BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 11–17
doi:10.1093/bioinformatics/btm547

*Genome analysis*

# Determination and validation of principal gene products

Michael L. Tress[1,*], Jan-Jaap Wesselink[1], Adam Frankish[2], Gonzalo López[1], Nick Goldman[3], Ari Löytynoja[3], Tim Massingham[3], Fabio Pardi[3], Simon Whelan[4], Jennifer Harrow[2] and Alfonso Valencia[1]

[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Madrid, Spain [2]HAVANA Group, The Sanger Institute, [3]The European Bioinformatics Institute, Cambridge and [4]Faculty of Life Sciences, University of Manchester, Manchester, UK

Received on August 17, 2007; revised on October 17, 2007; accepted on October 22, 2007

Advance Access publication November 15, 2007

Associate Editor: Alex Bateman

**ABSTRACT**

**Motivation:** Alternative splicing has the potential to generate a wide range of protein isoforms. For many computational applications and for experimental research, it is important to be able to concentrate on the isoform that retains the core biological function. For many genes this is far from clear.

**Results:** We have combined five methods into a pipeline that allows us to detect the principal variant for a gene. Most of the methods were based on conservation between species, at the level of both gene and protein. The five methods used were the conservation of exonic structure, the detection of non-neutral evolution, the conservation of functional residues, the existence of a known protein structure and the abundance of vertebrate orthologues. The pipeline was able to determine a principal isoform for 83% of a set of well-annotated genes with multiple variants.

**Contact:** mtress@cnio.es

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Alternative messenger RNA (mRNA) splicing (Black, 2000; Gilbert, 1987) allows for the generation of a diverse range of mature RNAs. Studies have suggested that at least 60% of human genes can produce differently spliced mRNAs (Harrow *et al*., 2006; Scherer *et al*., 2006) and that alternative splicing has the potential to more than double the number of different proteins in the cell.

While alternative splicing can produce a range of differently spliced protein isoforms, there is conflicting evidence about their biological relevance. It has been suggested that the purpose of alternative splicing is to expand functional complexity and that the multiple variants are likely to encode functional proteins (Graveley, 2001; Hui and Binderief, 2005). Proteins with new functions are generated at different stages of development and in different tissues by a sophisticated regulation of the splicing process (Florea, 2006; Smith and Valcarcel, 2000). Many splice variants are hypothesized to function as dominant negative isoforms that regulate the pathways in which the main functional form is involved (Arinobu *et al*., 1999; Stojic *et al*., 2007).

However, recent work (Rodriguez-Trellez *et al*., 2005) has suggested that gene expression is not as tightly related to protein function as had been thought. In addition we have shown that despite widespread evidence for the expression of alternative transcripts, there was little to indicate this is translated into an increase in the repertoire of protein functions (Tress *et al*., 2007). Many of the proteins that result from alternative exon use would almost certainly have substantially rearranged structures with respect to their constitutively spliced counterparts (Tress *et al*., 2007; Talavera *et al*., 2007) and these changes are likely to have profound effects on the location and function of these alternative gene products. Predicting the effect of these changes on the cell is complicated, not least because heavy selection pressure would not normally tolerate such large transformations (Xing and Lee, 2006).

In addition it seems possible that gene expression is not as tightly related to protein function as has been thought (Rodriguez-Trellez *et al*., 2005). Recent work from this group showed that despite widespread evidence for the expression of alternative transcripts, there was little to indicate this translated into an increase in the repertoire of protein functions (Tress *et al*., 2007).

At present the SwissProt database (Bairoch *et al*., 2004), part of UniProtKB (The UniProt Consortium, 2007), provides the best organization of the complicated web of alternative variants. Even if the SwissProt database is relatively small, it is the *de facto* gold standard of protein databases because entries are manually curated.

As part of the manual curation of the proteins in the SwissProt database, all UniProt variants from the same gene are merged into a single entry. One crucial step of the merging process is the selection of one of the merged sequences as the 'display' sequence for the entry. The display sequence is selected after careful inspection and remaining merged sequences are tagged as alternative splice variants of the corresponding display sequence. The longest variant is often chosen as the display sequence, not necessarily because it is the principal functional isoform but because this allows annotators to map

*To whom correspondence should be addressed.

more features to the sequence (Amos Bairoch, personal communication).

Each SwissProt gene entry brings together experimental and predicted information, including domain definitions, functional annotation, cellular location, post-translational modifications and disease association. The entries are extensively cross-referenced to a range of external sources. All this information is associated to a single display isoform.

SwissProt display sequences are ideally suited for the goals of annotators, but there are many purposes for which it is important to know which of a gene's transcripts codes for the principal functional isoform.

Although many genes have been studied in depth, there are still a considerable number of genes with little experimental evidence. For these genes it is important to know which variant is likely to have the principal functional activity in order to design experiments to determine the structure and function of a protein. Labelling one of the splice variants as the principal isoform will allow research groups to concentrate their efforts on the main functional isoform.

In addition, identifying a principal splice isoform for a gene would allow bioinformatics groups to make more reliable predictions of function and structure. In particular, automatic prediction pipelines need reliable input data. One good example of this is the structure prediction for the protein PTPA_HUMAN by ModBase (Pieper *et al*., 2006). ModBase makes automatic predictions of structure using the SwissProt display sequence. The structure of SwissProt alternative isoform 2, missing the fourth protein coding exon of the display sequence, has already been solved. In order to model the structure of the display sequence with the inserted exon ModBase is forced to squeeze the extra 35 residues into an extended non-protein like loop.

Defining a principal functional isoform for each gene presents two problems. The first is that considerable experimental work would be required for each gene and the second is that it may be difficult to define a principal isoform for those genes where two (or more) variants might be regarded as equally important.

In order to define the principal coding variant for each gene, we had to make two assumptions. The first was that each gene has a single variant that gives rise to a principal functional isoform. The remaining annotated variants would then be alternatively spliced. This is a general assumption and comparative studies usually suggest that one isoform has the principal function or is expressed in most tissues or in most stages of development. While this is likely to be true for most genes, it will not be true for all genes.

The second assumption is that this principal variant is evolutionarily conserved between species. Alternative exons tend to be recent evolutionary developments (Alekseyenko *et al*., 2007), so this is a reasonable assumption. Again this may not always be true for all genes — the principal variant may have evolved (possibly through alternative splicing) towards a function distinct from those performed by the orthologous gene products in neighbouring species.

For the purposes of this study, we have defined the principal functional isoform as the isoform that performs an orthologous functional role across a wide range of related organisms.

Most of the methods used here are based on conservation between related proteins or transcripts. The success of these conservation-based methods depends on the evolutionary diversity of the species studied and on alternative exons evolving at measurably different rates. In those cases where there was no clear difference in the evolutionary rates of competing alternative exons, it was not possible to determine a principal isoform.

With the pipeline we were able to define a principal variant for 83% of the genes with multiple variants. Comparisons with SwissProt showed that the definitions from the pipeline concurred with the display sequences 75% of the time. In the majority of the cases where there was disagreement between our method and the SwissProt display sequences, the experimental and transcript evidence suggested that the definitions based on conservation point to the principal functional isoforms.

## 2 METHODS

The pilot project of the Encyclopaedia of DNA Elements (ENCODE) project (The ENCODE Consortium, 2004) analysed 44 regions of the human genome. The HAVANA group produced a manually curated set of annotated splice variants for these regions as part of the GENCODE project (Harrow *et al*., 2006). The October 2005 freeze of this HAVANA/GENCODE set was used as the reference set in this work. The set contained 1097 protein-coding transcripts from 434 distinct loci, but we were not able to use the entire HAVANA set in this study because 181 of these genes were annotated as having just a single splice variant. In addition, 38 genes had alternatively spliced transcripts that differed only in the 5′ or 3′ untranslated regions. These cases were ignored in this study, but could have been treated by two of the five methods. Here, we concentrate on the 215 loci that coded for at least two protein sequence distinct splice isoforms. This set comprised 804 variants, a mean of 3.74 variants per locus.

We used five separate methods to help determine the constitutive isoforms for the genes in the ENCODE project. Although it was not always possible to use each method for every gene, we attempted to take all methods into account when we made our selection.

As a working hypothesis, all of the HAVANA annotated variants in a locus had an equal chance to be the principal variant. Most of our methods were better at demonstrating which of the isoforms was unlikely to be the principal isoform and as a result they were mostly used as a means of rejecting the hypothesis that a given variant could be the principal variant. For some genes we were able to reject several variants, but could not determine which of the isoforms was the principal isoform.

### 2.1 Method 1: conservation of exonic structure

Transcripts that do not have conserved exonic structure between species are not likely to code for the principal isoform. Transcripts with exonic structure that was not conserved were rejected as candidates for the principal variant.

We measured the conservation of exonic structure between species for all the transcripts in the reference set. For the study, we compared the conservation of exon structure between humans and three other species, mouse, chimpanzee and macaque. The method is applicable to any sequenced eukaryotic genome. Here we detail the human–mouse comparison. An example is shown in Figure 1.

ENCODE loci were obtained from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/encode/). The mouse genome data came from mm5, the October 2004 freeze. Mouse homologues for the loci from the reference set were identified using

Blastn (Altschul *et al.*, 1997) with an *E*-value of $1E{-}10$. Annotated ENCODE loci transcripts were aligned to the mouse homologues using exonerate (Slater and Birney, 2005) and predicted transcripts were aligned using clustalw (Thompson *et al.*, 1994). The exonic structure of the transcripts was superimposed on the alignment. Exonic structure was regarded as conserved if all human and mouse intron positions in the transcripts could be aligned (Castelo *et al.*, 2005).

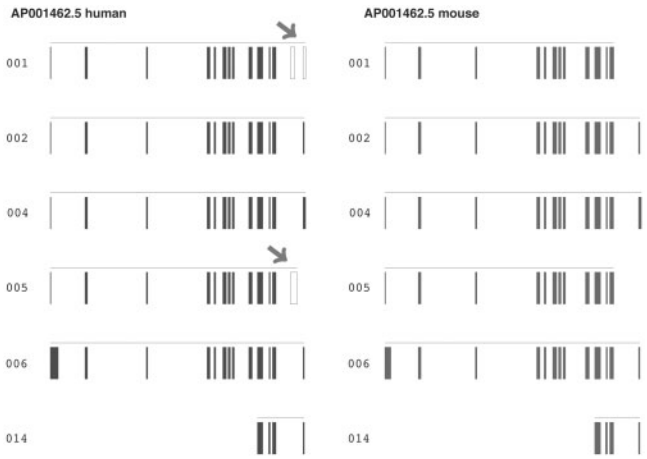


**Fig. 1.** Human and mouse exonic structure for the six variants of SF1. Human exonic structure for the SF1 variants (locus AP001462.5) shown left, mouse exonic structure on the right. Differences between human and mouse exonic structure are shown as white blocks and flagged by arrows. For variants 001 and 005 (the variant that codes for the SwissProt display sequence) exonic structure is not conserved.

## 2.2 Method 2: non-neutral evolution

Exons with unusual substitution patterns might indicate biological phenomena, such as the generation of a new function, but transcripts that contain one of these exons are unlikely to be the principal isoform. When one of the transcripts contained an exon with obvious non-standard conservation, we did not consider the variant transcript as a candidate for the principal isoform.

Exon conservation was assessed with prank (Löytynoja and Goldman, 2005) and SLR (Massingham and Goldman, 2005). Prank is a multiple alignment program based on an algorithm that avoids over-estimating deletions and correctly treats insertion events.

SLR detects non-neutral evolution through statistical tests that can identify positions that are unusually conserved and those that are unusually variable. SLR is based on studies that show that when the non-synonymous substitution rate is higher than the synonymous substitution rate at individual amino acid sites, this is an important indicator of positive selection.

Prank was used to align each mRNA variant in the reference set with homologues from vertebrate species and the resulting exon alignments were used to compute position-specific selection with SLR (see The ENCODE Project Consortium, 2007, for full details).

The prank alignments and SLR outputs are displayed in graphical form at http://www.ebi.ac.uk/goldman-srv/encode/encode_SEP2005_v3/.

There were clear differences in the substitution patterns of a small but significant fraction of individual exons. These exons have a pattern of substitution that does not appear to correspond to the usual evolutionary dynamics associated with protein-coding sequences and can be clearly seen in the visual output generated from SLR for each gene (see Fig. 2).

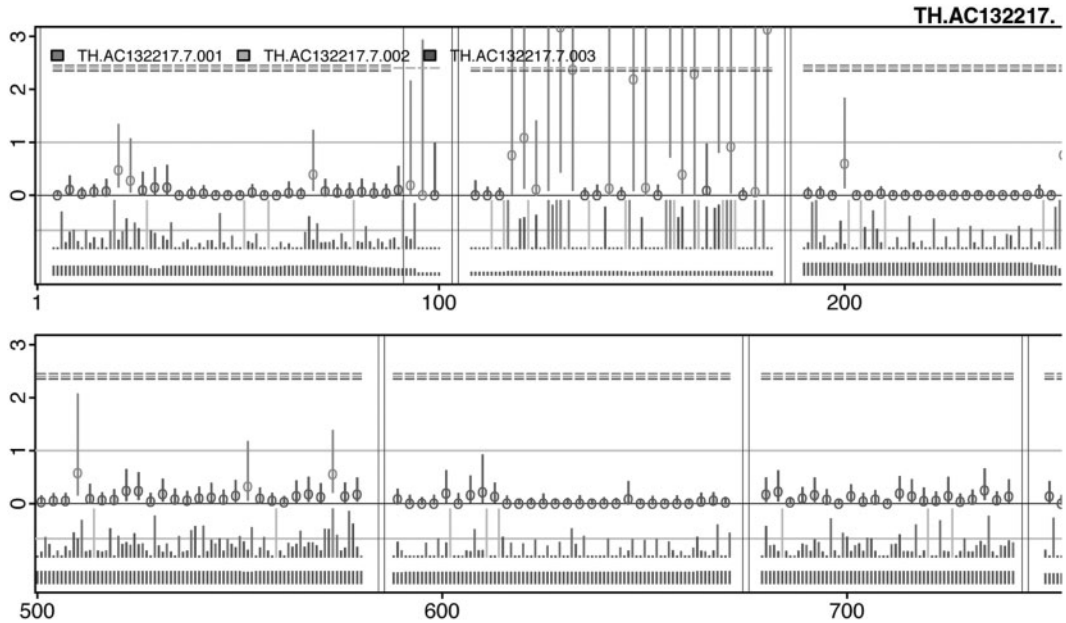


**Fig. 2.** The visual SLR output for the three variants of the TH locus (AC132217.7). At the top, the horizontal dashed lines indicate which exon belongs to which transcript (from top to bottom, 001 in red, 002 in green and 003 in blue). The colour of the circles (SLR score) and bars (confidence intervals) denote selection mode. Below this is a per nucleotide measure of conservation with abnormally fast sites coloured orange (third codon positions) or red (first or second codon positions). The black hatching at the foot of the record of the number of sequences available at each alignment position. The whole of the second coding exon between positions 100 and 200 is clearly differently conserved, suggesting that it is under unusual selective pressures. This exon is present in variant 002 and 003. On the basis of the SLR output, we rejected the hypothesis that these two variants could give rise to the principal functional isoform.

## 2.3 Method 3: protein structure mapping

Variants were also discounted as being principal functional variants if it was not possible to map their amino acid sequence onto a highly similar structural domain without having to introduce a deletion or insertion event resulting from an alternatively spliced exon. Variants that can be mapped to structure without these gaps have more chance of being the functional variant because we know that they are likely to fold properly. As of 2006, there were only five examples of alternative isoforms with resolved protein structures (Romero *et al*., 2006). A BLAST (Altschul *et al*., 1997) search of the Protein Databank (PDB, Berman *et al*., 2000) was sufficient to locate structures with amino acid sequences that were similar to those of the transcripts. If there are deletions/insertions in the alignment relative to the structure, the variant was discounted as the principal isoform (Fig. 3). We could map just over half the variants to similar homologous structures.

**Fig. 3.** Sequence to structure mapping. Human neurexin 2 is 82% sequence identical to neurexin 1 over the second LNS/LG domain. From the alignments between the two neurexins, we were able to map the sequence of the four variants of the NRXN2 gene (AP001092.3) onto the neurexin 1 structure (2h0b). The SwissProt display sequence (from variant 001) has an insertion of 15 residues relative to the structure of neurexin 1. These 15 residues would need to be squeezed in between the residues marked in red and yellow on the structure, thus breaking a beta sheet. This variant is therefore unlikely to be the primary variant.

## 2.4 Method 4: functional residue conservation

Exons that contain conserved functionally important residues are more likely to be part of the principal functional isoform of the protein. We used *firestar* (Lopez *et al*., 2007), a method that predicts functionally important residues in protein sequences. This method uses PSI-BLAST to align target sequences against sequence profiles pre-generated for a non-redundant set of PDB structures. Known functionally important residues are mapped onto the target sequences via the alignments (Fig. 4). The likely conservation of these functionally important residues in the target sequences is evaluated with SQUARE (Tress *et al*., 2004), a method for evaluating alignment reliability. Variants that were missing important conserved functional residues as a result of alternative splicing were rejected as possible primary variants.

## 2.5 Method 5: vertebrate alignments

Here we were looking for numbers of species, the more species that had a variant that aligned correctly to each transcript, the better. A good alignment was an alignment without insertions or deletions caused by alternative exons. Good alignments with more distant relatives (*danio*, *xenopus*, chicken) were regarded as more valuable than alignments with chimpanzee or dog. If the transcript is conserved over a greater evolutionary distance, it is more likely to be the constitutive variant. BLAST was used to search a non-redundant database of vertebrate sequences and the results of the search were not used to discard potential principal isoforms, but rather as a means of scoring each transcript.

## 2.6 Combining the methods

Most methods (the first four) are used to discard isoforms. If any one of the methods detects an unusual structure/conservation the isoform is rejected. There is no combination of these methods. If more than one isoform is left from the starting set of highly annotated variants, no decision can be made.

## 2.7 Evaluation of pipeline definitions

We analysed the pipeline definitions by inspecting aligned genomic sequences from a wide range of vertebrate species. We also analysed transcription evidence from multiple sources. All the data, including the genomic sequence alignments, are available from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

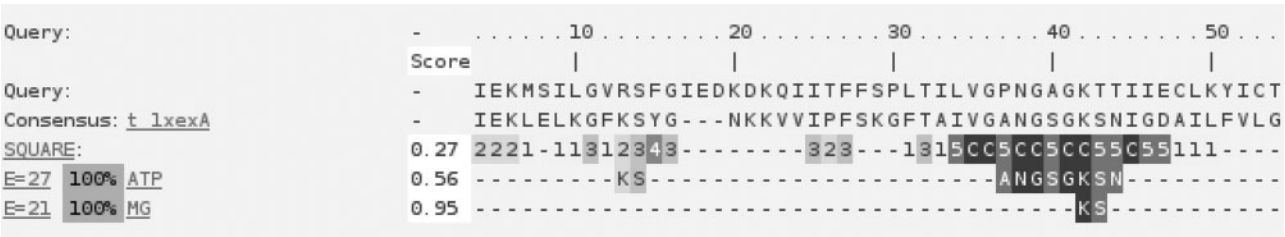


**Fig. 4.** Mapping functional residues. The output from *firestar*, showing the N-terminal end of the alignment between variant 001 of *RAD50* (AC0040401.1, labelled as 'Query') and a sequence with known structure (1xexA). The top line indicates the residue number of the variant. Below the alignment the numbers in coloured squares indicate locally conserved regions. The last two lines indicate the presence of ligand binding residues for ATP and magnesium. These last two rows are colour-coded: the darker the colours, the more conserved the residue. The structure clearly has conserved ATP binding residues and these are unlikely to have arisen by chance. Variant 002 (data not shown) is missing 139 residues from the N-terminal, including these functionally important residues, and was rejected as a candidate to be the principal variant.

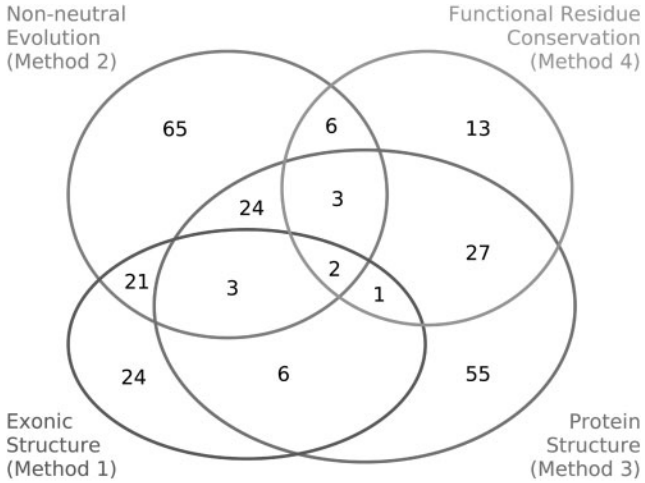


**Fig. 5.** The numbers of variants discounted by each of the first four pipeline methods, showing where variants were rejected by more than one method.

Where possible we also sought out experimental confirmation for the principal isoforms via PubMed searches.

## 3 RESULTS

We were able to use the first four methods in the pipeline to confirm principal isoforms for 99 genes. Prank/SLR was able to reject the hypothesis that a variant coded for the principal isoform in 124 cases, 121 isoforms could be discounted by structure mapping and 57 variants were rejected because they did not have conserved exonic structure in either mouse or chimpanzee. Evaluating conserved functional residues with *firestar* helped us to reject 52 variants. The overlap between these methods can be seen in Figure 5.

The principal isoform for a further 80 genes was selected based on the numbers of good alignments with other vertebrate species.

We were able to determine a principal isoform for 179 genes (83%). For 36 genes there was not enough information to determine a principal isoform. In most cases, this seemed to be either because the gene was evolving rapidly (as was the case of the immunoglobulin-like receptors) or was new in evolutionary origin. In both these cases, it is difficult to get good cross-species alignments. In a few cases, the results suggested that more than one variant might qualify as the principal isoform.

A total of 153 of the 179 genes for which we could define the principal isoform had a SwissProt display sequence. Here, it was possible to compare our definition of the principal functional splice isoform with the display sequence assigned by SwissProt.

The UniProtKB/SwissProt display sequence differed from the principal isoform selected by our pipeline in 37 of the 153 genes (Supplementary Table 1).

### 3.1 Examples

A close inspection of those cases where the selected principal isoform differed from the SwissProt display sequence shows that many of our selections are backed up by conservation evidence. For example the SwissProt display sequence for the TH gene (TY3H_HUMAN) has an extra exon in its coding sequence (CDS) that is conserved only in chimpanzee. The isoform selected by our method (isoform 001 in the reference set) is supported by SwissProt entries from cow (P17289) to eel (O42091). In addition, the donor splice site for the first coding exon of the SwissProt display sequence is not conserved for any species with sequence in this region.

The SwissProt display sequence CDS for PIP5K1A has a shifted acceptor splice site and a novel exon. While the shifted acceptor splice site appears to be conserved as far back as opossum, it is not universally conserved. The novel exon appears not to be very well conserved at all—only human, chimpanzee and baboon have all features required for a functional exon.

For a number of loci, it was possible to crosscheck our definitions against the experimental evidence. For example, the entire structure of human serine/threonine protein phosphatase 2A (PPP2R4) has been solved (Magnusdottir *et al.*, 2006) confirming our selected principal isoform. The SwissProt display sequence for PPP2R4 contains a short exon that shares a substantial overlap with a L1ME4 repeat sequence reported to be human specific (Jurka *et al.*, 2005). This exon is not present in the reported structure (2g62 in the PDB). The SwissProt display sequence for tafazzin (TAZ_HUMAN) contains an exon that appears to only be conserved in primates. Vaz *et al.* (2003) have shown that the principal isoform selected by the pipeline is the only one with full tafazzin activity.

Even where there is no confirmatory evidence for our selected principal isoform, there are sometimes clues to suggest that the SwissProt display sequence may not be the principal isoform. For example, SwissProt includes this comment about the display sequence for the SYT8 gene product: 'As it is truncated compared to orthologues, its function is not obvious'. In fact SYT8 is described as having two C2 domains (Li *et al.*, 1995). The truncated SwissProt display sequence does not have these two domains. In this case, the SwissProt display sequence is based on mRNAs that contain retained intron sequences. This intron retention leads to the incorporation of premature termination codons and the likely degradation of the transcripts via the nonsense-mediated decay pathway.

Several other SwissProt display sequences had little supporting evidence. For example, the only published support for the display sequence for HYPK (AC018512.7) is a large-scale sequencing paper. The transcript (AK000438.1) is chimeric; it links the upstream SERF2 locus to HYPK, and has been used to support an NMD variant of SERF2. The variant would contain the start codon of SERF2 inside the first intron. The start codon for the HYPK_HUMAN sequence is conserved, but as it lies in the middle of the SERF2 CDS, this is not unexpected. The pipeline selects a variant with 46 fewer N-terminal residues as the principal isoform.

The pipeline methods can efficiently label variants that are not likely to be the principal isoform, but they are less efficient at positively selecting the principal variant. In a number of cases where the SwissProt display sequence was rejected as the principal isoform, we were unable to determine a principal isoform with our methods. For example we can show that

the functional variant for the gene SF1 (Fig. 1) is likely to have a C-terminal different from that of the SwissProt display sequence, but we were unable to determine whether transcript 001 or 002 was the functional variant.

Not all human genes have been manually annotated in SwissProt, and we were able to determine a principal isoform for some of these unannotated genes. For example, we defined transcript 004 in the HAVANA set (Q8WXQ2) as the principal variant for the OSCAR gene (AC012414.3), while we rejected the hypothesis that the only sequence associated to the HISPPD2A gene in UniProtKB (Q6PFW1) might be the functional variant. For OSCAR, the splice sites of transcript 004 are conserved as far back as platypus and the transcript evidence points quite strongly to this variant. For HISPPD2A, a 12 base splice site shift in the transcript for Q6PFW1 is not universally conserved (it is missing in marmoset, rat, mouse, cow, opossum and zebrafish) while the 'regular' splice acceptor is conserved back to zebrafish.

In total we were able to confirm 11 of our definitions with evidence from external sources, but the results showed that adding more methods to our pipeline would improve the detection. Our definition of the principal isoform for NFS1 was different from the SwissProt display sequence, and the PRANK alignments showed that the N-terminal exon (60 residues) is not conserved among other species. However, we believe that the SwissProt display sequence is indeed the principal isoform because the NFS1 gene product is a mitochondrial protein and the N-terminal is the mitochondrial signal sequence (Biederbick *et al*., 2006). Although the mitochondrial signal sequence is clearly important to the function of the protein in the mitochondria, it is not as well conserved as the rest of the protein.

The sequences in SwissProt undergo a constant review process. Since we made these selections, SwissProt has changed the display sequence for two of the genes in this set. The display sequences for ASCL6_HUMAN and TKTL1_HUMAN were corrected in early 2007 and it would not be surprising to find that other entries have been updated since this manuscript was submitted.

## 4 DISCUSSION

As a result of this work, we have developed a pipeline that allows us to select the likely principal functional variant for well-annotated human genes. The pipeline employs a range of computational methods that choose the principal functional variant from among the known variants.

We have used five complementary methods to select the likely primary variant for 179 human genes. This represents 83% of the genes with multiple alternative variants. For those genes annotated by SwissProt, approximately 25% of the selected principal variants differed from the SwissProt display sequence.

It should be pointed out that these figures can not be extrapolated to the whole of the human genome for a number of reasons. The ENCODE pilot project set included 434 genes and we have only been able to look at 215 of these. The remaining 219 genes were not assigned protein sequence distinct alternative splice variants by HAVANA. The ENCODE regions were biased towards genes that had a single transcript (The ENCODE Project Consortium, 2007).

The methods used in this study were complementary. The majority of methods were conservation-based, requiring evolutionary information in the form of genomic and protein sequences. Two methods (structure mapping and *firestar*) also required structural information in the form of homologous proteins with known structure.

None of the methods was able to verify or reject every variant—the methods were inconclusive where the necessary evolutionary and structural information did not exist. Inevitably this meant that there were gaps in the evidence. While we were able to detect the principal variant for a high proportion of genes, for several it was impossible. For example, the ENCODE regions include several clusters of immunoglobulin-based receptors that are exclusively found in higher mammals. These receptors are evolving very rapidly, which meant that it was difficult to use conservation-based data for these genes.

For the majority of genes where the principal variant selected by our method did not agree with the SwissProt display sequence, we were able to suggest which of the alternative variants annotated by HAVANA was likely to be the principal isoform. However, in some cases our automatic methods were not able to point to a principal isoform. These cases would clearly require further intervention.

Where the principal isoform and the display sequences differed, the principal isoform chosen by our pipeline tended to be shorter than the SwissProt display sequence. This bears out the tendency of SwissProt to choose the longest variant as the display sequence. Three quarters of the principal isoforms we selected in Supplementary Table 1 were shorter than the corresponding SwissProt display sequence.

The results also highlight the importance of defining a principal isoform. At present, the selection of the SwissProt display sequence is of huge importance and has implications beyond SwissProt. The SwissProt display sequence is linked to many external servers and databases, so all the functional and structural information associated to these servers is linked to the display sequence. On top of that, those isoforms designated as alternative variants by SwissProt when the entries are merged, are removed from many versions of UniProt and these sequences also disappear from external databases. For example, the only isoform of PTPA_HUMAN that is included in the Pfam functional domains database (Finn *et al*., 2006) is the display sequence that has the extra exon. The principal isoform does not exist in Pfam. In turn, the PTPA_HUMAN display sequence is built into the seed alignment generated for the PTPA domain and as the Pfam seed alignments are regarded as the gold standard for alignments, the alignment with the incorrect principal isoform for PTPA_HUMAN is propagated further.

All these methods used to select the principal isoform are already automated or can easily be automated. Together with reliable annotations of splice variants, such as those from the HAVANA group for the genes in the ENCODE regions, they form the basis of a pipeline that will be used to confirm, refine and extend the annotation of the principal functional variants from the human genome. The method will be able to flag

unusual variants and will allow annotators to select the principal variants more easily. The method could also be extended to work with any number of well-annotated genomes.

Many human genes have not yet been annotated in SwissProt. The designation of one of the variants as the principal isoform will be of importance for the design of experimental work and a vital first step for large-scale studies of the human genome, such as the ENCODE project, where it is important to predict the effect of alternative splicing on structure, function and location.

## ACKNOWLEDGEMENTS

This work was funded by the BioSapiens Network of Excellence (grant number LSHG-CT-2003-503265) and by the National Institute of Bioinformatics (www.inab.org), a platform of 'Genoma España'. NG thanks the Wellcome Trust for financial support, and IBM for the grant of an IBM eServer BladeCenter to the EMBL-EBI for use in its research work.

*Conflict of Interest*: none declared.

## REFERENCES

Alekseyenko,A.V. *et al.* (2007) Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*, **13**, 661–670.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Arinobu,Y. *et al.* (1999) Antagonistic effects of an alternative splice variant of human IL-4, IL-4delta2, on IL-4 activities in human monocytes and B cells. *Cell Immunol.*, **191**, 161–167.

Bairoch,A. *et al.* (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Biederbick,A. *et al.* (2006) Role of human mitochondrial Nfs1 in cytosolic iron-sulfur protein biogenesis and iron regulation. *Mol. Cell Biol.*, **26**, 5675–5687.

Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.

Blanchette,M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Castelo,R. *et al.* (2005) Comparative gene finding in chicken indicates that we are closing in on the set of multi-exonic widely expressed human genes. *Nucleic Acids Res.*, **33**, 1935–1939.

Finn,R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Florea,L. (2006) Bioinformatics of alternative splicing and its regulation. *Brief. Bioinformatics*, **7**, 55–69.

Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 901–905.

Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.

Harrow,J. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**, S4.

Hui,J. and Bindereif,A. (2005) Alternative pre-mRNA splicing in the human system: unexpected role of repetitive sequences as regulatory elements. *Biol. Chem.*, **386**, 1265–1271.

Jurka,J. *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.

Li,C. *et al.* (1995) Ca2+-dependent and -independent activities of neural and non-neural synaptotagmins. *Nature*, **375**, 594–599.

Lopez,G. *et al.* (2007) *firestar* – prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.

Löytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA*, **102**, 10557–10562.

Magnusdottir,A. *et al.* (2006) The crystal structure of a human PP2A phosphatase activator reveals a novel fold and highly conserved cleft implicated in protein-protein interactions. *J. Biol. Chem.*, **281**, 22434–22438.

Massingham,T. and Goldman,N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.

Pieper,U. *et al.* (2006) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.

Rodriguez-Trelles,F. *et al.* (2005) Is ectopic expression caused by deregulatory mutations or due to gene-regulation leaks with evolutionary potential? *Bioessays*, **27**, 592–601.

Romero,P.R. *et al.* (2006) Alternative splicing in concert with protein intrinsic disorder enables increase functional diversity in multicellular organisms. *Proc. Natl Acad. Sci. USA*, **103**, 8390–8395.

Scherer,S.E. *et al.* (2006) The finished DNA sequence of human chromosome 12. *Nature*, **440**, 346–351.

Slater,G. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Stojic,J. *et al.* (2007) Three novel ABCC5 splice variants in human retina and their role as regulators of ABCC5 gene expression. *BMC Mol. Biol.*, **8**, 42.

Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.

Talavera,D. *et al.* (2007) The (in)dependence of alternative splicing and gene duplication. *PLoS Comput. Biol.*, **3**, 33.

The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tress,M.L. *et al.* (2004) SQUARE – determining reliable regions in sequence alignments. *Bioinformatics*, **20**, 974–975.

Tress,M.L. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.

Vaz,F.M. *et al.* (2003) Only one splice variant of the human TAZ gene encodes a functional protein with a role in cardiolipin metabolism. *J. Biol. Chem.*, **278**, 43089–43094.

Xing,Y. and Lee,C. (2006) Alternative splicing and RNA selection pressure – evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, **7**, 499–509.