top ten genes with most potentially identifiable peptides (1192) and yet none of the peptides were identified, suggesting that it was either expressed at very low abundance or not at all in the tissues or cell lines that were analyzed (see also `http://tinyurl.com/Obscurin`). In contrast, *Plectin-1*



(a) Cumulative gene identification rate as a function of the number of potential identifiable *in silico* digested peptides per protein coding gene.



(b) Cumulative exon identification rate as a function of the number of potential identifiable *in silico* digested peptides per protein coding exon.

Figure 5.4: Theoretical gene and exon validation rate. Note: considered peptides where fully tryptic, ranged from 8-30 residues and were unique to a genomic locus.

(ENSMUSG00000022565), a cytoskeletal protein that is more widely expressed, has a similar number of potential peptides (1447), but has the highest number of identified peptides (280). Other genes amongst the top five genes with highest peptide coverage (> 170 distinct peptides) include: *Spectrin alpha chain 2 (Spna2) brain* (ENSMUSG00000057738), *Bassoon (Bsn)* (ENSMUSG00000032589), *Cytoplasmic dynein 1 heavy chain 1 (Dync1h1)* (ENSMUSG00000018707) and *Spectrin alpha chain brain 1 (Spna1)* (ENSMUSG00000020315). All of these proteins are very large (275-533 kDa) and therefore smaller proteins at the same expression level would always result in lower peptide coverage.

It is important to note that the consideration of missed cleavages makes a significant difference. Allowing missed cleavages results in better gene coverage, which can be explained by the fact that peptides with missed cleavages tend to be longer. Trypsin is a very specific enzyme, but is not always 100% efficient. In fact, 31.7% and 9.9% of all identified peptides in this study have one and two missed tryptic cleavage sites respectively and only 58.4% have no missed cleavage sites, hence about 90% of the peptides have none or one missed tryptic cleavage site.

Overall I show that proteomics MS data is of significant value for confirming genes, some of which could be validated with extensive peptide coverage. Considering that currently eukaryotic proteomes are far from being saturated (de Godoy *et al.*, 2006), gene validation coverage of proteogenomics data will further increase as improved methods and instrumentation allow for deeper proteome sequencing, theoretically enabling validation of considerable portion of the genes.

5.3.2.3 Gene structure validation

A similar analysis at the exon level, using the same peptide properties as before, revealed that 15.1%, 10.0% and 9.0% of all Ensembl protein coding exons do not contain detectable peptides when nil, one or two missed cleavages are allowed, respectively. In addition, 93.6%, 47.8%, and 30.4% of the protein coding Ensembl

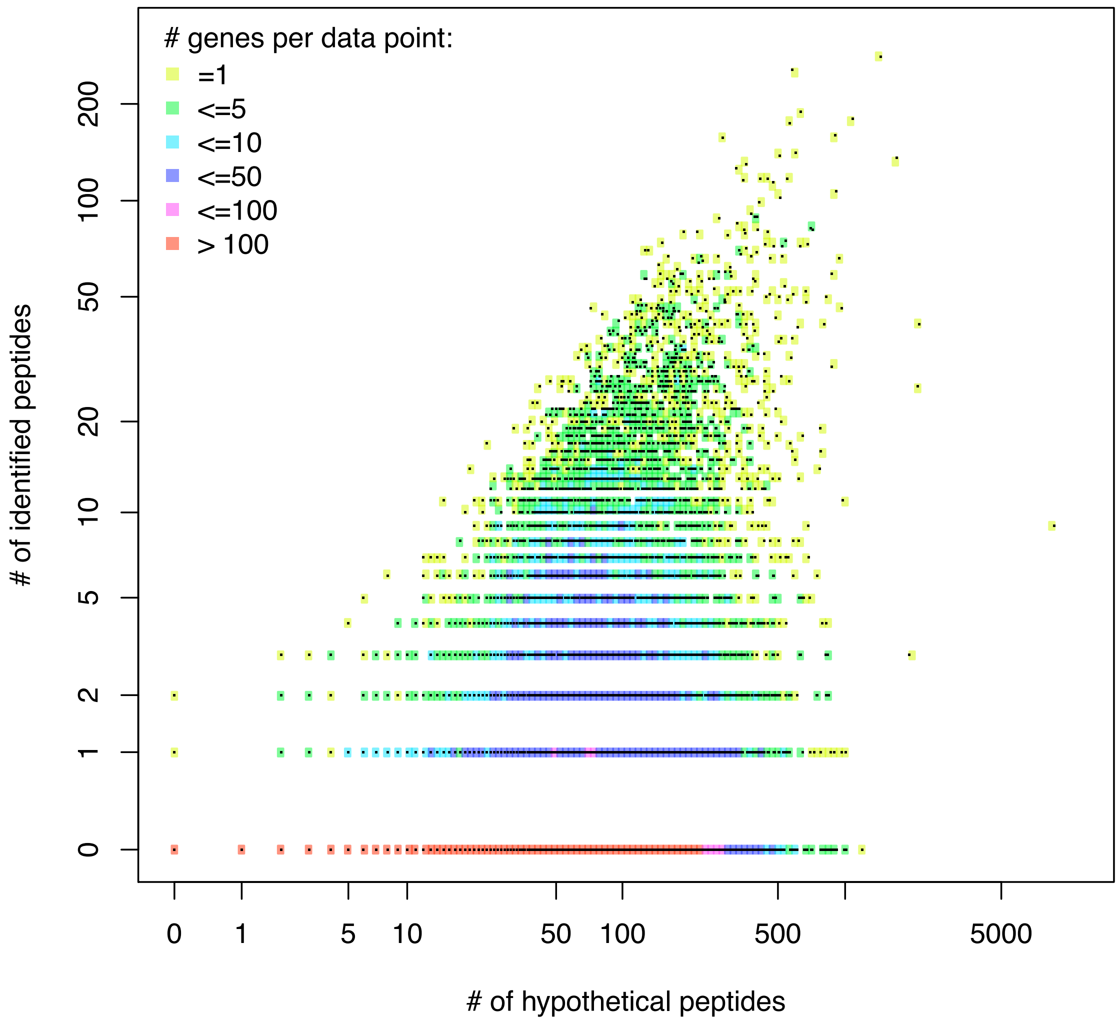
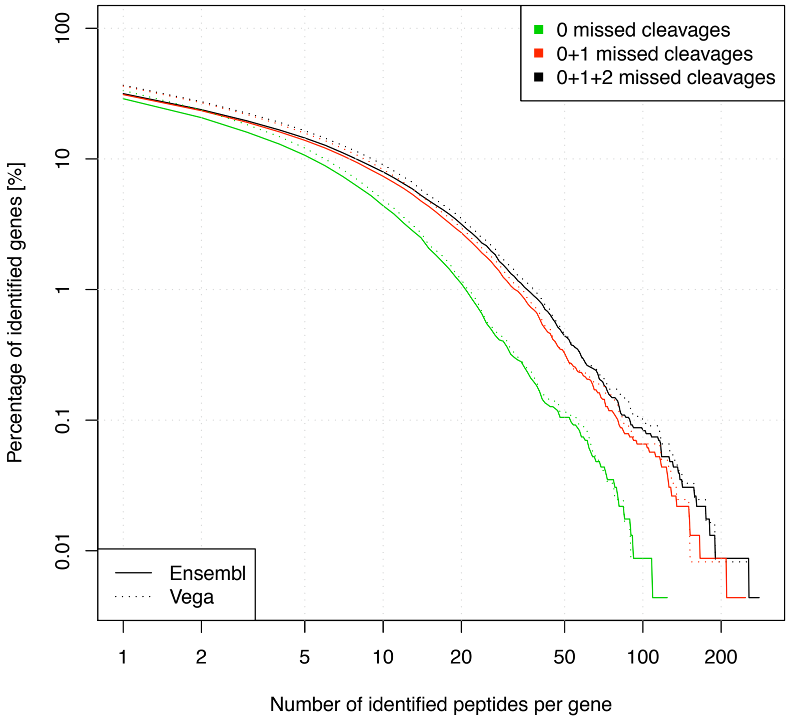


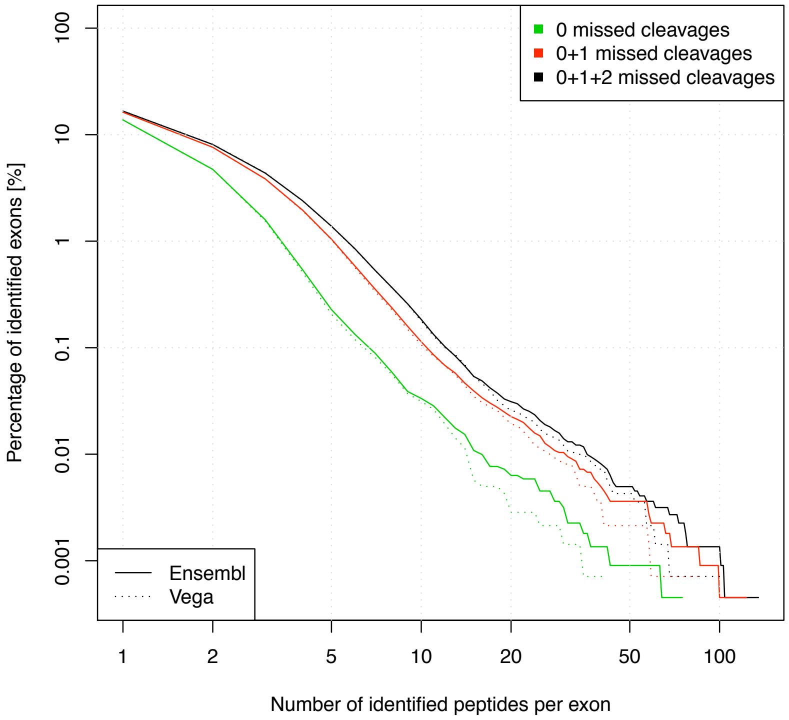
Figure 5.5: Correlation analysis between the number of identified peptides and the number of potential identifiable peptides per gene. Since many data points have the same x-y values, the number of overlaying data points (genes) is encoded with the color gradient (available from the legend).

exons contained five or fewer peptides, respectively (figure 5.4b). The lower peptide coverage compared with complete genes can be explained by the fact that the average protein coding exon count per gene in mouse is around 9.7.

A total of 16.7% of the total 222,378 Ensembl protein coding exons could be validated by peptide identifications. About 8.0% and 1.4% of Ensembl exons were validated by at least two and five peptides, respectively (figure 5.6b). Validation rates for Vega were insignificantly different.



(a) Inverse cumulative validation rate of all protein-coding genes as a function of the number of peptides identified per gene.



(b) Inverse cumulative validation rate of all protein-coding exons as a function of the number of peptides identified per exon.

Figure 5.6: Observed gene and exon validation rate using identified peptides.

A more difficult challenge is to validate annotation of introns, since this requires a fully tryptic and unique peptide spanning splice boundaries to be identified. Defining the accurate splice donor and acceptor sites is not trivial and a peptide spanning these sites not only validates them, but implicitly also validates the joined exons and thereby significantly contributes to gene structure validation.

Of the 202,205 (131,336) introns in Ensembl (Vega) that span a protein-coding splice boundary, up to 70.9% and 86.2% could theoretically be confirmed by peptides, allowing for one or two missed cleavages, respectively. However, when only peptides without missed cleavages are considered, the theoretical validation rate drops to 46%.

Using the subset of identified peptides that span a splice site, a total of 14,426 (9,347) Ensembl (Vega) introns could be confirmed, corresponding to 7.1% of all splice sites that join protein coding exons in both Ensembl and Vega, 1.3% of which were validated with two or more distinct peptides.

Clearly, the value of translational evidence is indispensable for independent gene structure validation. Notably up to 91.0% of all protein coding exons and 86.2% of all introns could theoretically be confirmed with peptides obtained in typical proteomics experiments. Applying the peptides identified in this study, 16.7% of all exons and 7.1% of all introns could be confirmed, highlighting that with relatively moderate efforts a significant proportion of gene structures can be validated.

5.3.2.4 Evidence of alternative translation

Until recently, only limited evidence was available of alternatively expressed transcripts at the protein level (Tress *et al.*, 2008). The detection of these variants by standard MS proteomics experiments is hindered by the fact that the majority of protein sequence is shared between transcripts, differing only in small parts of the translation products. Validation of alternative translation requires identification of at least one “signature” peptide for each protein isoform. 8,877 (40%) protein coding Ensembl genes code for alternative products, but only 16,664 transcripts from 1,542 genes could theoretically be discriminated by 168,726 “signature” peptides. For example, *Catenin delta-1* (ENSMUSG00000034101) has 25 alternative transcripts annotated as coding, but only nine “signature” peptides could theoretically distinguish the alternative translation of three protein isoforms.

Nevertheless, protein evidence for alternatively translated genes from tryptic digests was shown recently; Tanner *et al.* (2007) found evidence for 16 human genes, Castellana *et al.* (2008) found evidence for 47 Arabidopsis genes and Tress *et al.* (2008) identified 130 drosophila genes that express at least two protein isoforms. Here, a total of 370 peptides enabled discrimination of 112 transcripts in 53 genes, corresponding to 3.4% of all protein coding genes with multiple isoforms that can be discriminated by a peptide. *UDP-glucuronosyltransferase 1-2 Precursor* (ENSMUSG00000054545), which has 12 alternative coding transcripts within one locus, is unusual as all variants have an alternative 5’ exon spliced to a common set of downstream constant exons. These variable first exons confer diverse functional mRNAs with different tissue specific expression profiles (Zhang *et al.*, 2004). Figure 5.7 shows an overview of this complex locus with evidence for expression of five alternative protein isoforms from 27 “signature” peptides. Other examples with evidence for three alternative gene products include: *ankyrin 2 brain isoform 2* (ENSMUSG00000032826), *Synaptotagmin-7* (ENSMUSG00000024743) and *Core histone macro-H2A*.1 (ENSMUSG00000015937). Two alternative isoforms were validated for each of the remaining 49 genes.

Even though the overall rate of peptide identifications that could be attributed to alternative protein isoforms is low in proteogenomic studies due to only few available “signature” peptides that are unique to one isoform, these results demonstrate evidence for the presence of alternative splice variants *in vivo*. It would be interesting to follow-up this study with a more sensitive hypothesis driven targeted proteomics approach (Anderson *et al.*, 2009; Arnott *et al.*, 2002), in which the mass spectrometer is directed to scan specifically for “signature” peptides of individual protein isoforms.

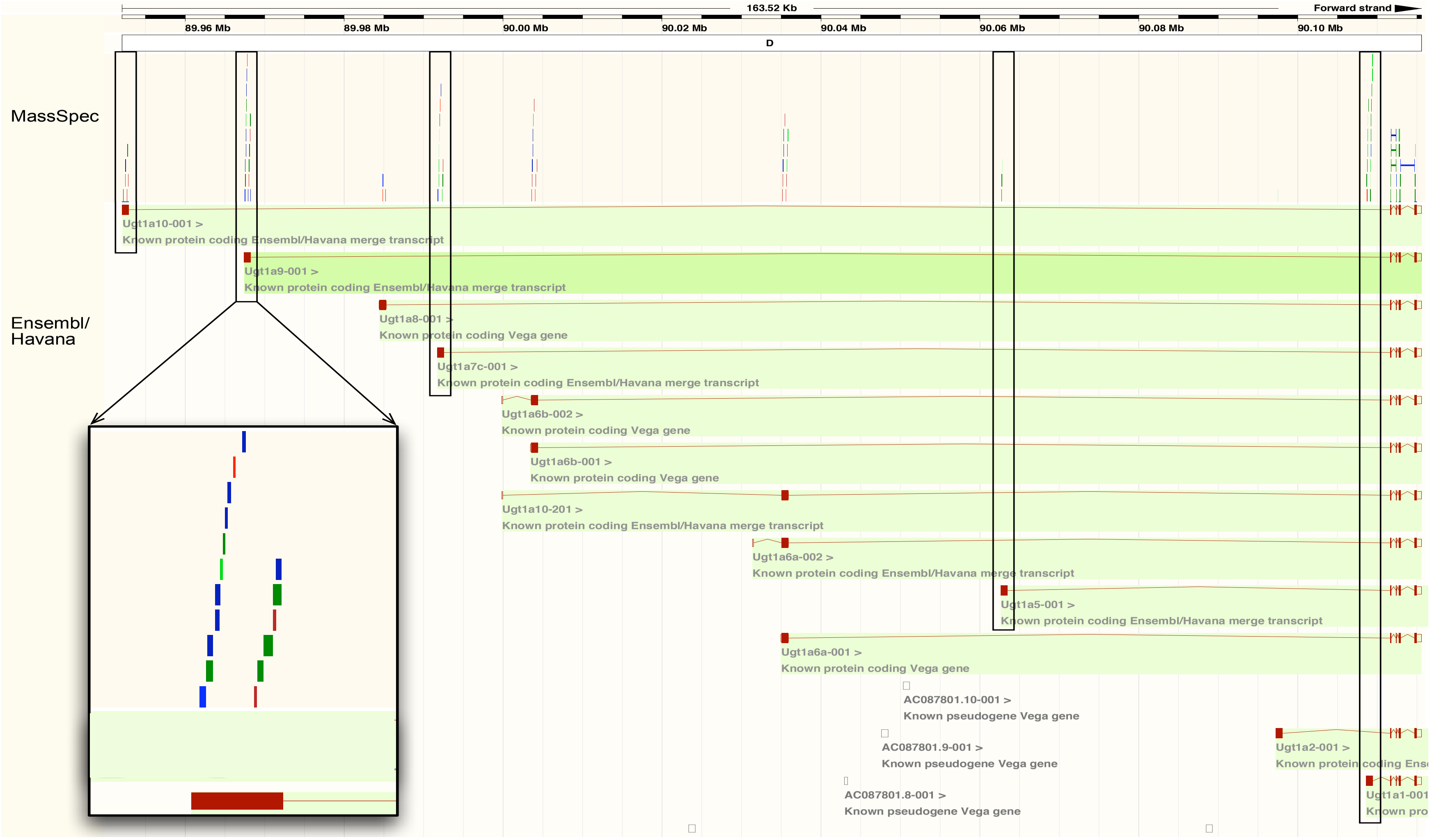

163.52 Kb
Forward strand
89.96 Mb
89.98 Mb
90.00 Mb
90.02 Mb
90.04 Mb
90.06 Mb
90.08 Mb
90.10 Mb
D
MassSpec
Ensembl/ Havana
Ugt1a10-001 >
Known protein coding Ensembl/Havana merge transcript
Ugt1a9-001 >
Known protein coding Ensembl/Havana merge transcript
Ugt1a8-001 >
Known protein coding Vega gene
Ugt1a7c-001 >
Known protein coding Ensembl/Havana merge transcript
Ugt1a6b-002 >
Known protein coding Vega gene
Ugt1a6b-001 >
Known protein coding Vega gene
Ugt1a10-201 >
Known protein coding Ensembl/Havana merge transcript
Ugt1a6a-002 >
Known protein coding Ensembl/Havana merge transcript
Ugt1a5-001 >
Known protein coding Ensembl/Havana merge transcript
Ugt1a6a-001 >
Known protein coding Vega gene
AC087801.10-001 >
Known pseudogene Vega gene
AC087801.9-001 >
Known pseudogene Vega gene
Ugt1a2-001 >
Known protein coding Ens
AC087801.8-001 >
Known pseudogene Vega gene
Ugt1a1-001
Known pro

Figure 5.7: Peptide evidence that is specific to five alternative gene products of *UGT1A2*. Each isoform was identified with a set of signature peptides that were unique to only one variant (highlighted in figure, one example magnified). This assumes complete annotation since novel gene variants may have shared peptides with existing gene variants. Peptide colour codes, see figure 4.4.

5.3.2.5 Nonsense mediated decay

Nonsense mediated decay (NMD) is a translational-coupled mechanism that eliminates mRNAs containing premature translation-termination codons (PTCs) (Brogna and Wen, 2009) and is estimated to effect 75-90% of human genes (McGlincy and Smith, 2008). The exact mechanism of how NMD occurs in mammals is still under debate (Brogna and Wen, 2009). Some known proteins e.g. NRAS have transcripts that appear to escape the NMD since they contain a PTC but still a functional protein appears to be produced. Since the Vega database contains annotation of transcripts predicted to be subject to NMD, I used the MS data to test whether any of the NMD transcripts actually produced a detectable translated protein. The search database contained 2,690 NMD transcripts, which would allow identification. However, only 1,704 NMD transcripts could theoretically be validated by 9,202 potential "signature" peptides. Interestingly, I have not been able to identify any "signature" peptides that would suggest the translation of NMD transcripts. Using Fisher's exact test, this result is significantly different (p-value 5×10^{-9}) from what would be expected by chance (20 peptides) when compared with the conservative peptide identification rate that could be attributed to alternative transcripts. This reinforces the theory that transcripts flagged with NMD indeed undergo degradation with a short half life. On the other hand, these proteins may not be expressed at all or at a very low level, hindering detection by the MS instrument.

5.3.3 Gene model correction

Peptide identifications are also of great value for correcting gene structures, only limited by the fact that the protein sequence needs to be in the search database to be identified in the first place. For this reason, the search database was supplemented with Augustus predictions, containing about ten-fold the number of peptides compared to Ensembl (see methods and figure 5.1a). Moreover, peptides derived from

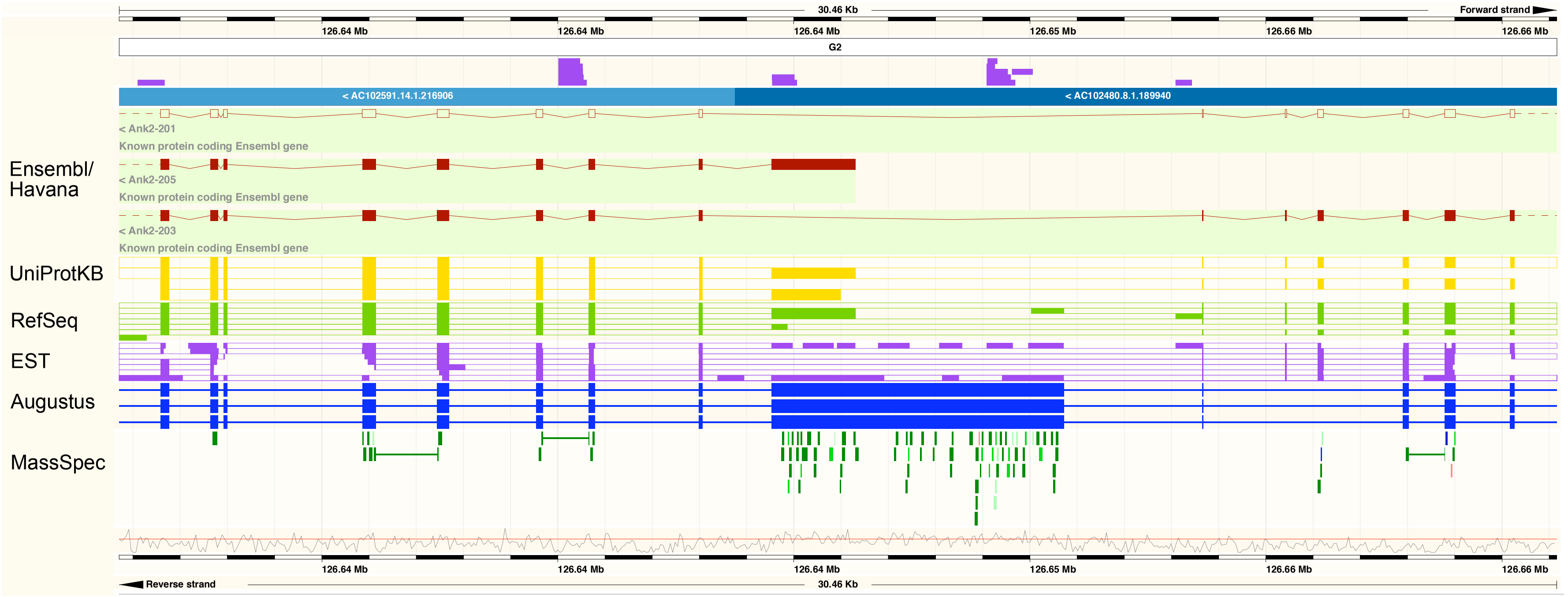


Figure 5.8: Gene *Ankyrin 2* (brain isoform 2) is annotated in Ensembl with three alternative transcripts (red). The Augustus runs predicted a large exon, not present in Ensembl/Havana, that was validated by multiple peptide identifications (“MassSpec” track), suggesting either incorrect Ensembl annotation or a new alternative isoform. This was further supported by transcriptional evidence (RefSeq and EST track).

the IPI database could be indicative of differences between the genomic annotation and the protein database. Therefore the superset of Ensembl, Vega, Augustus, IPI and the cRAP contaminants database was used to search the MS data with Mascot for a second round for extended analysis.

1.9% of all peptide identifications matched neither Ensembl nor Vega, but were present in the IPI database or Augustus gene predictions, indicating a significant number of identifications that contribute to gene structure refinements or novel genes (figure 5.1b). These identifications do not fall into the expected number of incorrect identifications (0.12% false discovery rate at the chosen 1% PEP threshold) and were therefore further investigated.

5.3.3.1 Gene model refinements

Predicting the correct gene structure and defining the exact donor and acceptor splice site remains one of the most difficult problems in genome annotation. Using peptide data that was searched against the Augustus database, a total of 168 intron refinements could be made, which include the correction of splice donor and acceptor sites, skipping exons, as well as the introduction or refinement of novel exons.

Figure 5.8 shows one example where Augustus predicted an exon extension that was not annotated in any of the Ensembl/Vega transcripts but was validated by 52 distinct peptide identifications. This clearly suggests that either the existing annotation was incorrect or a novel isoform was found. This example demonstrates the power of searching tandem-MS data against an over-predicted genome to detected flaws or missing annotation.

Refinements identified in this work will be manually investigated by the HAVANA team in-house and future Vega releases will have validated refinements incorporated.

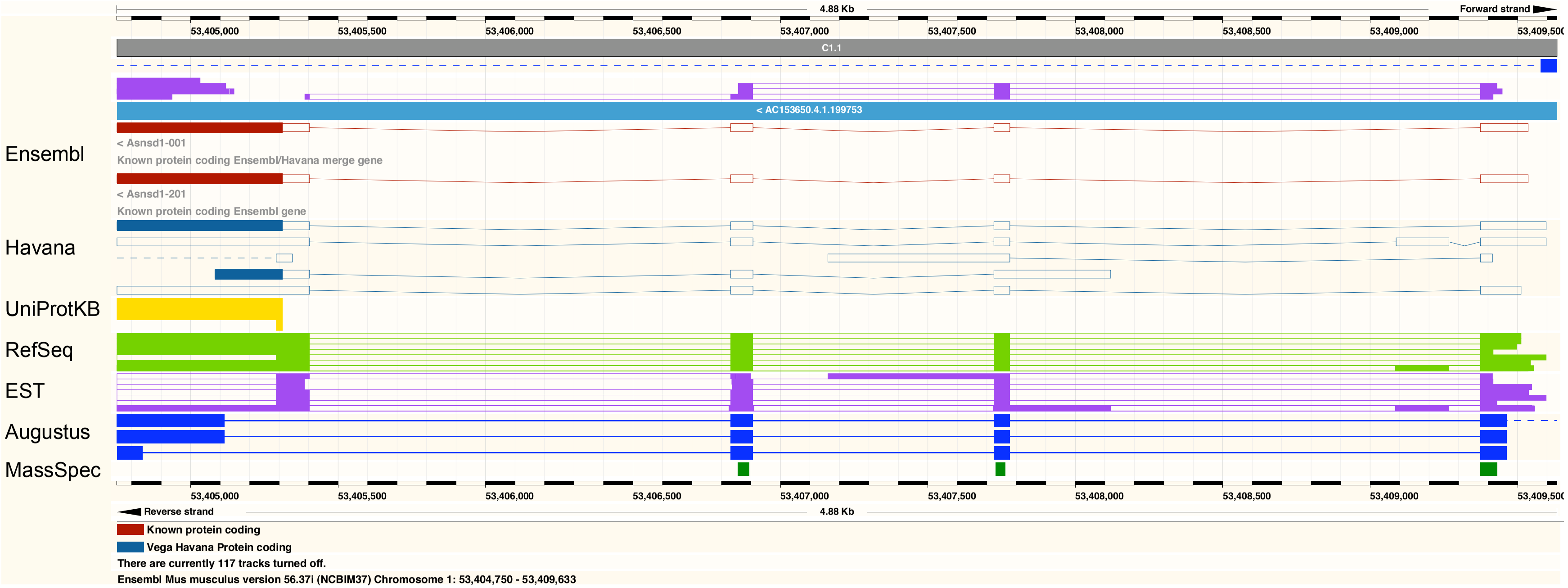


Figure 5.9: Three exons, part of the annotated 5' UTR of gene *Asnsd1* (ENSMUSG00000026095), were confirmed as coding by high confident peptide identifications ($PEP < 10^{-4.5}$), which are indicated in the "MassSpec" track. EST and cDNA evidence (EST and RefSeq track) support these exons, however, the existing protein evidence (UniProtKB track) suggests a translational start site downstream of these identified peptides, which may have led to the existing Ensembl/Havana annotation.

5.3.3.2 Translational evidence for annotated non-coding regions

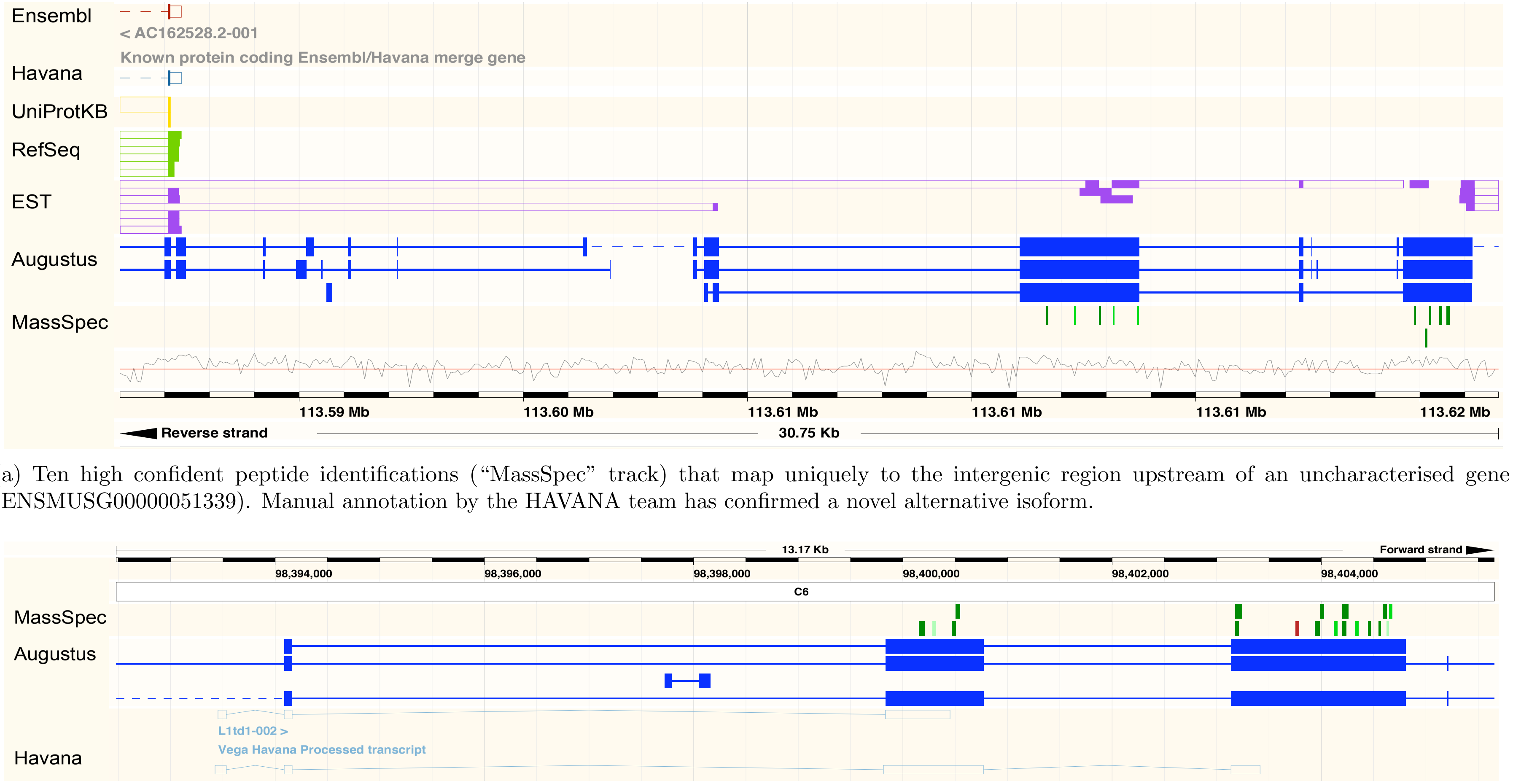
The accurate identification of the UTR and protein coding regions is another challenge in genome annotation. For example, cDNA sequences are often truncated and protein sequences from protein databases are not validated by mass spectrometry, which can lead to wrongly annotated UTRs or protein coding regions.

Data of this pilot study revealed translational evidence either within the UTR or in adjacent intergenic regions for 101 genes, suggesting incorrectly defined coding or gene boundaries. Of the 39 genes that were manually investigated, 85% had additional peptides matching upstream the 5' end.

Figure 5.9 illustrates one example where peptide identifications map uniquely to three exons in the 5' UTR of Ensembl/Vega gene *Asparagine synthetase domain-containing protein 1* (ENSMUSG00000026095), suggesting that either the UTR is incorrectly annotated or an alternative protein isoform exists.

Another example is illustrated in figure 5.10a, where ten peptides map to the intergenic region upstream of an uncharacterised gene (ENSMUSG00000051339). Gary Sounders from the HAVANA team investigated this region manually and built an *ab initio* gene model, which was supported by EST evidence and the ten identified peptides. EST *Em:BY593944.1* fused this novel upstream region with the existing annotation of ENSMUSG00000051339. The translation of an orthologous gene in human showed extensive sequence conservation, further supporting this novel variant.

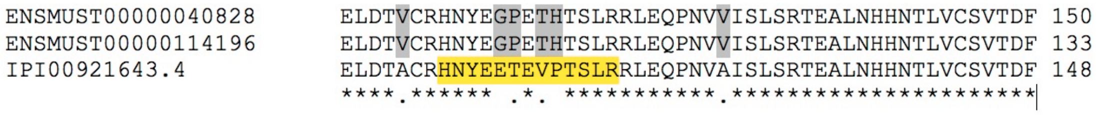
Moreover, pseudo and processed genes in Vega were predicted by Augustus to be protein coding. Strikingly, for 55 of these, translational evidence in the form of peptide identifications was found, suggesting incorrect Ensembl/Vega annotation. Figure 5.10b shows one example where gene *LINE-1 type transposase* (OTTMUST00000019654) was annotated as processed, but a significant number of peptide identifications clearly demonstrated translation. Similar proteogenomic findings of translated pseudogenes were recently demonstrated by Castellana *et al.* (2008) in *Arabidopsis thaliana* and by Merrihew *et al.* (2008) in *C. elegans*.



(a) Ten high confident peptide identifications ("MassSpec" track) that map uniquely to the intergenic region upstream of an uncharacterised gene (ENSMUSG00000051339). Manual annotation by the HAVANA team has confirmed a novel alternative isoform.

(b) The Vega gene OTTMUST00000019654 *(LINE-1 type transposase)* was annotated as a processed transcript without translation ("Havana" track). However, extensive peptide evidence ("MassSpec" track) that is unique to the locus was found, suggesting that there is at least one form of the gene that is translated. The blue tracks represent collapsed Augustus transcripts.

Figure 5.10: Examples of translational peptide evidence in annotated non-coding regions.



```
ENSMUST00000040828      ELDTVCRHNYEGPETHTSLRRLEQPNVVISLSRTEALNHHNTLVCSVTDF 150
ENSMUST00000114196      ELDTVCRHNYEGPETHTSLRRLEQPNVVISLSRTEALNHHNTLVCSVTDF 133
IPI00921643.4           ELDTACRHNYEETEVPTSLRRLEQPNVAISLSRTEALNHHNTLVCSVTDF 148
                        ****.****** .*. ***********.**********************
```

Figure 5.11: One example where the peptide sequence in Ensembl was different to the sequence in IPI (highlighted with grey). The identified peptide (indicated in yellow, $PEP = 1 \times 10^{-4}$) matches perfectly the IPI database but differs in four residues to the Ensembl translations.

5.3.3.3 Protein database derived peptide matches

Genotyping projects over the last number of years have populated large SNP databases, but although these are large, they are yet far from being complete, especially for the mouse genome. Insertion of SNPs into protein databases inflates the search space significantly since multiple variants of one peptide need to be searched, thereby reducing identification sensitivity. In a recent study, Tanner *et al.* (2007) searched a corpus of 18.5 million MS spectra (human) against a database incorporating coding SNPs, resulting in 1.2 million peptide identifications with only 0.02% (308) validated coding SNPs. I have decided to not include coding SNPs into the search databases, but currently there are large scale mouse sequencing efforts underway, potentially allowing strain specific search databases to be built in the future. Evaluation of their performance over a generic species-specific protein databases will be interesting to study.

However, since the search database included IPI protein sequences, some of which were not derived from genomic but from mRNA sequences, differences between IPI and the Ensembl/Vega protein sequences could be detected. 19 IPI proteins with peptide sequences not matching the reference genome were identified. In five cases, the sequence differences were caused by indels, with the remaining 14 cases caused by coding SNPs. Figure 5.11 shows one example where a peptide match was made against the IPI protein sequence, but the Ensembl/Vega reference sequence is different in four residues. This indicates that either the genome reference sequence was incorrect or four novel coding SNPs exist in this peptide.

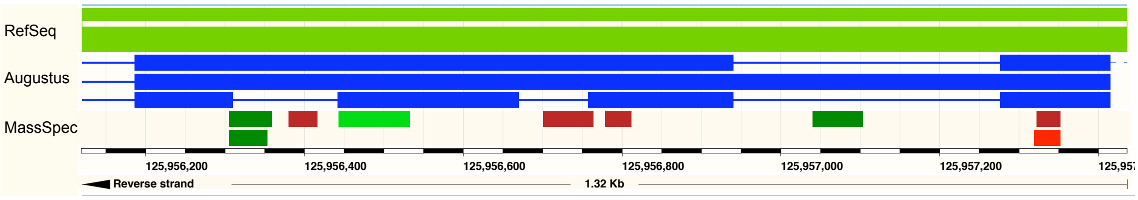


Figure 5.12: Four peptides match uniquely (green) to an intergenic region ("MassSpec" track). Together with full-length mouse cDNA evidence ("RefSeq" track) these data suggests a novel protein coding region.

5.3.3.4 Novel genes

Peptides matching to intergenic regions are of great interest to further complement the list of coding genes. The Ensembl genome annotation process is conservative (see section 1.2.5) and proteogenomic methods are ideally placed to identify such missing genes. The caveat is that the gene of interest must be present in the search database of Mascot to enable identification. As discussed above, the gene finding algorithm Augustus was employed to over-predict the genome. Peptides derived from these predictions and existing annotated protein coding sequences were used as a search database. Assuming that the Ensembl gene list is close to complete, the Augustus predictions contain 90% random sequence (figure 5.1a). Therefore, reliable and stringent peptide scoring together with subsequent filtering to exclude ambiguous matches are crucial to minimize and ultimately to exclude any false positive identifications. To reiterate, the worst peptide match considered in this study had a 1% probability to be incorrect, corresponding to a false discovery rate of less than 0.14%. For subsequent analysis, where peptides were not supported by any existing annotation, this was further constrained in that at least two peptides (one of which with a PEP of at least 0.01, the second of at least 0.05) had to be identified.

Using this approach, I propose 29 novel genes, supported by a total of 70 peptides. However, 12 of these genes have overlapping identifications with IPI protein entries, suggesting that the Ensembl/Vega annotation process missed these genes. The remaining 17 novel genes do not overlap with any known Ensembl/Vega genes or

IPI entries, but six are potentially an extension of known Ensembl/Vega genes, four at the 5' and two at the 3' region. For nine of these genes, Pfam-A domains (Sonnhammer *et al.*, 1997) could be detected with high significance, and based on this it is likely that most of these genes are RNA/DNA binding proteins. Some cDNA or EST evidence was observed for 50% of these novel regions, but vertebrate conservation was generally absent for all of these predictions (figure 5.12).

I have not further investigated an additional set of 50 potential new protein coding genes that were supported by only one peptide. Nevertheless, these peptides are strong matches, with PEP values ranging from 4×10^{-13} to 1×10^{-3}. Even though I am hesitant to identify these regions as novel genes only based on peptide identifications, these predictions together with the proposed new genes are available as a DAS annotation track for the HAVANA annotation team who currently investigate these cases manually and potentially demand additional experimental evidence to complete annotation in Vega.

5.4 Conclusion

Mass spectrometry has become the method of choice to identify peptides and infer proteins in a high-throughput manner and it is therefore a consequent development to incorporate these data into genome annotation pipelines as translational evidence. I have shown that, theoretically, peptide evidence could validate up to 96.5% of all protein coding genes, 91.0% of all protein coding exons and 86.2% of all exon-exon junctions.

However, the mouse proteome is far from being saturated by MS based peptide identifications. Even if every organ with all its regions, cell types and organelles could be isolated and analyzed, there would probably be a significant set of genes that would be missed because expression of these may be only activated under specific and transient cellular activation. There have not been systematic analyses at these levels of complexity, but if I compare studies from ten years ago with the latest achievements, it is clear that MS data becomes richer and more valuable for genome annotation every day.

Using the proteomics datasets readily available for this study, comprising about 10 million spectra, I could validate 31.6% of all protein coding genes, 16.7% of all protein coding exons and 7.1% of all exon-exon junctions in Ensembl, with similar numbers in Vega, significantly contributing to the validation of the mostly automatically annotated mouse genome. Interestingly, for 53 genes I have shown evidence of expression of alternatively spliced isoforms, yet I have also shown that MS data is not always sufficient to fully validate protein isoforms, since many share coding sequence and do not always allow the variants to be distinguished.

It is still not clear of whether transcripts that are flagged to undergo nonsense mediated decay could be translated into stable proteins. I have not detected a single peptide that was unique to NMD transcripts. This could be interpreted in two ways: either these transcripts indeed undergo degradation and cannot be detected or they are translated at very low abundance and were not sampled by the MS instrument for this reason.

Beyond validation, peptide identifications contributed to the identification of potential incorrect annotation. 129 regions were attributed to donor or acceptor splice site refinements or the introduction of novel exons, 101 genomic loci were identified that mapped outside the coding region of genes (mostly at the 5' region) and 55 pseudo- and processed genes were found to be coding. Lastly, I propose 29 protein coding genes, 12 of which are already present in IPI but not in Ensembl/Vega and 6 cases could be coding extensions of known genes.

Overall I have highlighted the possibilities and the limitations of the use of "bottom-up" proteomics for genome annotation and demonstrated the use of available MS data for incorporation into automatic genome annotation pipelines such as Ensembl as an additional layer of evidence.