HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2016 February 26.

Published in final edited form as:

J Proteome Res. 2015 April 3; 14(4): 1880–1887. doi:10.1021/pr501286b.

Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform

Iakes Ezkurdia[†], **Jose Manuel Rodriguez**[§], **Enrique Carrillo-de Santa Pau**[||], **Jesús Vázquez**[‡], **Alfonso Valencia**[*,§,||], and **Michael L. Tress**[*,||]

[†]Unidad de Proteómica, Centro Nacional de Investigaciones Cardiovasculares, 28029 Madrid, Spain

[‡]Laboratorio de Proteómica Cardiovascular, Centro Nacional de Investigaciones Cardiovasculares, 28029 Madrid, Spain

[§]National Bioinformatics Institute, Spanish National Cancer Research Centre, 28029 Madrid, Spain

[||]Structural Biology and Bioinformatics Programme, Spanish National Cancer Research Centre, 28029 Madrid, Spain

Abstract

Although eukaryotic cells express a wide range of alternatively spliced transcripts, it is not clear whether genes tend to express a range of transcripts simultaneously across cells, or produce dominant isoforms in a manner that is either tissue-specific or regardless of tissue. To date, large-scale investigations into the pattern of transcript expression across distinct tissues have produced contradictory results. Here, we attempt to determine whether genes express a dominant splice variant at the protein level. We interrogate peptides from eight large-scale human proteomics experiments and databases and find that there is a single dominant protein isoform, irrespective of tissue or cell type, for the vast majority of the protein-coding genes in these experiments, in partial agreement with the conclusions from the most recent large-scale RNAseq study. Remarkably, the dominant isoforms from the experimental proteomics analyses coincided overwhelmingly with the reference isoforms selected by two completely orthogonal sources, the consensus coding sequence variants, which are agreed upon by separate manual genome curation teams, and the principal isoforms from the APPRIS database, predicted automatically from the conservation of protein sequence, structure, and function.

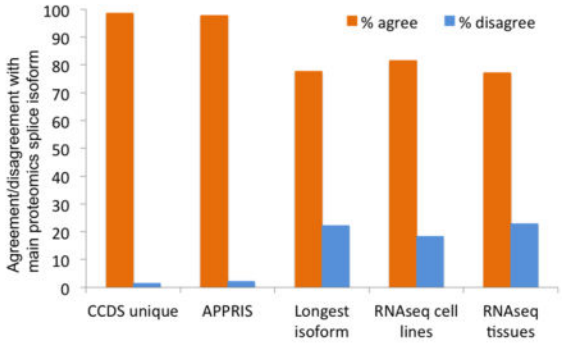
[*]Corresponding Authors: mtress@cnio.es. Tel: +34 91 732 80 00 (x3017). Fax: +34 91 224 69 76. Valencia@cnio.es.

Notes

The authors declare no competing financial interest.

Supporting Information

Data from individual analyses and graph of the number of peptides used to select main proteomics isoforms. This material is available free of charge via the Internet at http://pubs.acs.org.



Keywords

Large-scale proteomics; RNAseq; Alternative splicing; Dominant isoforms; Protein structure; Protein function

INTRODUCTION

Alternative splicing of mRNA can generate a wide range of mature RNA transcripts. It has been estimated that alternative splicing of pre-mRNA occurs in 95% of multiexon human genes.[1,2] EST and cDNA sequence evidence[3] and microarray data[4] strongly support the expression of multiple alternative mRNA transcripts from the same gene, and manual genome annotation projects are confirming ever more alternative splice variants: the GENCODE 20[5] human gene set has slightly more than 20 000 protein-coding genes but over 93 000 coding transcripts. As long as they are translated into stable proteins,[6] the 73 000 alternative transcripts have the potential to expand the cellular protein repertoire.[7]

Although there is abundant evidence for the expression of multiple transcripts in cells, it is less clear whether these transcripts are expressed more or less equally across tissues or whether it would be biologically relevant to designate one transcript per gene as dominant and the rest as alternative. The question of whether genes have dominant transcripts is one that is becoming ever more important with the growth in the numbers of alternative transcripts annotated in the databases.

Three contrasting large-scale studies came to different conclusions. An EST-based study with 13 different tissues[8] predicted that primary tissues generally had a single dominant transcript per gene. By way of contrast, a large-scale study using RNAseq[9] found that more than three-quarters of protein-coding genes had cell-line-specific dominant transcripts and that those genes with the most splice variants had more dominant transcripts. Most recently, a second study of RNAseq data from the Illumina Human BodyMap project found that approximately half the genes expressed in the 16 tissues studied had the same major transcript in all tissues,[10] whereas another third of the genes had major transcripts that were tissue-dependent. One curious result was that the major transcript was noncoding in close to 20% of the protein-coding genes.

To date, no equivalent study has been carried out at the protein level. Indeed, the extent to which the expression of alternative transcripts affects cellular protein diversity is still open to question. Peptide evidence from MS/MS proteomics experiments has been used to demonstrate the expression of alternative protein isoforms, but no reliable study has identified more than 150 pairs of alternative isoforms.[6]

At least part of the reason for the low numbers of alternative isoforms detected in these studies is the relatively low coverage of peptides from proteomics experiments; mass spectrometry only identifies a fraction of the peptide ions in protease digests.[11] However, many experiments identify many fewer alternative isoforms than would be expected, even if the low peptide coverage is taken into account.[6,12–14]

Ning and Nesvizhskii[15] carried out a study on the feasibility of combining data from RNAseq and proteomics to search for alternative isoforms. Although they found a relationship between RNAseq and proteomics expression at the gene level, the fraction of alternative isoforms identified from the corresponding RNAseq data was "substantially lower than the number expected." The results from all these proteomics studies suggest that alternative isoforms may be expressed infrequently, in very few tissues, or have very short half-lives.

Here we have attempted to determine whether protein-coding genes have dominant isoforms using reliable peptide evidence from eight separate large-scale MS analyses. We mapped at least two peptides to 63.9% of the human gene set and identified alternative splice isoforms for just 246 human genes; this clearly suggests that the vast majority of genes express a single main protein isoform. These main experimental isoforms found strong support from both consensus coding sequence (CCDS) transcripts[16] and APPRIS principal isoforms, [17] whereas the agreement with dominant transcripts from RNAseq data was less clear.

PROCEDURES

We collected the peptides for the analysis from eight distinct large-scale proteomics data sets: the PeptideAtlas[18] and NIST (http://peptide.nist.gov/) databases and six published large-scale experiments.[6,19–23] For the Wilhelm analysis,[23] we only included the peptides from the publically available Cellzome experiments on human tissues.

The eight studies covered a huge range of tissues and cell types: the peptides from the PeptideAtlas database cover 51 different tissues, cell types, and developmental stages, whereas the Geiger[21] study interrogated 11 different cell types. The spectra from the PeptideAtlas database were only part of the NIST database and the Ezkurdia[6] analyses. The Kim[22] and Wilhelm analyses peptides were generated from 30 and 35 distinct tissues types (51 tissues in total).

FILTERING LOW-RELIABILITY PEPTIDE IDENTIFICATIONS

The peptides from the eight data sets were subjected to a series of rigorous filters to remove as many likely false-positive peptides as possible from all analyses. First, we eliminated all nontryptic and semitryptic peptides; only peptides that were fully tryptic for at least one

annotated isoform were included. Fully tryptic peptides included C- and N-terminal peptides and N-terminal peptides with the initial methionine cleaved. We only allowed missed cleavages when they were supported by the presence of one the cleaved tryptic subpeptides. The same rules were applied to the peptides cleaved by the enzyme LysC in the Nagaraj[20] and Wilhelm analyses. Search engines cannot easily distinguish leucine from isoleucine; so for the purposes of our analysis, leucine and isoleucine were considered indistinguishable from each other. Peptides that mapped to more than one gene were not included in the experiment.

Multiple search engines can be used separately to increase coverage[22–24] or combined to improve reliability.[25] To be more rigorous, we only included peptides identified by the intersection of two search engines where possible. In practice, this meant that we excluded peptides identified by just one of the five search engines included in the NIST database, those peptides identified by just one of the two search engines in the Kim analysis, and those peptide-spectra matches with an Andromeda[26] score of less than 100 in the Nagaraj, Geiger, and Wilhelm analyses. Excluding peptides below this score improves the false-positive rate because peptides identified by Andromeda with scores of less than 100 are not always in agreement with those identified by Mascot[27] for the same spectra.[26]

The peptides from the Ezkurdia analysis had a peptide FDR of 0.1%, whereas the PeptideAtlas peptides have a peptide-spectrum match (PSM) FDR of 0.0002%. The peptides from the Munoz analysis had a peptide FDR of 1%.

Even after the filtering we carried out for this analysis, peptides identified just once over eight such large-scale experiments have a much higher probability of being false-positive spectra matches, so we only considered peptides that were identified in two or more of the eight data sets. The main isoforms were determined by counting the number of experimental peptides that mapped to each isoform. A single peptide was enough to discriminate between isoforms in a gene.

We identified a total of 149 954 highly reliable gene discriminating peptides. The peptides were mapped to the protein isoforms annotated in the GENCODE 20 human gene set. The number of gene-discriminating peptides that mapped to each gene in the GENCODE 20 set is shown in Figure S1 (Supporting Information). The manual GENCODE annotations are highly enriched in alternative isoforms; the gene set has a mean of four protein-distinct isoforms per gene. GENCODE 20 was filtered for pseudoautosomal genes and for read-through transcripts. The GENCODE 20 gene set we used is annotated with 19 906 protein-coding genes that can produce 83 229 sequence-distinct protein isoforms. A total of 15 548 genes are annotated with more than one splice isoform.

The numbers of peptides, genes, and dominant transcripts identified from each of the experiments after filtering is shown in Table 1 of the Supporting Information.

PRINCIPAL SPLICE ISOFORMS

The APPRIS database determines principal splice isoforms on the basis of the conservation of protein features, including protein structural and functional data and information from

cross-species conservation. Splice isoforms in APPRIS are annotated with protein structural information via mapping to structural homologues and with functional information from the conserved, functionally important amino acid residues predicted by *firestar*[28] and from Pfam functional domains.[29] The conservation information used to select principal isoforms comes from protein alignments between vertebrate orthologues. The highest-scoring isoform from the analyses is chosen as the principal splice isoform.

The features that determine the principal isoform are useful for discriminating between splice isoforms because they have a high degree of conservation. For example, protein structural and functional domains and motifs have evolved over huge evolutionary timeframes, so isoforms that have lost these features are not likely to be the principal isoform.

RNASEQ COUNTS

The in-house RNAseq analysis was carried out with CD14-positive, CD16-negative classical monocyte sample C000S5B1 from the BLUEPRINT consortium. We used the alignments generated for release November 8, 2014, as aligned by the consortium.[30]

We removed reads with a Phred score lower than 15 as well as duplicate reads and used the intersectBed tool included in bedtools, version 2.18.2,[31] with parameter c. We counted the number of reads that mapped on transcripts annotated with a coding sequence (CDS); for this purpose, we used the annotation file from GENCODE19.[5] We only counted reads that mapped to CDS exons. The transcript with the most reads was determined to be the major variant.

RESULTS AND DISCUSSION

We collected peptides from eight large proteomics data sets and subjected them to a series of rigorous filters as detailed in the section above. After filtering, there were 149 954 highly reliable peptides, 111 382 of which discriminated between isoforms from the same gene. We mapped these peptides to the annotations in the GENCODE 20 gene set (equivalent to Ensembl 76).[32] We detected at least two peptides for 12 716 (63.9%) of the protein-coding genes but found alternative protein isoforms for just 246 genes (1.2%), which meant that the vast majority of genes had peptide evidence for just one protein isoform. This is in line with findings from similar proteomics experiments.[6,13,14]

Peptides do not provide complete sequence coverage (even when they do, they are not evenly distributed across the sequences), so counting abundances in a way similar to RNAseq reconstruction methods is inappropriate. Instead, we determined a main proteomics isoform by counting the total number of peptides that mapped to each splice isoform annotated for a gene. The isoform with the highest number of peptides was the main proteomics isoform. In this way, we could identify a unique main proteomics isoform for 5011 genes. Of the remaining 7705 identified genes, 3977 were annotated with a single protein coding isoform, 3703 had too few isoform discriminating peptides to identify isoforms, and just 25 had evidence of alternative splicing where the main isoform could not

be distinguished because the total number of peptides was the same for the two splice isoforms.

HOW DOES THE MAIN PROTEOMICS ISOFORM COMPARE WITH OTHER REFERENCE ISOFORMS?

The number of peptides that map to each isoform is a particularly simple estimation of cellular dominance, one that can only work because we detect few alternative isoforms. To determine whether these counts had a biological reference, we investigated the relationship between the main proteomics isoform and three different reference isoforms: the isoform with the longest sequence, the unique CCDS variants, and the APPRIS principal isoforms. The results of all the comparisons are shown in Table 1 and the results for the individual experiments are shown in supplementary Table 1.

CCDS variants are based on genomic evidence and are variants that are mutually agreed on by teams of manual annotators from the National Center for Biotechnology Information, the Wellcome Trust Sanger Institute, the European Bioinformatics Institute and the University of California Santa Cruz. A total of 13 297 GENCODE 20 genes were annotated with a single CCDS variant. This unique manually curated variant agreed with the main proteomics isoform for a remarkable 98.6% of the 3331 genes that we compared.

APPRIS annotates principal isoforms on the basis of conservation of structure and function and selected a main isoform for 15 172 of the GENCODE 20 coding genes. We were able to compare the APPRIS principal isoforms and the main proteomics isoforms over 4186 genes and found that the main proteomics isoform agreed with the isoform with the most conserved protein features for 97.8% of these genes (4093).

The longest sequence is the method of choice for selecting a representative variant for publicly available databases (except Ensembl) and for practically all large-scale experiments. Here the agreement between the longest isoform and the main proteomics isoform will be high if the peptides we find are randomly distributed among the isoforms. We compared the longest isoform to the main proteomics isoform over all 5011 genes with a defined main proteomics isoform. The longest isoform coincided with the main proteomics isoform for 89.6% of the genes. The agreement between the main proteomics isoform and the reference isoforms selected by the other two methods was substantially higher than with the longest isoform.

We further investigated the agreement between the APPRIS, CCDS, and main experimental variants, looking at those 3015 genes where all three data sets had a single dominant isoform. Here, the CCDS variant and the main proteomics isoform were in agreement for 99.37% of the 3015 genes, whereas the agreement between the main proteomics isoform and the APPRIS principal isoform was 99.5% over the same genes.

For those few genes where there was disagreement between the main isoforms, several disagreements were due incomplete gene models. One example is the gene *KIAA1468*, which codes for a LisH domain and HEAT-repeat-containing protein. This gene is annotated

with four variants in the GENCODE human annotation, two of which are of the most interest: KIAA1468-001 and KIAA1468-002 (Figure 1). The transcripts differ in two regions: KIAA1468-002 has an inserted exon (exon 18), whereas the two transcripts also possess a pair of mutually exclusive homologous exons (exon 21a in KIAA1468-001 and exon 21b in KIAA1468-002). CCDS and APPRIS select KIAA1468-001 as their main isoform, whereas more peptides map to KIAA1468-002. However, the reason that more peptides map to KIAA1468-002 is that there is more peptide evidence for exon 21b than for exon 21a; there were no peptides for exon 18. APPRIS favors KIAA1468-001 because KIAA1468-002 has no evidence in the protein databases in contrast to KIAA1468-001, which is conserved back to *Danio* (Figure 1b), whereas the inserted exon 18 would break the likely 3D structure (a HEAT-repeat solenoid, Figure 1c). The mutually exclusive homologous exons 21a and 21b are both found in teleosts, meaning that the splicing event occurred over 400 million years ago. The conservation of these two exons strongly suggests a functional relevance. Exon 18 is only conserved in macaque. This evidence suggests that the gene model for KIAA1468 is missing a variant that has conserved exon 21b but does not have inserted exon 18 (Figure 1).

The close agreement between these three orthogonal methods of selecting a reference isoform, proteomics, APPRIS, and CCDS, clearly demonstrates that the dominant isoforms from all three methods are highly reliable and reflect the biological reality for most genes.

PROTEOMICS AND RNASEQ AGREEMENT

We also looked at the agreement between the main proteomics isoform and the dominant isoforms form the most recent RNAseq analysis, the Human BodyMap study.[10] Here, the authors found that 4199 genes had a major dominant transcript (defined as a transcript with at least 5-fold higher abundance than any other) that recurred across cell lines, whereas 5228 genes had 5-fold dominant transcripts that recurred across all tissues. The authors carried out a comparison with the APPRIS principal isoform as part of the paper and found that the approximately 5000 genes with 5-fold abundance agreed with the APPRIS principal isoforms in over 60% of the genes, whereas those with 2-fold dominance agreed with the APPRIS principal isoform only 45.6% of the time.

We compared the main proteomics isoform from our study with the 5-fold dominant RNA variants from the BodyMap study. These are the transcripts that had 5-fold dominance across cell lines or tissues, i.e., the dominant transcripts with the clearest evidence. We could only compare over a limited set of genes because the Gonzalez-Porta analysis was carried out with annotations from Ensembl 66 (GENCODE 11). We excluded all genes and transcripts with changed identifiers and those genes where the RNAseq data had chosen a noncoding transcript as the 5-fold dominant variant. The main proteomics isoform agreed with the tissue-recurring 5-fold dominant transcripts from the Human BodyMap study for 77.2% of the 1038 genes that could be compared, whereas the agreement with the cell-line-recurring 5-fold dominant transcripts was 81.6% over 762 comparable genes (Table 1).

The agreement between RNAseq 5-fold dominant transcripts and the main proteomics isoform was better than the agreement between the 5-fold dominant transcripts and the

APPRIS principal isoforms that was reported by the authors of the Human BodyMap study but still was some way short of the agreement between the main proteomics isoform and the APPRIS and CCDS reference isoforms.

It is somewhat surprising that the agreement between the main proteomics isoform and the 5-fold dominant RNAseq transcripts is considerably worse than the agreement with APPRIS principal isoforms and CCDS unique variants. One reason may be that there are substantial differences between what is expressed at the transcript level and what is expressed at the protein level. The Human BodyMap study[10] found that 20% of the 5-fold dominant transcripts were noncoding transcripts and surmised that a substantial portion of the transcripts expressed are never destined for translation and may have a function related to gene expression. Other groups have suggested that discrepancies between proteomics and RNAseq data[14,15] could be due to factors such as mRNA and protein turnover rates, post-transcriptional editing,[33] or translational efficiency.[34]

In this study, there seems to be a clear difference in the length of the main variants determined from the two experimental methods. The main proteomics isoform agrees with the longest isoform in almost 90% of comparable genes, whereas the Gonzalez-Porta study[10] reported that the dominant transcripts only agreed with longest CDS in 50% of the genes. This squares with our results: in 27 of the 36 genes where neither the 5-fold dominant transcript nor the longest isoform agrees with the main proteomics isoform, the 5-fold dominant variant does not coincide with the longest variant.

As an example, the gene *CRIP2*, cysteine-rich protein 2, has been well studied.[35] The structure of the first domain has been solved (entry 2CU8 in the Protein Database,[36]) and the second domain is a duplication of the first. The main proteomics isoform also has support all the way back to *Danio rerio* in the protein databases. Both the longest isoform and the 5-fold dominant variant would break the 3D structure of the CRIP2 zinc-binding domain (Figure 2). The isoform from the 5-fold dominant transcript is considerably shorter than the main proteomics isoform.

The gene *PSMD13*, a regulatory subunit of the 26S proteasome[37] that is involved in the degradation of ubiquitinated proteins, has nine coding transcripts. In this gene, the 5-fold dominant transcript selected by the RNAseq study is again much shorter and is tagged as a nonsense-mediated decay variant (Figure 2). Although the 3D structure of the *PSMD13* gene product has not been solved, similar structures exist for the whole protein. The main proteomics isoform maps well to the structure; the 5-fold dominant transcript, if translated, would only have 72 residues and would break the 3D structure. The 5-fold dominant transcript does not have support in the protein databases, whereas the main proteomics isoform is supported all the way back to the first eukaryotes because homologues of PSMD13-001 can be found in both fungi and plants.

To investigate why isoforms from 5-fold dominant transcripts tended to be shorter than the main proteomics isoform, we looked at raw RNAseq reads from the BLUEPRINT project. We mapped the raw reads to genes from the GENCODE 19 gene set and used these reads to select a dominant transcript in the same way as what we had used to calculate the main

proteomics isoform, i.e., we determined the dominant transcript to be the transcript that had the highest total number of reads that mapped to the CDS. The agreement with the main proteomics isoform is shown in Table 1.

Without any filtering of the RNAseq data, the agreement between the dominant transcript calculated from raw reads and the main proteomics isoform was 85.8%. This agreement is already better than that between the proteomics main isoform and 5-fold dominant transcripts taken from the Human BodyMap study, where only 5-fold major dominant transcripts were compared. When we looked at those genes that had one transcript with twice as many reads as the next (1018 genes), the agreement between proteomics and RNAseq major variants was 95%. This indicates that RNAseq reads do indeed contain a signal that can be used to select the main isoform.

CONCLUSIONS

We mapped highly reliable peptides from eight large-scale proteomics studies to 64% of protein coding genes. For the vast majority of these genes, the peptides mapped to a single splice isoform, and we could distinguish a main protein isoform from the peptide data for 5011 genes. The main isoform from these proteomics experiments was simply the isoform with the most mapped peptides. This simple method for selecting the main isoform works well because large-scale proteomics experiments identify very few alternative protein isoforms when false-positive matches are rigorously filtered.

This main isoform agreed with reference isoforms from two very different sources, the principal isoforms from the APPRIS database and the unique CCDS variants. Indeed, for those genes where all three sources selected a main variant, the agreement between the APPRIS principal isoform, the unique CCDS variant, and the main proteomics isoform was almost perfect (over 99%).

The main proteomics isoform was determined from an extensive set of rigorously filtered experimental peptides, CCDS variants are chosen from genomic evidence by agreement between separate manual curation teams, and APPRIS principal isoforms are generated automatically using the conservation of protein features. The clear agreement between three orthogonal sources significantly reinforces the probability that the main proteomics isoform is the dominant protein isoform in the cell.

The agreement between the main proteomics isoforms and APPRIS principal isoforms demonstrates that the cellular machinery tends to express the most conserved splice isoform and the one that best preserves the conserved structural and functional features of the protein. This opens the door to using computational predictions for reference isoforms for all protein coding genes. APPRIS currently houses annotations for six species but could be extended to predict principal isoforms for more genomes. This would enable the protein isoforms that are most likely to be the main cellular isoform to be prioritized in experiments and large-scale data analyses.

The discrepancy between the carefully selected RNAseq major dominant transcripts and proteomics main isoforms is interesting. Even though we made the comparison only with

coding transcripts that had 5-fold dominance across cell lines or tissues, the agreement did not exceed 82%. The dominant transcripts from the raw RNAseq reads appear to agree more often with the main proteomics isoform. Although part of the reason for this will be that more RNAseq reads map to longer sequences, it does suggest that either transcript expression is very different from protein expression for many genes or that transcript reconstruction methods may not be interpreting the RNAseq reads correctly. Recent large-scale comparisons have been critical of the accuracy of much of the transcript reconstruction process.[38,39] It may be that RNAseq deconvolution algorithms have a preferential bias for shorter transcripts. The set of approximately 3000 genes with reliably identified dominant isoforms that we have generated in this experiment would make an ideal gold standard set for validating RNAseq reconstruction algorithms.

The dominant protein isoform that we identify with the peptide data is expressed across the range of tissues and cell lines interrogated by the eight proteomics analyses. For those more than 95% of genes with a main experimental proteomics isoform, the remaining annotated transcripts will be alternative isoforms that are likely to be expressed in lower quantities, in limited tissues, or have a limited half-lives, if expressed at all. Our results reaffirm the concept of the gene, underline the importance of protein-level conservation, and will have a substantial effect on our understanding of cellular biology. We believe that these results open a new line of research that will be followed by many other experimental and computational investigations into the distribution and molecular functions of splice isoforms.

Acknowledgments

This work was supported by the National Institutes of Health (NIH, grant no. U41 HG007234). Additional funding from the Spanish Ministry of Economics and Competiveness (grant nos. BIO2012-40205, BIO2012-37926, RD07-0067-0014-COMBIOMED, RETICS-RD12-0042-0056, and PRB2-ProteoRed- PT13/0001/0017) and by the EU-FP7 Project BLUEPRINT (282510). J.M.R. is supported by the Spanish National Institute of Bioinformatics (www.inab.org), a platform of the Instituto de Salud Carlos III (INB-ISCIII, PRB2).

References

1. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008; 40:1413–1415. [PubMed: 18978789]
2. Wang E, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore S, Schroth G, Burge C. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. [PubMed: 18978772]
3. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7:S4. [PubMed: 16925838]
4. Johnson J, Castle J, Garrett-Engele P, Kan Z, Loerch P, Armour C, Santos R, Schadt E, Stoughton R, Shoemaker D. Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. Science. 2003; 302:2141–2144. [PubMed: 14684825]
5. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:760–774.
6. Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia A, Tress ML. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. Mol Biol Evol. 2012; 29:2265–2283. [PubMed: 22446687]

7. Smith CW, Valcárcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci. 2000; 25:381–388. [PubMed: 10916158]
8. Taneri B, Snyder B, Gaasterland T. Distribution of alternatively spliced transcript isoforms within human and mouse transcriptomes. J OMICS Res. 2011; 14:1–5.
9. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. Nature. 2012; 14:101–108. [PubMed: 22955620]
10. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013; 14:R70. [PubMed: 23815980]
11. Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet. 2009; 10:617. [PubMed: 19687803]
12. Chang K, Georgianna D, Heber S, Payne G, Muddiman D. Detection of alternative splice variants at the proteome level in *Aspergillus flavus*. J Proteome Res. 2012; 9:1209–1217. [PubMed: 20047314]
13. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hübner N, van Breukelen B, Mohammed S, et al. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 2013; 5:1469–1478. [PubMed: 24290761]
14. Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol Cell Proteomics. 2013; 12:2341–2353. [PubMed: 23629695]
15. Ning K, Nesvizhskii A. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinf. 2010; 11:S14.
16. Harte RA, Farrell CM, Loveland JE, Suner MM, Wilming L, Aken B, Barrell D, Frankish A, Wallin C, Searle S. Tracking and coordinating an international curation effort for the CCDS Project. Database. 2012:bas008. [PubMed: 22434842]
17. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, Valencia A, Tress ML. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013; 41:D110–D117. [PubMed: 23161672]
18. Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang CY, Moritz RL. The state of the human proteome in 2012 as viewed through PeptideAtlas. J Proteome Res. 2013; 12:162–171. [PubMed: 23215161]
19. Munoz J, Low TY, Kok YJ, Chin A, Frese CK, Ding V, Choo A, Heck AJ. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol. 2011; 7:550. [PubMed: 22108792]
20. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Pääbo S, Mann M. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011; 7:548. [PubMed: 22068331]
21. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics. 2012; 11 M111.014050.
22. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. A draft map of the human proteome. Nature. 2014; 509:575–581. [PubMed: 24870542]
23. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509:582–587. [PubMed: 24870543]
24. Ezkurdia I, Vázquez J, Valencia A, Tress ML. Analyzing the first drafts of the human proteome. J Proteome Res. 2014
25. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining results of multiple search engines in proteomics. Mol Cell Proteomics. 2013; 12:2383–2393. [PubMed: 23720762]

26. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011; 10:1794–1805. [PubMed: 21254760]
27. Koenig T, Menze BH, Kirchner M, Monigatti F, Parker KC, Patterson T, Steen JJ, Hamprecht FA, Steen H. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. J Proteome Res. 2008; 7:3708–3717. [PubMed: 18707158]
28. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML. *firestar*–advances in the prediction of functionally important residues. Nucleic Acids Res. 2011; 39:W235–W241. [PubMed: 21672959]
29. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–D301. [PubMed: 22127870]
30. For details, see Palumbo E, Djebali S, Breschi A. Blueprint Project CRG Pipeline. Feb 18.2013 ftp://ftp.ebi.ac.uk/pub/databases/blueprint//protocols/Analysis_protocols/README_rnaseq_analysis_crg
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
32. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. Ensembl 2013. Nucleic Acids Res. 2013; 41:D48–D55. [PubMed: 23203987]
33. Farajollahi S, Maas S. Molecular diversity through RNA editing: a balancing act. Trends Genet. 2010; 26:221–230. [PubMed: 20395010]
34. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012; 13:227–232. [PubMed: 22411467]
35. Tsui SK, Chan PP, Cheuk CW, Liew CC, Waye MM, Fung KP, Lee CY. A novel cDNA encoding for a LIM domain protein located at human chromosome 14q32 as a candidate for leukemic translocation. Biochem Mol Biol Int. 1996; 4:747–754. [PubMed: 8843343]
36. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2012; 28:235–242. [PubMed: 10592235]
37. Hoffman L, Gorbea C, Rechsteiner M. Identification; molecular cloning, and characterization of subunit 11 of the human 26S proteasome. FEBS Lett. 1999; 449:88–92. [PubMed: 10225435]
38. Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, Harrow J, Bertone P. RGASP Consortium Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013; 10:1177–1184. [PubMed: 24185837]
39. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, Grant GR, Hogenesch JB. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014; 15:R86. [PubMed: 24981968]

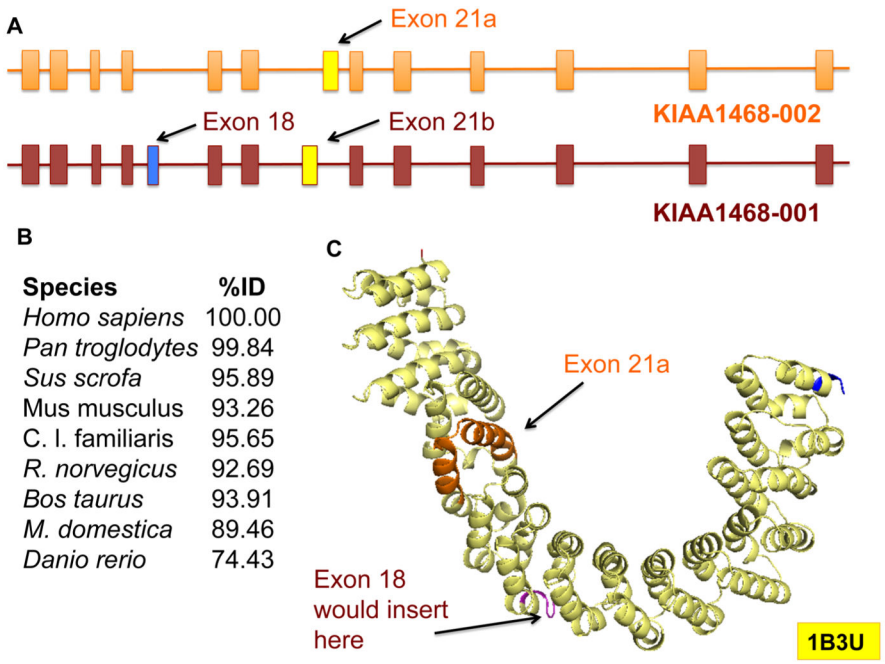


Figure 1.

The main proteomics, CCDS, and APPRIS isoforms differ because of the gene model. (A) The 3′ exons from two KIAA1468 transcripts. The arrows highlight the differences in the two transcripts (KIAA1468-001 and KIAA1468-002), inserted exon and exon 18 in KIAA1468-002 and a pair of mutually exclusively spliced exons (exons 21a and 21b). We find peptides for both mutually spliced exons but not for exon 18. Both CCDS and APPRIS select KIAA1468-001 because it does not have exon 18, but there are more peptides for mutually exclusive exon 21b from KIAA1468-002 than for exon 21a. (B) The orthologues found by the APPRIS database that align without gaps to the sequence of KIAA1468-001. (C) The structure of a protein similar to that encoded by *KIAA1468*, 1B3U. The region coded by the mutually exclusive homologous exons is shown in orange, the region where exon 18 from KIAA1468-002 would produce an insertion is shown in purple.

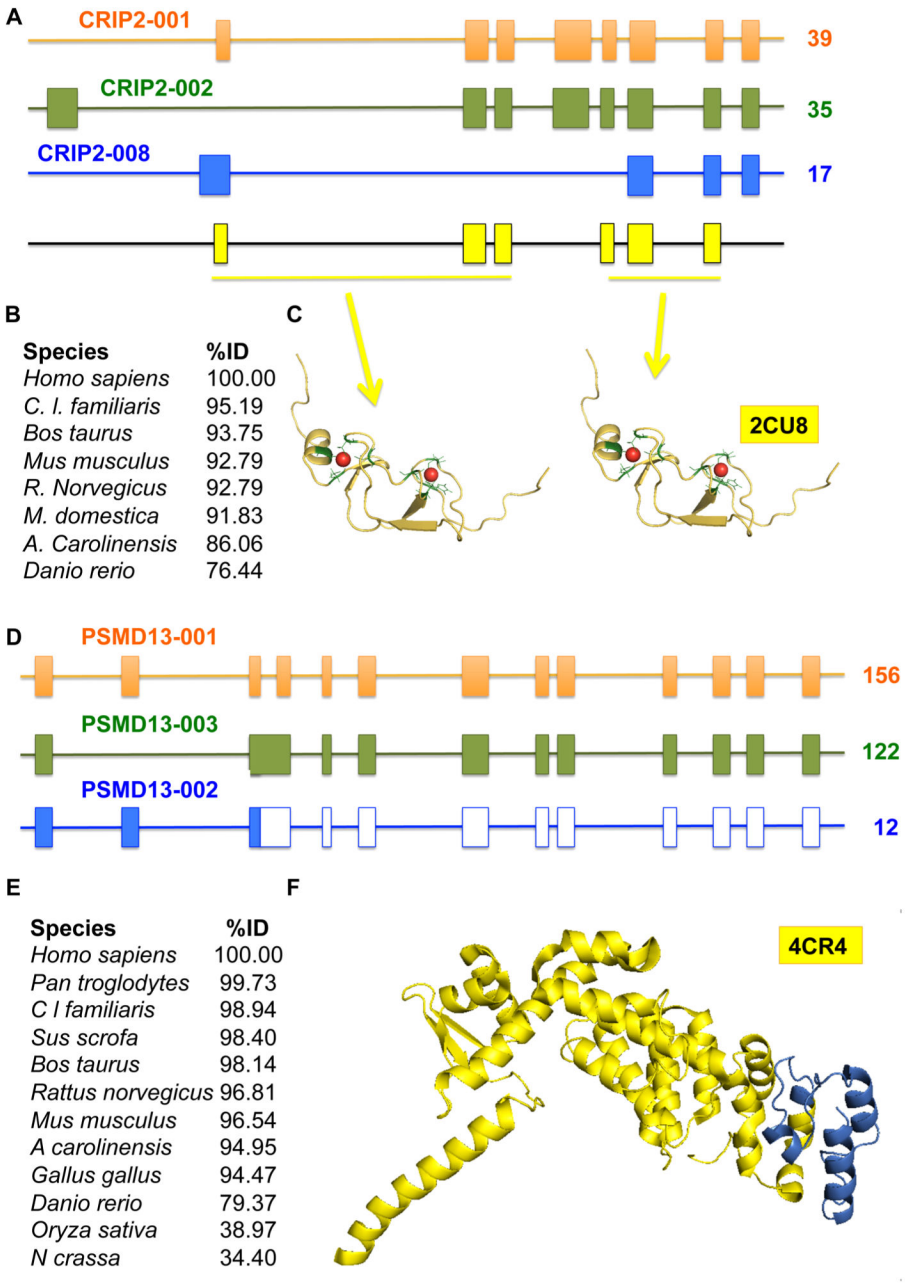


Figure 2.

The main proteomics isoform, the longest isoform, and the 5-fold dominant variants for *CRIP2* and *PSMD13*. (A) Transcripts from the GENCODE gene model of *CRIP2*. The transcripts selected by the proteomics experiment (CRIP2-001 in orange), the RNAseq experiment (CRIP2-008 in blue), and the longest variant (CRIP2-002 in green) are compared and the number of peptides detected for each isoform are shown in the right. The exons in yellow show the exons for which a 3D structure has been solved. (B) The orthologues found by the APPRIS database that align without gaps to the sequence of CRIP2-001. (C) The structure of the first domain of *CRIP2*, 2CU8, highly similar to domain 2 of *CRIP2*. (D) Transcripts from the GENCODE gene model of *PSMD13*. The transcripts selected by the

proteomics experiment (PSMD13-001 in orange), the RNaseq experiment (PSMD13-002 in blue), and the longest variant (PSMD13-003 in green) are compared and the number of peptides detected for each isoform are shown in the right. The nonfilled exons are not translated. (E) Model organism orthologues that align without gaps to the sequence of PSMD13-001 only. (F) The structure of a protein similar to PSMD13-001, 4CR4. The residues that would be coded by nonsense-mediated decay variant PSMD-002, if translated, are shown in blue.

Table 1

Comparison of Main Proteomics Isoform to Different Reference Isoforms

	genes[a]	comparable[b]	disagree[c]	% agree[d]	% disagree[e]
CCDS unique	13297	3331	46	98.6	1.4
APPRIS principal	15172	4186	93	97.8	2.2
longest isoform	20462	5011	520	89.6	10.4
HBM cell lines	4199	762	140	81.6	18.4
HBM tissues	5228	1038	237	77.2	22.8
unfiltered reads	15950	4618	656	85.8	14.2
2-fold reads	2465	1018	62	95.4	4.6

[a] Number of genes for which we could determine a reference isoform.

[b] Number of genes that we could compare with the main proteomics isoform.

[c] Number of genes in which the reference isoform disagreed with the main proteomics isoform.

[d] Percentage of genes in which there was agreement between the reference isoform and the main proteomics isoform.

[e] Percentage of genes in which there was no agreement between the reference isoform and the main proteomics isoform.