Published in final edited form as:

Proteomics. 2010 March ; 10(6): 1127–1140. doi:10.1002/pmic.200900445.

Investigating protein isoforms *via* proteomics: a feasibility study

Paul Blakeley, Jennifer A Siepen, Craig Lawless, and Simon J Hubbard¹

Faculty of Life Sciences, Michael Smith Building, University of Manchester, Manchester M13 9PT, UK

Abstract

Alternative splicing (AS) and processing of pre-messenger RNAs 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 alternative splicing exceeds 33% of the alternately spliced genes in the human and worm genomes. Examining AS in chicken for the first time, we find proteomic 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 (10-51% 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.

Keywords

alternate splicing; protein isoforms; peptide identifications

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 [1-8]. 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 [9], Ensembl [10], TrEMBL [11], and RefSeq [12]. 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 [13]. 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 [14]. Sequencing of such libraries can also generate ESTs which provide partial coverage of their

¹Corresponding author. Simon.hubbard@manchester.ac.uk.

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 [15, 16]. 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 [17, 18]. Indeed, the most recent data suggests as many as ~95% of human multi-exon genes undergo AS [19]. The high levels of tissue specificity and functional diversity in higher organisms is thought to be consistent with their higher levels of AS [20]. For example, AS is thought to play a prominent role in the regulation of neuronal development [21, 22], metabolism [23], immunity [24], and may modify protein-protein interaction domains to regulate signalling pathways at key switch points [21, 25]. For example, Kruppel transcription factors contain Kruppel associated box domains which appear to be removed by AS, directly affecting their function [25].

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 [26]. In contrast, some argue that there is little evidence to suggest that AS can generate functional protein isoforms on a large scale [27]. 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* [28]. 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 [29], 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 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. This data is cross-referenced with extant peptide identifications found in the

PeptideAtlas repository [2], as well as a large data set of chicken mass spectra [30]. 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.

Materials and Methods

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 (www.ensembl.org). This was done for the following species; human, mouse, chicken, *Drosophila*, and *C. elegans*. This represents a range of highly annotated eukaryotic species which display varying levels of alternate splicing and 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 paper (for a full description, see [13]), but briefly, Ensembl determines genes using automatic annotation. This involves both computer and biological 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 (www.ebi.ac.uk/integr8).

Peptide identifications

Apart from chicken, all peptide identifications were taken from the PeptideAtlas database (<http://peptideatlas.org/builds/>). PeptideAtlas organises its data into “builds” centred on a particular species or tissue. For this work, we used the human build (www.peptideatlas.org/builds/human/200903/APD_Hs_all.fasta March 2009), containing 104,744 peptides, the Mouse plasma build (www.peptideatlas.org/builds/mouse/plasma/2_007_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 (www.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 [31].

Chicken samples and data analysis

Peptide identifications were derived from a proteomic analysis of the DT40 Chicken cell line [30] 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 (www.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* [32] retaining all peptide identifications at an FDR of 0.01. In total, 14,612 high quality identifications were retained.

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 1.). Exons were considered to be either 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 in to account ionisation and missed cleavage propensity [33].

A full list of peptide numbers at each stage is provided in Table 1.

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 [34] 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

Results and discussion

Proteomic data and alternate 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 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 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 in to 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 [9], 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 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 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 [17, 18] this does represent a consistent set of high quality annotations, ranging from 12-41% of the genome (Table 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 [35-38], the majority of the genome still lacks direct proteomic experimental evidence.

This study has focused on characterising AS events in genes that are derived from differential RNA processing leading to different protein sequences. Figure 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, this 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.

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 [6]. The relative fraction of tryptic peptides that span introns in a digest of the entire proteome is shown in Figure 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 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 [33]. 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 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 [39], supporting the functional diversification of vertebrate proteomes. It also, in part, reflects the level of annotation currently available for the different proteomes.

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 [40]. 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 physio-chemical properties such as hydrophobicity, secondary structure and overall charge and represent peptides likely to be detected in a mass spectrometer [33]. 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 [28]. 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.

Finally we note a further small reduction in the relative fraction of AS-informative peptides when solely experimental data is considered, apart from in the worm. However, the reduction is small and it should not be forgotten that these data were not acquired with the characterisation of AS in mind. 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 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 [41-43] 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 indeed 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.

Examples of alternate splicing with supporting proteomic evidence

Two example cases illustrating the utility of proteomic annotation for AS characterisation are shown in Figures 5 and 6. Figure 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 2 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 4 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 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 2 alternative exons of the ENSGALT00000022043 and ENSGALT00000040695 transcripts (Figure 6). A detailed view of the peptide sequences is supplied in an alignment in Supplementary Figure S1. 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 [44], 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 Figures 5 and 6.

Protein Isoforms in UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot represents a complete and 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. Therefore, we investigated the proportion of Ensembl protein isoforms that are also present in the human component of Swiss-Prot. Table 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 similarly annotated 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 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.

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 only a small proportion (<2%) of identified peptides are unique to one of the two annotation sets. For example, the human TGF β gene (Supplementary Figure S2) is assigned two isoforms by Ensembl48 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 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 the isoforms in instances similar to that shown in Supplementary Figure S2.

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 9 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.

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 [1, 3, 28, 45-48]. 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 (SRM) [49]. This also opens up the possibility of a fully quantitative approach, using SRM (sometimes referred to as MRM) approaches or labelled internal standard peptides (AQUA [50-52] or QconCAT [53, 54]) to quantify individual isoforms. Although these approaches are not without their problems [55-57], a rigorous design strategy and careful experimentation can produce useful results.

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 [58]. We calculated the theoretical digested human proteome for three other enzymes in reasonably common use, as shown in Table 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.

In this study, we have not restricted our calculations and used the so-called “one-hit wonders” [59], where a protein isoform may be characterised by a single peptide identification. This is an inherent problem facing studies of this type, where perhaps 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.

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 [60]. 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.

To fully characterise novel and previously unannotated AS events requires algorithms that can search MS data against the complete genome or predicted genes. Searching against a six-frame translation of an entire genome has the advantage of potentially discovering novel genes and splice forms and is independent of previous (potentially erroneous) annotations. Such methods have been demonstrated to be highly effective in annotating bacterial genomes [4], 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 of seminal fluid genes in *Drosophila*, where false positives were eliminated through the experimental verification of novel gene models [61]. However, without thorough experimental verification, such methods may not be reliable for searching against large eukaryotic genomes because the immense search spaces involved would increase false positive identifications. An alternative strategy is to limit the search space by using *de novo* gene finding algorithms or by searching against putative (predicted) ORFs. This has been demonstrated to be effective in *C. elegans*, where 6350 gene models were identified at the protein level. In addition, 429 novel coding sequences were discovered, including novel intragenic protein-coding regions and genes that had been incorrectly annotated as pseudogenes [8].

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 (www.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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thanks Kathryn Lilley for making DT40 mass spectrometry data available to us, and to BBSRC for funding, via grant BB/F004605/1 to JAS and SJH, and via a studentship to PB.

References

- [1]. Castellana NE, Payne SH, Shen ZX, Stanke M, et al. Discovery and revision of Arabidopsis genes by proteogenomics. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:21034–21038. [PubMed: 19098097]

- [2]. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology*. 2005; 6
- [3]. Gupta N, Benhamida J, Bhargava V, Goodman D, et al. Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Research*. 2008; 18:1133–1142. [PubMed: 18426904]
- [4]. Gupta N, Tanner S, Jaitly N, Adkins JN, et al. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Research*. 2007; 17:1362–1377. [PubMed: 17690205]
- [5]. Savidor A, Donahoo RS, Hurtado-Gonzales O, VerBerkmoes NC, et al. Expressed peptide tags: An additional layer of data for genome annotation. *J. Proteome Res*. 2006; 5:3048–3058. [PubMed: 17081056]
- [6]. Tanner S, Shen ZX, Ng J, Florea L, et al. Improving gene annotation using peptide mass spectrometry. *Genome Research*. 2007; 17:231–239. [PubMed: 17189379]
- [7]. Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, et al. Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*. *Bmc Genomics*. 2009; 10:61. [PubMed: 19193216]
- [8]. Merrihew GE, Davis C, Ewing B, Williams G, et al. Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome Research*. 2008; 18:1660–1669. [PubMed: 18653799]
- [9]. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, et al. The International Protein Index: An integrated database for proteomics experiments. *Proteomics*. 2004; 4:1985–1988. [PubMed: 15221759]
- [10]. Hubbard TJP, Aken BL, Ayling S, Ballester B, et al. Ensembl 2009. *Nucleic Acids Research*. 2009; 37:D690–D697. [PubMed: 19033362]
- [11]. Bairoch A, Consortium U, Bougueleret L, Altairac S, et al. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*. 2009; 37:D169–D174. [PubMed: 18836194]
- [12]. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*. 2007; 35:D61–D65. [PubMed: 17130148]
- [13]. Potter SC, Clarke L, Curwen V, Keenan S, et al. The ensembl analysis pipeline. *Genome Research*. 2004; 14:934–941. [PubMed: 15123589]
- [14]. Okazaki Y, Furuno M, Kasukawa T, Adachi J, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*. 2002; 420:563–573. [PubMed: 12466851]
- [15]. Hubbard SJ, Grafham DV, Beattie KJ, Overton IM, et al. Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. *Genome Research*. 2005; 15:174–183. [PubMed: 15590942]
- [16]. Frith MC, Bailey TL, Kasukawa T, Mignone F, et al. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol*. 2006; 3:40–48. [PubMed: 17114936]
- [17]. Modrek B, Lee C. A genomic view of alternative splicing. *Nat Genet*. 2002; 30:13–19. [PubMed: 11753382]
- [18]. Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*. 2003; 34:177–180. [PubMed: 12730695]
- [19]. Pan Q, Shai O, Lee LJ, Frey J, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*. 2008; 40:1413–1415. [PubMed: 18978789]
- [20]. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*. 2001; 17:100–107. [PubMed: 11173120]
- [21]. Jiao YX, Robison AJ, Bass MA, Colbran RJ. Developmentally regulated alternative splicing of densin modulates protein-protein interaction and subcellular localization. *Journal of Neurochemistry*. 2008; 105:1746–1760. [PubMed: 18248607]

- [22]. Boutz PL, Stoilov P, Li Q, Lin CH, et al. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & Development*. 2007; 21:1636–1652. [PubMed: 17606642]
- [23]. Passetti F, Ferreira CG, Costa FF. The impact of microRNAs and alternative splicing in pharmacogenomics. *Pharmacogenomics Journal*. 2009; 9:1–13. [PubMed: 19156160]
- [24]. Zikherman J, Weiss A. Alternative Splicing of CD45: The Tip of the Iceberg. *Immunity*. 2008; 29:839–841. [PubMed: 19100695]
- [25]. Resch A, Xing Y, Modrek B, Gorlick M, et al. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome Res*. 2004; 3:76–83. [PubMed: 14998166]
- [26]. Birzele F, Csaba G, Zimmer R. Alternative splicing and protein structure evolution. *Nucleic Acids Research*. 2008; 36:550–558. [PubMed: 18055499]
- [27]. Tress ML, Martelli PL, Frankish A, Reeves GA, et al. The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104:5495–5500. [PubMed: 17372197]
- [28]. Tress ML, Bodenmiller B, Aebersold R, Valencia A. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology*. 2008; 9
- [29]. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, et al. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *Comptes Rendus Biologies*. 2005; 328:882–899. [PubMed: 16286078]
- [30]. Hall SL, Hester S, Griffin JL, Lilley KS, Jackson AP. The organelle proteome of the DT40 lymphocyte cell line. *Mol Cell Proteomics*. 2009; 8:1295–1305. [PubMed: 19181659]
- [31]. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol*. 2005; 1 2005 0017.
- [32]. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 2008; 7:29–34. [PubMed: 18067246]
- [33]. Mallick P, Schirle M, Chen SS, Flory MR, et al. eComputational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*. 2007; 25:125–131.
- [34]. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 1990; 215:403–410. [PubMed: 2231712]
- [35]. Mann K. The chicken egg white proteome. *Proteomics*. 2007; 7:3558–3568. [PubMed: 17722208]
- [36]. Mizukami M, Kanamoto T, Souchelnytskyi N, Kiuchi Y. Proteome profiling of embryo chick retina. *Proteome Science*. 2008; 6
- [37]. Buza TJ, McCarthy FM, Burgess SC. Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *Bmc Genomics*. 2007; 8
- [38]. Zheng XJ, Hong LL, Shi LX, Guo JQ, et al. Proteomics analysis of host cells infected with infectious bursal disease virus. *Molecular & Cellular Proteomics*. 2008; 7:612–625. [PubMed: 18056921]
- [39]. Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*. 2007; 35:125–131. [PubMed: 17158149]
- [40]. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100:189–192. [PubMed: 12502788]
- [41]. Jones P, Cote RG, Cho SY, Klie S, et al. PRIDE: new developments and new datasets. *Nucleic Acids Res*. 2008; 36:D878–883. [PubMed: 18033805]
- [42]. Mathivanan S, Pandey A. Human Proteinpedia as a resource for clinical proteomics. *Mol Cell Proteomics*. 2008; 7:2038–2047. [PubMed: 18573810]
- [43]. Mead JA, Bianco L, Bessant C. Recent developments in public proteomic MS repositories and pipelines. *Proteomics*. 2009; 9:861–881. [PubMed: 19212957]

- [44]. Marden JH. Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity*. 2008; 100:111–120. [PubMed: 17006532]
- [45]. Loevenich SN, Brunner E, King NL, Deutsch EW, et al. The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. *BMC Bioinformatics*. 2009; 10:59. [PubMed: 19210778]
- [46]. Menon R, Zhang Q, Zhang Y, Fermin D, et al. Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res*. 2009; 69:300–309. [PubMed: 19118015]
- [47]. Power KA, McRedmond JP, de Stefani A, Gallagher WM, Gaora PO. High-throughput proteomics detection of novel splice isoforms in human platelets. *PLoS One*. 2009; 4:e5001. [PubMed: 19308253]
- [48]. Schrimpf SP, Weiss M, Reiter L, Ahrens CH, et al. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol*. 2009; 7:e48. [PubMed: 19260763]
- [49]. Schiess R, Wollscheid B, Aebersold R. Targeted proteomic strategy for clinical biomarker discovery. *Mol Oncol*. 2009; 3:33–44. [PubMed: 19383365]
- [50]. Gerber SA, Kettenbach AN, Rush J, Gygi SP. The absolute quantification strategy: application to phosphorylation profiling of human separase serine 1126. *Methods Mol Biol*. 2007; 359:71–86. [PubMed: 17484111]
- [51]. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A*. 2003; 100:6940–6945. [PubMed: 12771378]
- [52]. Kirkpatrick DS, Gerber SA, Gygi SP. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods*. 2005; 35:265–273. [PubMed: 15722223]
- [53]. Pratt JM, Simpson DM, Doherty MK, Rivers J, et al. Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat Protoc*. 2006; 1:1029–1043. [PubMed: 17406340]
- [54]. Rivers J, Simpson DM, Robertson DH, Gaskell SJ, Beynon RJ. Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. *Mol Cell Proteomics*. 2007; 6:1416–1427. [PubMed: 17510050]
- [55]. Duncan MW, Yergey AL, Patterson SD. Quantifying proteins by mass spectrometry: the selectivity of SRM is only part of the problem. *Proteomics*. 2009; 9:1124–1127. [PubMed: 19253279]
- [56]. Pan S, Aebersold R, Chen R, Rush J, et al. Mass spectrometry based targeted protein quantification: methods and applications. *J Proteome Res*. 2009; 8:787–797. [PubMed: 19105742]
- [57]. Sherman J, McKay MJ, Ashman K, Molloy MP. How specific is my SRM?: The issue of precursor and product ion redundancy. *Proteomics*. 2009; 9:1120–1123. [PubMed: 19253278]
- [58]. de Godoy LM, Olsen JV, Cox J, Nielsen ML, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*. 2008; 455:1251–1254. [PubMed: 18820680]
- [59]. Veenstra TD, Conrads TP, Issaq HJ. What to do with “one-hit wonders”? *Electrophoresis*. 2004; 25:1278–1279. [PubMed: 15174049]
- [60]. Bell AW, Deutsch EW, Au CE, Kearney RE, et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat Methods*. 2009
- [61]. Findlay GD, MacCoss MJ, Swanson WJ. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Research*. 2009; 19:886–896. [PubMed: 19411605]

Figure 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 they 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. Most peptides that cover unique exons are specific to a particular isoform.

Figure 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 from those that don't. 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 1), which is broadly in line with the percentage of exons which are not constitutive.

Figure3. 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 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.

Figure 5. Peptide evidence for alternative protein isoforms encoded by a hypothetical gene
The gene model FAM136A is predicted to encode 2 different isoforms. Ensembl transcripts (top) and enlarged exons (bottom) are shown in grey. Mapped tryptic peptides are shown as black 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 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 grey. Mapped tryptic peptides are shown as black 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.

sp P56181 NDUV3_HUMAN	MAAPCLLRQGRAGALKTMLQEAQVFRGLASTVSLSAESGKSEKGPQNSK	50
ENSP00000342895	MAAPCLLRQGRAGALKTMLQEAQVFRGLASTVSLSAESGKSEKGPQNSK	50
ENSP00000346196	MAAPCLLRQGRAGALKTMLQEAQVFRGLASTVSLSAESGKSEKGPQNSK	50

sp P56181 NDUV3_HUMAN	KQSPPK-----	56
ENSP00000342895	KQSPPK-----	56
ENSP00000346196	KQSPPKNVVEPKERKLLATQTAAELSKNLSSPSSYPVAVNKGRK VASPS	100

sp P56181 NDUV3_HUMAN	-----	
ENSP00000342895	-----	
ENSP00000346196	PSGSVLFTDEGVPK FLSRKTLVEFPQKVLSPFRKQGSSEARQVGRK VTS	150
	1	
sp P56181 NDUV3_HUMAN	-----	
ENSP00000342895	-----	
ENSP00000346196	PSSSSSSSSDSESDDEADVSEVTPR VVSKGRGGLRKPEASHSFENRAPR	200
	2, 3	
sp P56181 NDUV3_HUMAN	-----	
ENSP00000342895	-----	
ENSP00000346196	VTVSAKEKTLQKPHVDITDPEKPHQPKKKGSPAKPSEGRENARPKTTMP	250
sp P56181 NDUV3_HUMAN	-----	
ENSP00000342895	-----	
ENSP00000346196	RSQVDEEFLKQSLKEKQLQKTRFLNEIDKESQKPFVKGPLPVHTK SGLS	300
sp P56181 NDUV3_HUMAN	-----	
ENSP00000342895	-----	
ENSP00000346196	APPKGPAPAVLAEEAR AEGQLQASPPGAAEGHLEKPVPEPQR KAAPPLP	350
	4 5	
sp P56181 NDUV3_HUMAN	-----	
ENSP00000342895	-----	
ENSP00000346196	RKETSGTQGIIEGHLKGGQAIVEDQIPPSNLETVPVENNHGFHEK TAALKL	400
	6 7 10	
sp P56181 NDUV3_HUMAN	----- KPAPVPAEPFDNTTYK NLQHHDYSTYFL	85
ENSP00000342895	----- KPAPVPAEPFDNTTYK NLQHHDYSTYFL	85
ENSP00000346196	EAEGEAMEDAAAPGDDRGGTQEPAPVPAEPFDNTTYK NLQHHDYSTYFL	450
	8 : ***** 9	
sp P56181 NDUV3_HUMAN	DLNLELSKFRMPQPSSGRESRPH	108
ENSP00000342895	DLNLELSKFRMPQPSSGRESRPH	108
ENSP00000346196	DLNLELSKFRMPQPSSGRESRPH	473

Figure 7. Evidence for an additional isoform in the human NDUV3 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 segment in ENSP00000342895 that is missing in ENSP00000346196 and in the equivalent Swiss-Prot entry (P56181). A total of 9 peptides (bold regions) were unambiguously mapped to ENSP00000346196, thus confirming an exon-skipping event that has not been annotated in Swiss-Prot. The mitochondrial targeting sequence, common to all predicted isoforms, is shown as a grey box at the beginning of the sequence.

Table 1

Peptide data sets calculated or used in this study

	Tryptic Peptide subsets (including upto 1 missed cleavage)			
Ensembl species	all unambiguously mapped peptides	length filtered peptides	Proteotypic peptides	Experimentally determined, mapped peptides
human	1571922	1400266	694492	61871
mouse	1156002	810503	474869	9988
chicken	1243182	1109471	530106	13546
<i>D.melanogaster</i>	1105525	681090	325441	3845
<i>C.elegans</i>	921281	868143	233138	12820

Table 2

Protein isoforms in Ensembl 48

a) Total genes				
organism	protein coding genes	proteins	average number of isoforms per gene	
human	22,839	41,021	1.80	
mouse	23,429	35,628	1.52	
chicken	16,736	21,871	1.31	
<i>D.melanogaster</i>	14,039	17,104	1.22	
<i>C.elegans</i>	20,140	23,334	1.16	

b) Genes encoding more than one non-redundant protein isoform				
organism	protein coding genes	proteins	average number of isoforms per gene	% of genes with >1 isoform
human	9,343	27,903	2.99	41
mouse	7,280	19,876	2.73	31
chicken	3,776	8,990	2.38	23
<i>D.melanogaster</i>	2,079	5,276	2.54	15
<i>C.elegans</i>	2,337	5,706	2.44	12

Table 3

Relative fractions of gene sets with associated peptide evidence

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

^a only high quality peptide identifications from PeptideAtlas or [30]

^b only high quality peptide identifications from PeptideAtlas or [30] 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.

Table 4

Comparative analysis of different proteome annotations for alternate splicing and proteomic peptide evidence

	HPI Swiss-Prot		Ensembl48	
Genes annotated with multiple non-redundant protein isoforms	7,439 ^a		9,343	
	Uniquely multi-isoform	Shared multi-isoform	Uniquely multi-isoform	Shared multi-isoform
	1345	5671 ^b	3717	5626 ^b
Total number of associated protein isoforms	3152	17358	10429	20514
Total number of mapped peptides	4281	23939	16607	24181
Number of mapped “unique” peptides (present in only a single isoform)	927	2853	3605	2531
Number of genes with “unique” isoform-specific peptides mapped	271	911	746	818

^aSeveral 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.

^bA few Swiss-Prot genes were equivalenced to a single Ensembl gene, and *vice versa*, and hence the shared gene total is not identical in the two annotations.

Table 5

Suitability for alternative peptidases for AS characterisation in humans

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