UNIVERSIDAD AUTÓNOMA DE MADRID

Facultad de Medicina
Departamento de Bioquímica

Protein Isoforms: Functional Importance and Tissue Specificity

TESIS DOCTORAL

Jose Manuel Rodríguez Carrasco

Director
Michael Tress

Centro Nacional de Investigaciones Oncológicas (CNIO)

AGRADECIMIENTOS

A mis padres por su amor y sacrificio.
A mis hermanas y familia por su amor y compañía.
A mis profesores por sus enseñanzas y paciencia.
A Michael Tress *for his friendship and knowledge*.
A Fera y mis hijos, Laura y Lucas… por todo.
A todo aquel que me ha dado algo que aprender, que apreciar y que amar.

ABSTRACT

The number of protein coding genes in the human reference gene sets has stabilized at slightly more than 20,000 genes in recent years, principally as a result of painstaking manual curation efforts. Although the three main gene sets, Ensembl/GENCODE, RefSeq, and UniProtKB, have similar numbers of genes, it is not clear how many of these genes coincide between the three sets.

Many researchers were surprised by the relatively low numbers of human coding genes and some have sought other explanations for an assumed human complexity such as alternative splicing. The alternative splicing of messenger Ribonucleic acid (RNA) is a fundamental molecular process that regulates eukaryotic gene expression and can generate a wide range of mature RNA transcripts. Many thousands of alternatively spliced transcripts are routinely detected in RNA-seq studies, although reliable large-scale mass spectrometry-based proteomics analyses identify only a small fraction of annotated alternative isoforms. Indeed, proteomics experiments strongly suggest that most genes have a single main protein isoform.

In this thesis, we present three papers on the functional description of coding genes, and of the principal and alternative protein isoforms derived from alternative splicing. In the first publication, we present the updates to the APPRIS Database. APPRIS selects a single protein isoform, the *principal isoform*, as the reference for each gene based on protein structural and functional features and information from cross-species conservation. Experimental evidence shows that the APPRIS principal isoform almost always coincides with the main cellular protein isoform. In the paper we detail the expansion of gene sets for multiple species, refinements in the core methods that make up the annotation pipeline and the merge of individual Ensembl/GENCODE, RefSeq, and UniProtKB reference gene sets. APPRIS now provides a measure of reliability for individual principal isoforms and updates with each release of the reference sets.

In the second paper, we analyse human protein-coding genes in the three main reference sets: Ensembl/GENCODE, RefSeq and UniProtKB. We find that one in eight of these genes are classified differently in at least one of the reference sets. Evidence from various sources suggests that many of the 22,210 genes in the union of the three sets are unlikely to code for functional proteins.

In the final publication, we carried out a reanalysis of a large-scale proteomics study of human tissues in order to determine to what extent tissue-specific alternative splicing can be detected at the protein level. We found evidence of significant tissue-specific differences across more than a third of the splice events that we interrogated. Tissue specific alternative protein forms were particularly abundant in nervous and muscle tissues. By contrasting the proteomics evidence with data from a large-scale transcriptomics analysis, we found that more than 95% of tissue specific events in which proteomics and RNA-seq analyses agree on tissue-specificity evolved over 400 million years ago. Our results suggest that tissue specific alternative splicing has played a crucial role in the development of the brain and the heart in vertebrates.

RESUMEN

El número de genes humanos que codifican a proteínas dentro de las bases de datos (BD) de referencia humanos se ha estabilizado en un poco más de 20,000 genes en los últimos años. Principalmente como resultado de minuciosos esfuerzos de curación manual. Aunque las tres BD de referencia, Ensembl/GENCODE, RefSeq y UniProtKB, tienen un número similar de genes, no está claro cuántos de estos genes coinciden entre los tres conjuntos.

El empalme alternativo del ácido ribonucleico mensajero (ARN) es un proceso molecular fundamental que regula la expresión de genes eucariotas y puede generar una amplia gama de transcripciones de ARN. Aunque muchos miles de transcritos de empalme alternativamente se detectan de forma rutinaria en los estudios de *RNA-seq*[1], los análisis de proteómica basados en espectrometría de masas identifican solo una pequeña fracción de isoformas alternativas. De hecho, los experimentos de proteómica sugieren que la mayoría de los genes tienen una única isoforma proteica. En esta tesis presentamos tres artículos sobre la descripción funcional de genes codificantes y de las isoformas proteicas principales y alternativas derivadas del empalme alternativo.

En la primera publicación, presentamos las actualizaciones de APPRIS. Algoritmo que selecciona una única isoforma proteica, la isoforma principal, como referencia para cada gen, en función de las características estructurales y funcionales de las proteínas y la información de la conservación entre especies. La evidencia experimental muestra que la isoforma principal APPRIS casi siempre coincide con la isoforma principal de la célula. En el artículo detallamos la expansión de las anotaciones para múltiples especies, la mejora de los métodos, y la creación de una fusión de genes basado en las tres BD de referencia. Además, proporciona una medida de fiabilidad para isoformas principales.

En el segundo artículo, analizamos genes humanos que codifican a proteínas en las tres BD de referencia: Ensembl/GENCODE, RefSeq y UniProtKB. Encontramos que uno de cada ocho de estos genes se clasifica de manera diferente en al menos uno de las BD de referencia. La evidencia de diversas fuentes sugiere que es poco probable que muchos de los 22,210 genes de los tres conjuntos codifiquen a proteínas funcionales.

En la publicación final, llevamos a cabo un nuevo análisis de un estudio proteómico a gran escala de tejidos humanos con el fin de determinar hasta qué punto se puede detectar el empalme alternativo específico de tejido. Encontramos diferencias significativas específicas de tejido en más de un tercio de los eventos. Las isoformas de proteínas alternativas eran particularmente abundantes en los tejidos nerviosos y musculares. Al contrastar la evidencia de proteómica con datos de transcriptómica, encontramos que más del 95% de los eventos específicos de tejidos que coinciden entre ambos análisis, evolucionaron hace más de 400 millones de años. Nuestros resultados sugieren que el empalme alternativo específico de tejido ha jugado un papel crucial en el desarrollo del cerebro y el corazón de los vertebrados.

[1]Siglas en inglés para *RNA sequencing*

(English) TABLE OF CONTENTS / (Español) ÍNDICE

Sections in English: orange colour
Apartados en español: color azul

ABBREVIATIONS

Term	Description
AS	Alternative Splicing
BLAST	Basic Local Alignment Search Tool
BPS	Branch Point Sequence
CCDS	Collaborative Consensus coDing Sequence
cDNA	complementary DNA
CDS	CoDing Sequence
CHESS	Comprehensive Human Expressed Sequences
CNV	Copy Number Variants
DNA	DeoxyriboNucleic Acid
ENCODE	ENCyclopedia Of DNA Elements
ESE	Exonic Splicing Enhancer
ESS	Exonic Splicing Silencer
EST	Expressed Sequence Tag
GENCODE	GENome enCyclopedia Of DNA Elements
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
hnRNP	heterogeneous nuclear RiboNucleoProtein
indel	insertion/deletion polymorphism
ISE	Intronic Splicing Enhancer
ISS	Intronic Silencing Silencer
mRNA	Messenger RNA
MS/MS	tandem mass spectrometry
NCBI	National Center for Biotechnology Information
NMD	Nonsense-Mediated Decay
ORF	Open Reading Frame
PED	PEptides from Different experiment
Pfam	Protein families database
PI	Principal Isoforms
PPT	PolyPyrimidine Tract
PDB	Protein Data Bank
PTB	Polypyrimidine Tract-Binding
pre-mRNA	precursor Messenger RiboNucleic Acid
RefSeq	The Reference Sequence database
RNA	RiboNucleic Acid
RNA-seq	RNA sequencing
snRNP	small nuclear RiboNucleoProtein
SS	Splice Site
TIGR	The Institute for Genomic Research
TSL	Transcript Support Level
UniProtKB	Universal Protein Resource KnowledgeBase

INTRODUCTION

The concept of a “gene” is basic to the understanding of genetics and molecular biology. At the time when the term was coined, it was seen from the phenotypic perspective as a distinct region, a “locus”, on a chromosome explaining mechanisms of heredity, development, and physiological function. Later, with the discovery of Deoxyribonucleic acid (DNA) and the publication of the “Central Dogma” of molecular biology (Crick, 1970), a gene became a physical entity that is transcribed and finally translated into protein.

In this model a gene is the region of DNA that contains the necessary information for the expression of a protein or other molecule that ultimately helps in the survival, reproduction and function of the organism (Figure 1). The transcription of protein-coding genes in eukaryotes generates a precursor messenger RNA (pre-mRNA) which is converted into mature messenger RNA (mRNA) ready for translation into protein. Eukaryotes have elaborated a complex mechanism of modifying their primary RNA transcripts, called “splicing”. Pre-mRNA transcripts contain intervening sequences, known as *introns*, which do not become part of the final mRNA. Regions of pre-mRNA that are retained and ligated for translation are known as *exons* (Gilbert, 1978). In the process of mRNA maturation introns are selectively excised out and exons are ligated together. The spliced mRNA molecule forms a continuous protein-coding region ready to be translated into a protein molecule.

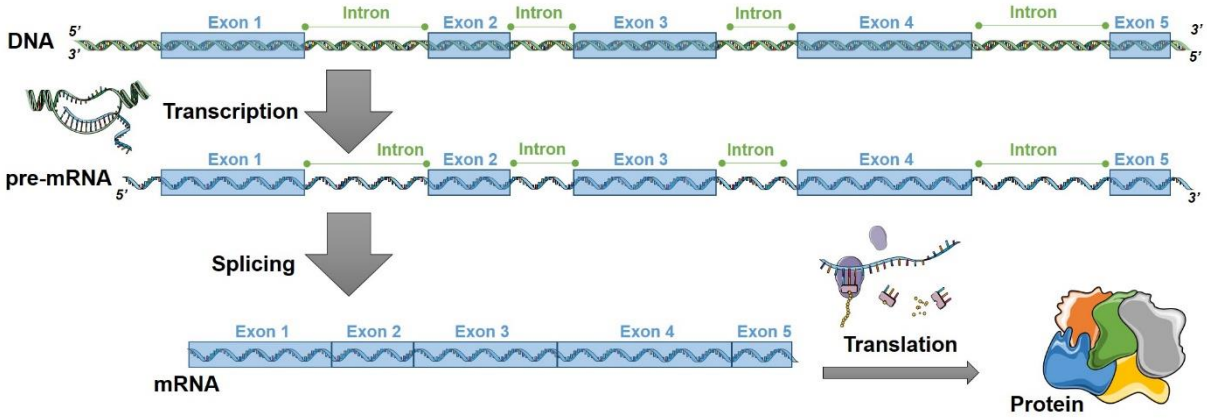


Figure 1. ***Transcription of protein-coding genes in eukaryotes****.* ***Transcription)*** *By this model, the transcription generates a pre-mRNA with exons and introns.* ***Splicing)*** *Pre-mRNA introns are removed by the spliceosome (see later), and a mature mRNA is generated.* ***Translation)*** *The mature mRNA is translated into a protein by the ribosome complex.*

However, the question “what constitutes a gene?” has been much debated in recent years (Brosius, 2009; Gerstein *et al.*, 2007; Gingeras, 2007; Mattick, 2003; Mercer & Mattick, 2013; Pearson, 2006). When the “completion” of the human genome sequence was announced (Collins *et al.*, 2003), the gene still was a genomic region with clear structural boundaries. The current view of transcription is becoming more complicated. In particular, a locus may generate multiple transcripts due to alternative splicing (AS). Alternative splicing can change the genotype-phenotype relationship, because it has the potential to generate different protein isoforms, implying different physiological functions derived from the same gene. This complexity complicates the work of scientists tasked with describing the human genome. To this end, Gerstein *et al.* proposed that *“a gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products”* (Gerstein *et al.*, 2007).

The Human Reference Gene Set

Genome annotation comprises all efforts to assign biological functions, mechanistic and structural roles, and observations linked to genomic positions to every nucleotide in the genome. The Encyclopedia of DNA Elements project (ENCODE) (The ENCODE Project Consortium *et al.*, 2012) was established to annotate the human genome with all possible functional information.

As part of this project, the GENCODE consortium (Harrow *et al.*, 2012) was formed to identify and map all protein-coding genes within the ENCODE regions. The GENCODE consortium is composed of several groups that are dedicated to producing high-accuracy annotations of evidence-based gene features based on manual curation, computational analyses and targeted experiments. The consortium initially focused on 1% of the human genome in the Encyclopedia of DNA Elements pilot project (The ENCODE Project Consortium *et al.*, 2007) and expanded this to cover the whole genome (The ENCODE Project Consortium *et al.*, 2012). GENCODE is now part of Ensembl (Zerbino *et al.*, 2018) and their annotations are regularly released as the Ensembl/GENCODE gene sets. They are also accessible via the Ensembl and UCSC Genome Browsers (Haeussler *et al.*, 2019).

In addition, there are other large-scale gene annotation projects in progress on the human genome. The RefSeq project (O'Leary *et al.*, 2016) at the National Center for Biotechnology Information (NCBI) combines manual and automated processes, and collaboration to produce a standard set of stable, non-redundant reference sequences.

For each "gene set" or "genebuild" produced, the vast majority of models are based upon transcript evidence. A recent approach, CHESS (Comprehensive Human Expressed Sequences) (Pertea *et al.*, 2018), has taken this to an extreme by assembling the gene models for their catalog of human genes and transcripts entirely from deep RNA sequencing experiments by the Genotype-Tissue Expression (GTEx) (Lonsdale *et al.*, 2013) project. CHESS is an entirely automatic annotation project and is not subject to any manual scrutiny.

In 2002, the UniProt Knowledgebase (UniProtKB) (The UniProt Consortium, 2018) was created. The UniProt Knowledgebase consists of two sections: a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and a section with computationally analyzed records that await full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as "UniProtKB/Swiss-Prot" (reviewed, manually annotated) and "UniProtKB/TrEMBL" (unreviewed, automatically annotated), respectively. UniProtKB protein sequences are not all based on genomic coordinates and it has been noted that some of the RefSeq and UniProtKB sequences are inconsistent with the sequences expected from the coding regions on the human genome (Farrell *et al.*, 2014; Harte *et al.*, 2012). Although the differences in the RefSeq mRNA and UniProtKB protein sequences from the reference genome have been pointed out several times, the cause of this discordance has not been well characterized (Shirota & Kinoshita, 2016).

The different methods employed by these public resources can result in distinct representations of genes, transcripts, and proteins. However, the collaborative consensus coding sequence (CCDS) project (Pruitt *et al.*, 2009) tracks identical coding sequence (CDS) annotations in RefSeq and Ensembl mouse and human genomes and ensures that they are consistently represented on the NCBI, Ensembl/GENCODE, and UCSC Genome Browsers with a stable identifier.

The Gradual Downward Trend of the Human Protein Gene Count

Estimating the number of human genes dates back to the 1940s when the genetic code and even the structure of DNA were unknown (Figure 2). In 1948, James N. Spuhler estimated the number of human genes (Spuhler, 1948) based on the chromosomal length occupied by genes comparing with the fruit fly (then 42,000 genes) and extrapolating the number derived from X-linked lethal mutations (19,890-30,420 genes). At about that time, Muller (Muller, 1950) estimated the number of human genes between 5,000 to 20,000 genes. In 1964, Friedrich Vogel (Vogel, 1964) calculated the number of genes dividing the length of the human genome by the gene-length of 50,000 nucleotides inferred from the length of genes in Dipteran giant chromosomes (60,000 human genes). Shortly thereafter, Muller revised his earlier estimate to “not much more than 30,000” genes based on newer data on spontaneous mutations and frequencies of X-ray induced mutations (Muller, 1966).

In 1990, the U. S. Human Genome Project claimed to have sequenced the human genome and to have located the suspected 50,000-100,000 human genes without providing any data or reference for this estimate (U.S. Department of Health and Human Services & Department of Energy, 1990). The success of sequencing and high-throughput technologies provided further numbers, including 20,000-40,000 genes implied by the measurement of RNA re-association kinetics (Benjamin Lewin, 1990), 80,000 genes implied by determining and extrapolating CpG island coverage (Antequera & Bird, 1993), and 64,000 genes implied by expressed sequence tag (EST) sequencing followed by clustering and extrapolation (Fields *et al.*, 1994).

As the release of the first draft of the human genome was approaching, researchers from The Institute for Genomic Research (TIGR) predicted 110,000 to 134,000 genes made available in the TIGR Gene Index based on massive expressed sequence tag (EST) (Liang *et al.*, 2000). In the same journal issue, other researchers predicted 33,630 to 34,700 genes based on similar EST data (Ewing & Green, 2000) and 28,000-34,000 genes by comparison with pufferfish (Crollius *et al.*, 2000).

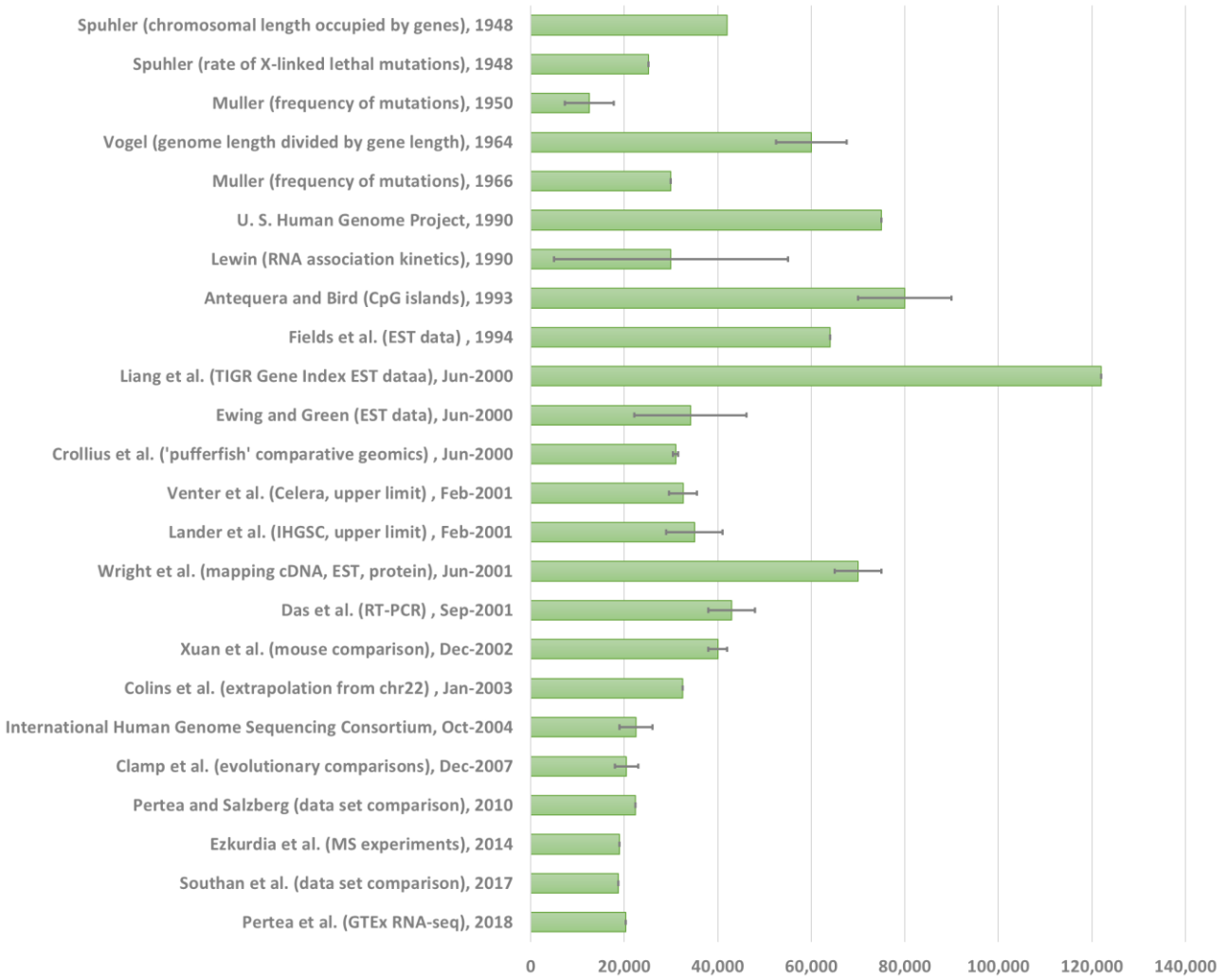


Figure 2. ***Estimates of human protein-coding number based on different lines of evidence.*** *Taken from (Hatje et al., 2019; Southan, 2004) with the addition of more recent papers (Ezkurdia et al., 2014; Pertea et al., 2018; Pertea & Salzberg, 2010; Southan, 2017).*

The publication of the draft human genomes in early 2001 did not stop the speculation about higher gene numbers, though estimates were close to the ranges predicted 26,588-38,588 genes (Venter *et al.*, 2001) and 30,000-40,000 genes (Lander *et al.*, 2001). Extrapolation of RT-PCR data of chromosome 22 predicted 41,000-45,000 genes (Das *et al.*, 2001) and mapping of available complementary DNA (cDNA), EST, and protein data combined with gene predictions suggested 65,000-75,000 genes (Wright, F. A., *et al.*, 2001). Still in 2003, when the human genome sequence was “finished”, researchers predicted 29,000-36,000 genes based on the extrapolation of a refined annotation of chromosome 22 (Collins *et al.*, 2003) and up to 40,000 protein-coding genes based on analysis of conserved sequence elements between human and mouse (Xuan *et al.*, 2003).

With the publication of the final draft of the Human Genome Project (International Human Genome Sequencing Consortium, 2004), the number of protein-coding genes was revised downwards again to between 20,000 and 25,000. In 2007, Clamp and co-workers (Clamp *et al.*, 2007) used evolutionary comparisons to suggest that the most likely figure for the number of protein-coding genes would be at the lower end of this continuum, just 20,500 genes. The Clamp analysis suggested that a large number of annotated open reading frames (ORFs) were not protein coding because they had features resembling non-coding RNA and lacked evolutionary conservation. The study suggested that there were relatively few novel mammalian protein-coding genes and that the 24,500 genes annotated in the human gene catalogue at the time would end up being cut by 4,000.

The number of protein-coding genes annotated in the Ensembl/GENCODE database (Frankish *et al.*, 2019; Harrow *et al.*, 2006) has also been on a downward trend since its inception. More than two thousand automatically predicted genes have been removed from the reference genome as a result of the merge with the manual annotation, often by being re-annotated as non-coding biotypes. The most recent GENCODE release (GENCODE v35 08/2020) contains 19,954 protein-coding genes. Most recently, CHESS (Pertea *et al.*, 2018), which is an entirely automatic annotation project, predicted 20,352 protein-coding genes.

Validating Coding Potential

Coding genes need to produce functional proteins and the best way to validate whether they do that is by detecting evidence of the gene product. Manual annotation of protein-coding genes requires many different sources of evidence (Frankish *et al.*, 2019; Guigó *et al.*, 2006). The most convincing evidence, experimental verification of cellular protein expression, is technically challenging to produce. Proteomics technology has improved considerably over the last decades (Aebersold & Mann, 2003; Mallick & Kuster, 2010), and it has become an increasingly important tool in genome annotation (Brosch *et al.*, 2011; Deutsch *et al.*, 2015; Ezkurdia, Valencia, and Tress *et al.*, 2014; Tanner *et al.*, 2007).

The shotgun proteomics approach (Aebersold & Mann, 2003; Gygi *et al.*, 1999; Link *et al.*, 1999; Washburn *et al.*, 2001) has become the method of choice for identifying and quantifying proteins in most large-scale studies. This strategy is based on digesting proteins into peptides followed by peptide sequencing using tandem mass spectrometry (MS/MS) and automated database searching. Compared with methods of analysis based on extensive protein separation prior to MS-based identification, such as two-dimensional gels (Görg *et al.*, 2004), shotgun proteomics allows higher data throughput and better protein detection sensitivity. MS/MS experiments has become an increasingly important tool for validating the translation of protein-coding genes (Brosch *et al.*, 2011; Deutsch *et al.*, 2015; Ezkurdia *et al.*, 2014; Tanner *et al.*, 2007), and large-scale mass spectroscopy experiments are now the main source of evidence of alternative splicing at the protein level.

Ezkurdia *et al* analysed the human genome with seven sets of proteomics data and found peptide evidence to support 11,840 coding genes (Ezkurdia *et al.*, 2014). The study found that proteins with annotated protein functional domains, functional residues, homology to known structures or cross-species conservation were more likely to be detected in the proteomics experiments than proteins without these features.

Gene family age (the oldest phylogenetic division in which a gene from the same family is found) and gene age were also related to peptide detection. These ages were calculated using Ensembl Compara phylogenetic trees (Herrero *et al.*, 2016). Peptides were detected for 96.4% of genes that evolved in the Fungi-Metazoa clade and did not duplicate (1,136 genes). By contrast, the most recently evolved genes (those with primate gene family age) and the least conserved genes were much less likely to be detected in proteomics experiments. Discriminating peptides were found for just 0.9% of the 563 primate-specific genes and 2% of the 987 genes with a low conservation score in APPRIS (Rodriguez *et al.*, 2013).

Alternative Splicing

Alternative splicing (AS) is a fundamental molecular process regulating eukaryotic gene expression that results in a single gene coding for multiple proteins (Black, 2003; A. J. Lopez, 1998; Smith & Valcárcel, 2000). In this process, exons can be included or excluded in different combinations to create a diverse range of mRNA transcripts from a single pre-mRNA (Figure 3). It was first described in the 80's, when it was discovered that membrane-bound and secreted antibodies are encoded by the same gene (Alt *et al.*, 1980; Early *et al.*, 1980).

Splicing in general, and AS in particular, is also important for regulation of the levels and tissue specificity of gene expression and, if disrupted, can lead to disease (Cartegni *et al.*, 2002; Tazi *et al.*, 2009; Venables, 2004; Wang, G. S. & Cooper, 2007).

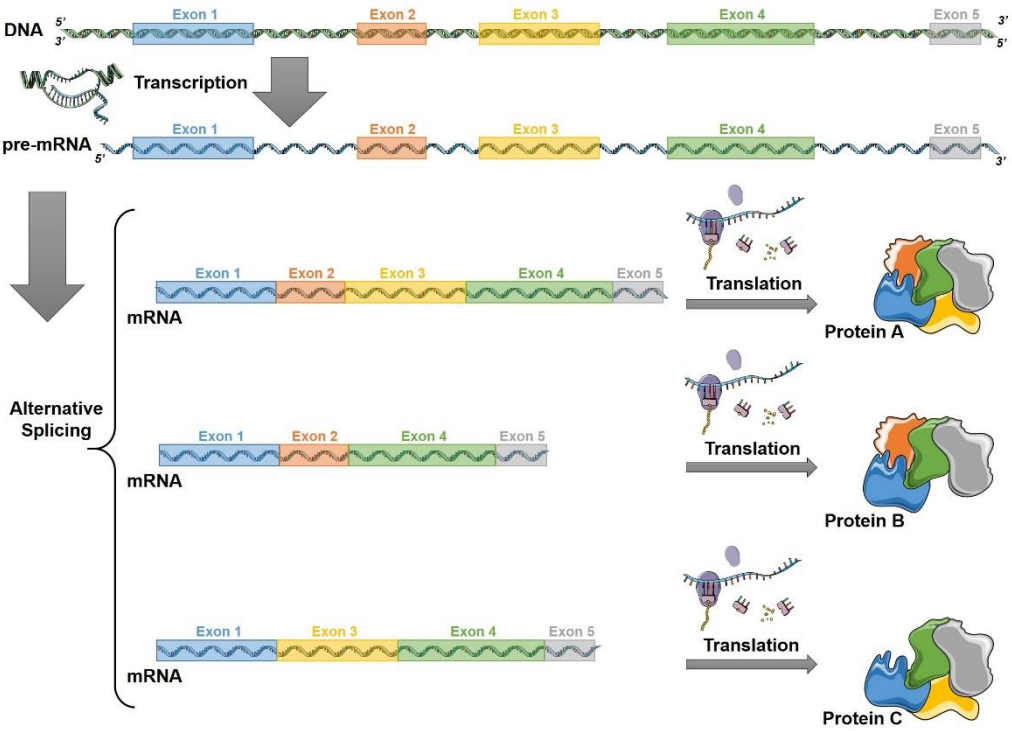


Figure 3. ***Alternative Splicing.*** *The figure shows the theoretical effect of AS with different combinations of coding exons producing three different proteins.*

The Splicing Machinery

The splicing reaction, which forms the central step in the production of mRNAs, involves the recognition of introns and exons by the splicing machinery. It can be regulated at many different levels, but most alternative splicing is a result of differential splice-site recognition by the *spliceosome*. The spliceosome is a complex composed of five small nuclear RNAs (U1, U2, U4, U5 and U6) that assemble with proteins to form small nuclear ribonucleoproteins (snRNPs) (Hoskins *et al.*, 2011; Jurica & Moore, 2003; Staley & Guthrie, 1998).

The spliceosomal machinery (Figure 4) is a coordinated series of RNA–RNA, RNA–protein and protein–protein interactions (Hoskins & Moore, 2012; Trowitzsch *et al.*, 2009). The spliceosome recognizes four conserved signals: the exon–intron junctions at the 5′ and 3′ ends of introns - the 5′ splice site (5′ SS) and 3′ splice site (3' SS) -, the branch point sequence (BPS) located upstream of the 3′ SS and the polypyrimidine tract (PPT) located between the

3′ SS and the BPS. First, U1 binds to the 5’ SS of an exon and U2 binds near BPS just upstream of the 3’ SS of the adjacent exon (Peled-Zehavi *et al.*, 2001). U2 snRNP is recruited to the branch region through interactions with the E complex component U2AF (U2 snRNP auxiliary factor). Later, a tri-snRNP complex, composed of U4/U6/U5, joins in and leads to the formation of an active complex that catalyzes splicing. Once the splicing is over, the spliceosome disassembles and all components are recycled for future splicing reactions (Hnilicová & Staněk, 2011).

Exons and introns also contain short, degenerate binding sites for auxiliary splicing proteins. These sites (Figure 4) are called exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs) and intronic silencing silencers (ISSs). Splice-site recognition is mediated by proteins that bind specific regulatory sequences, such as the serine/arginine (SR) proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs), polypyrimidine tract-binding (PTB) proteins, the *TIA1* RNA-binding protein, *Fox* proteins and *Nova* proteins (Chen & Manley, 2009; Hui, 2009; Licatalosi & Darnell, 2010).

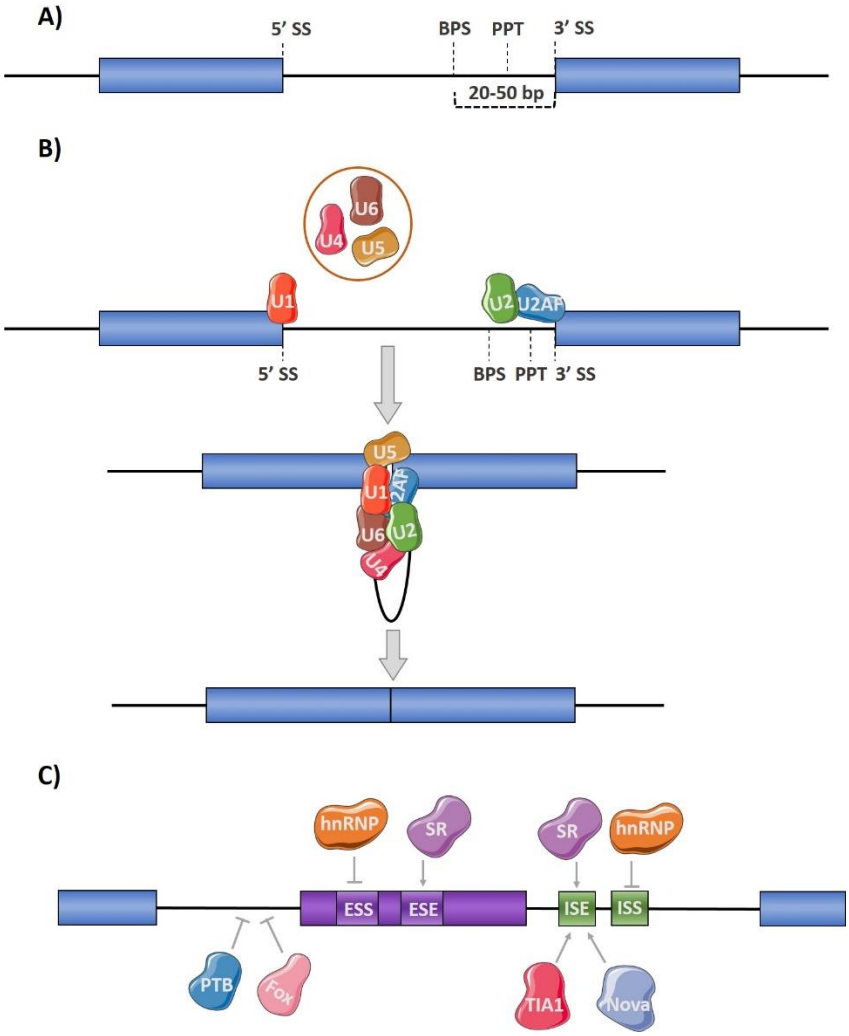


Figure 4. ***The Splicing Machinery****. Splicing is catalyzed by the spliceosome, which recognizes and assembles on exon–intron boundaries to catalyze intron processing. A) The four signals that enable recognition of RNA by the spliceosome are 5′ SS, 3’ SS, the BPS and the PPT. B) Several steps in the process of the splicing machinery with the action of multiple components. C) Exons and introns contain short, degenerate binding sites for splicing auxiliary proteins.*

Types of Alternative Splicing

Exons that are present in all variants within a gene are often referred to as *constitutive exons*. In this definition, *alternative exons* are exons that are not involved in all the variants of a gene. Although as annotation databases have grown, this definition has come to be somewhat problematic. The cornucopia of splicing variants for some genes can often be so great that almost all annotated exons end up defined as alternative.

Systematic analyses of ESTs and microarray data have revealed several types of alternative splicing (Pan *et al.*, 2008; Wang, E. T. *et al.*, 2008). These events can occur during the splicing process or as the mRNA is formed from the transcription step of the central dogma of molecular biology (Figure 5):

- **Exon Skipping** or **Cassette Exon**: In this case, exon(s) are included or excluded from the final gene transcript leading to extended or shortened mature mRNA variants. Exon skipping accounts for nearly 40% of AS events in higher eukaryotes but is extremely rare in lower eukaryotes (Alekseyenko *et al.*, 2007; Kim, E. *et al.*, 2007; Sugnet *et al.*, 2004).

- **Alternative 5' Splice Site (5’ SS) and 3' Splice Site (3’ SS)**: Alternative gene splicing includes joining of different 5' and 3' splice sites. In this kind of splicing, two or more alternative 5' splice sites compete for joining to two or more alternate 3' splice sites. Alternative 3′ SS and 5′ SS selection account for ~18% and ~8% of all AS events in higher eukaryotes, respectively (Kim, E. *et al.*, 2007; Koren *et al.*, 2007; Sugnet *et al.*, 2004).

- **Intron Retention**: An event in which an intron is retained in the final transcript. This is one of the rarest AS events in vertebrates and invertebrates, accounting for less than 5% of known events (Alekseyenko *et al.*, 2007; Kim, E. *et al.*, 2008; Sakabe & de Souza, 2007; Sugnet *et al.*, 2004). By contrast, intron retention is the most prevalent type of AS in plants, fungi and protozoa (Kim, E. *et al.*, 2008).

- **Mutually Exclusive Exons**: Another very uncommon splicing event. One of two exons (or one group out of two exon groups) is retained in mRNAs after splicing, while the other one is spliced out. Here, two (or more) splicing events are not independent any more, but are executed or disabled in a coordinated manner.

- **Alternative Promoters** and **Alternative polyadenylation**: Here transcription either starts or ends at different points. Alternative promoters are those pre-mRNA transcripts that have distinct 5' exons composition. Alternative polyadenylation occurs when distinct polyadenylation sites provide different 3' end points for transcripts. Both mechanisms can occur in combination with alternative splicing and provide additional variety in mRNAs derived from a gene (Ast, 2004; Black, 2003; Kim, E. *et al.*, 2008).

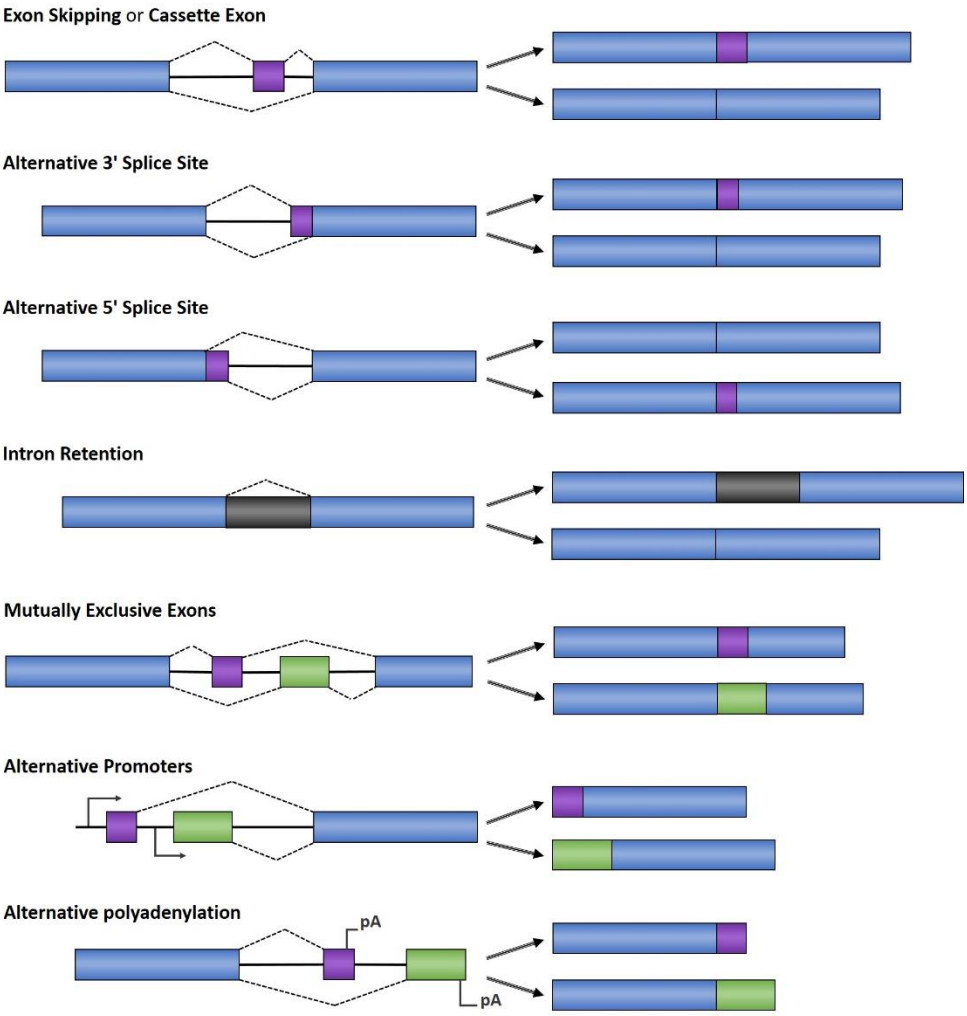


Figure 5. ***Types of alternative splicing events.*** *There are several different types of alternative splicing events. In the figure, exons and final transcripts are illustrated as boxes while lines represent introns. Constitutive exons are shown in blue, and alternatively spliced exons are depicted in green or purple. Retained introns occur in the absence of splicing, with the intervening intro (black) included in the final transcript.*

The Functional Impact of Alternative Splicing at Protein Level

One of the current priorities of the scientific community is the understanding of cellular responses specific to tissue, developmental stage or environmental conditions. Alternative splicing is a mechanism that has the potential to expand the cellular protein repertoire far beyond the one gene–one protein model (Nilsen & Graveley, 2010; Smith & Valcárcel, 2000) and has been linked to tissue and developmental differences.

The presence of multiple alternative mRNA transcripts from the same gene is unequivocally supported by EST and cDNA sequence evidence (Harrow *et al.*, 2012), microarray data (Sánchez-Pla *et al.*, 2012), and RNA-seq data (Juntawong *et al.*, 2014; Uhlén *et al.*, 2015). Despite the overwhelming evidence for alternative splicing at the transcript level, there is limited support for the translation of these alternative transcripts into protein isoforms. Individual experiments do provide evidence for the expression of isoforms for certain genes (Kelemen *et al.*, 2013).

Alternative splicing of messenger RNA produces a wide variety of differently spliced RNA transcripts that may be translated into diverse protein products. Some studies suggest that

human coding genes could generate on average more than ten alternative transcripts (Hu *et al.*, 2015; Pertea *et al.*, 2018). Assuming almost all of these transcripts are translated into functional alternative splice isoforms, we might expect the overall protein population to increase 10-fold from 20,000 (the number of human coding genes) to 200,000. This increase would have profound biological consequences.

Theoretically, all these coding transcripts could be translated into functional protein isoforms, which could in turn diversify the range of cellular functions. This possible expansion of function is often suggested to be the reason that humans have so few coding genes (Nilsen & Graveley, 2010; Smith & Valcárcel, 2000). However, although we have a limited understanding of the function of a small number of these alternative isoforms (Kelemen *et al.*, 2013), there is a general lack of knowledge about the functional roles of the vast majority of annotated splice isoforms. If translated to protein, most annotated splice variants are likely to produce isoforms with substantially altered 3D structure and drastic changes in biological function (Melamud & Moult, 2009; Tress *et al.*, 2007).

Initially, it was not clear whether alternative transcripts and proteins are expressed more or less equally across tissues, whether different transcripts or isoforms were dominant in different tissues, or whether it would be biologically relevant to designate one transcript or isoform per gene as dominant and the rest as alternative. Large-scale transcriptomics studies (Bahar *et al.*, 2011; Djebali *et al.*, 2012; Gonzàlez-Porta *et al.*, 2013) showed that genes have dominant transcripts but with contrasting results. While in some most genes had a single dominant transcript across all cell lines (Bahar *et al.*, 2011; Gonzàlez-Porta *et al.*, 2013), in others the majority of protein-coding genes had at least two different dominant transcripts depending on the cell (Djebali *et al.*, 2012).

Most Genes Have a Single Main Protein Isoform

Proteomics studies strongly suggest that most genes have a single main protein isoform (Abascal *et al.*, 2015). Abascal *et al* analysed peptides from eight large-scale data sets (Deutsch *et al.*, 2015; Ezkurdia *et al.*, 2014; Geiger *et al.*, 2012; Kim, M.-S. S. *et al.*, 2014; Munoz *et al.*, 2011; Nagaraj *et al.*, 2011; Wilhelm *et al.*, 2014) identifying 12,716 genes but just 282 alternative splice events. In total, these eight datasets covered over 100 distinct tissues and cell lines, yet less than 0.4% of the peptides mapped to alternative isoforms. Almost all peptides mapped to a single isoform per gene.

A related study investigated the relationship between this main proteomics isoform and other means of determining reference isoforms (Ezkurdia *et al.*, 2015). They found that all methods for selecting a reference isoform for a gene were better than random - a random selection of isoforms would have agreed with the main proteomics isoform 46% of the time (Figure 6) – but that methods based on RNA-seq evidence performed worse than manual annotators and worse than a method based on protein features and conservation.

Dominant RNA-seq transcripts, expressed five times more than across all tissues or cell lines (Gonzàlez-Porta *et al.*, 2013), agreed with the main proteomics isoform over 77.2% of comparable genes, while the Highest Connected Isoforms (Li *et al.*, 2015), based on transcript-level expression and interactions in a functional network, coincided with the main isoform over 78% of genes (Figure 6).

Both these methods performed worse than the strategy of selecting the longest isoform, the method of choice for selecting a reference isoform in practically all studies and databases

and the basis of UniProtKB display isoforms. Although it has no biological basis, the longest isoform coincided with the main experimental proteomics isoform across 89.6% of genes.

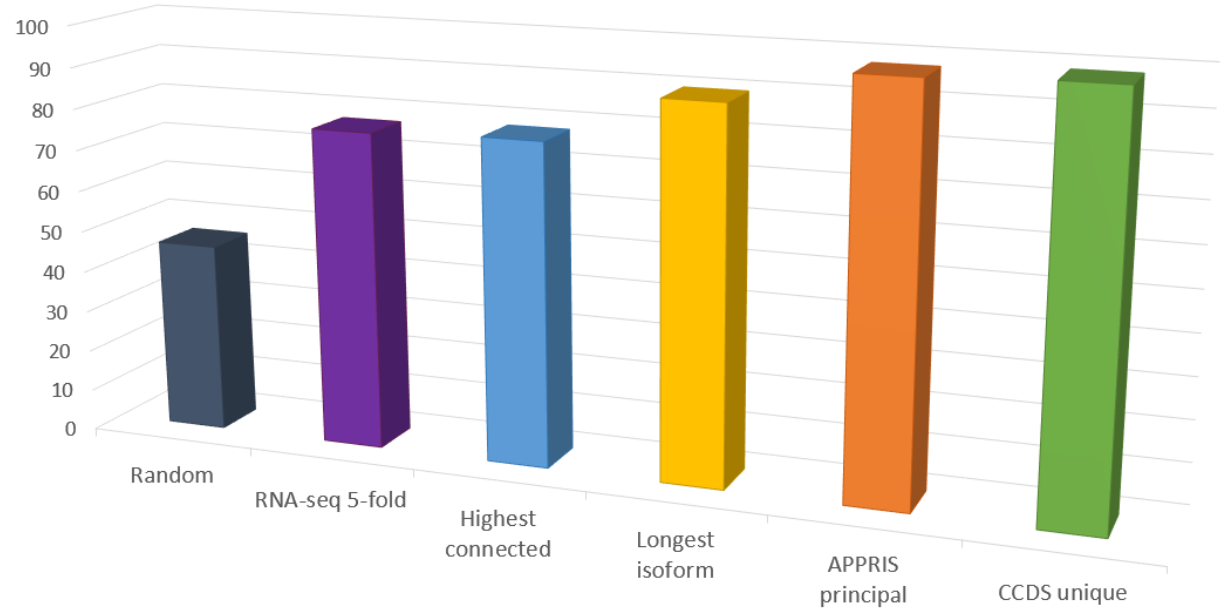


Figure 6. ***Comparison of Main Proteomics Isoform to Other Reference Isoforms.*** *Percentage of genes in which there was agreement between the reference isoform and the main proteomics isoform (Ezkurdia et al., 2015; Tress et al., 2017).*

The CCDS variants are based on cDNA evidence and agreed on by manual annotators (Pruitt *et al.*, 2009). For those genes where there was just a single CCDS variant per gene, the agreement with the main proteomics isoform was much higher at 98.6%. Finally, the APPRIS database (Rodriguez *et al.*, 2013) predicts principal isoforms based on the preservation of protein features and cross-species conservation. In the study the main proteomics isoforms agreed with the principal isoforms selected by APPRIS over 97.8% of genes (Figure 6).

Both single CCDS variants and APPRIS principal isoforms are reliable predictors of the main cellular isoforms.

APPRIS in Detail

APPRIS (Rodriguez *et al.*, 2013, 2015) is a computational system that provides annotations of alternative splice variants and identifies principal isoforms. The theory behind APPRIS was developed in 2008 (Tress *et al.*, 2008) and the database was developed over a four year period within the GENCODE consortium (Frankish *et al.*, 2019; Harrow *et al.*, 2012). APPRIS annotates alternative gene products with reliable, biologically relevant data.

APPRIS annotates splice isoforms in protein-coding genes with protein structural and functional features and information from cross-species conservation. Currently the annotation pipeline comprises six modules (Figure 7). The first four methods are referred to as the “core” methods in APPRIS:

- **Matador3D** detects similarity to structural homologs in the PDB (Rose *et al.*, 2017).
- **firestar** (G. Lopez *et al.*, 2011) predicts functionally important amino acid residues.

- **SPADE** identifies Pfam functional domains via the PfamScan algorithm (El-Gebali *et al.*, 2019).

- **CORSAIR** carries out BLAST (Altschul *et al.*, 1997) searches against vertebrate protein sequences to determine the number of orthologs that align correctly and without gaps.

- **THUMP** makes unanimous predictions of trans-membrane helices from three predictors (Jones, 2007; Käll *et al.*, 2004; Viklund & Elofsson, 2004).

- **CRASH** predicts the presence and location of signal peptides using the SignalP and TargetP programs (Emanuelsson *et al.*, 2000; Petersen *et al.*, 2011).

The pipeline also uses these features to select a single reference isoform for each protein-coding gene, here termed the *principal isoform*. This principal isoform has the most conserved protein features and the most evidence of cross-species conservation. At the same time isoforms that have lost conserved protein features or do not have cross-species conservation are flagged as alternative.

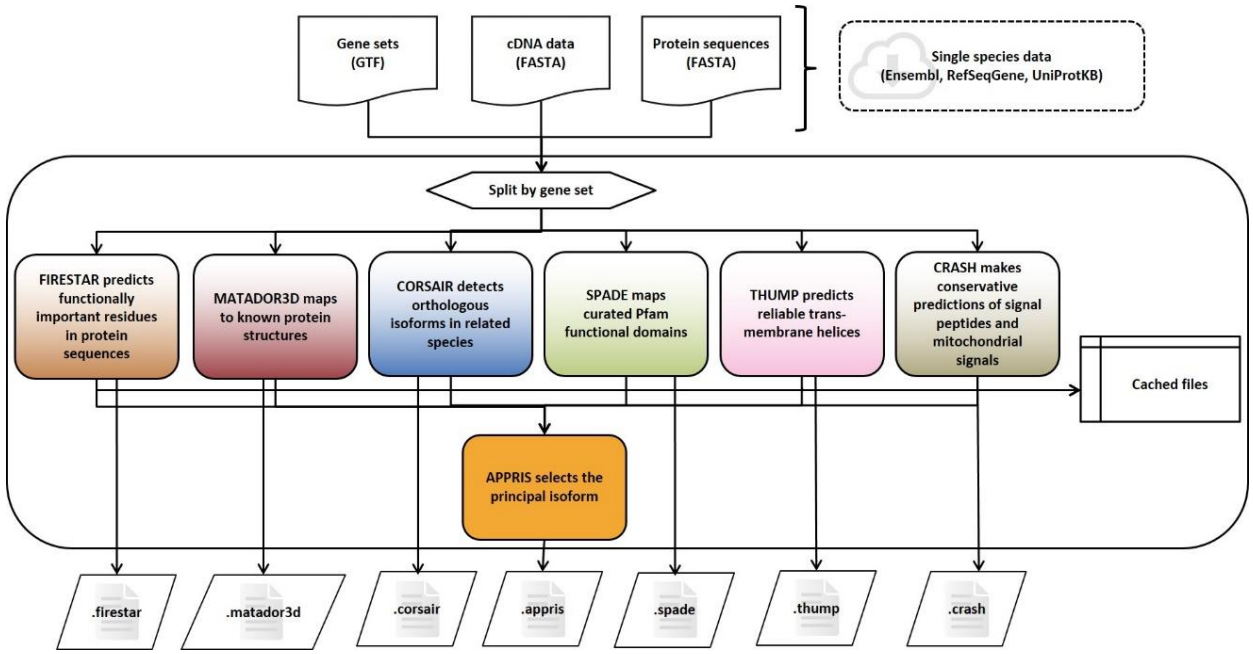


Figure 7. ***APPRIS pipeline.*** *The inputs to the pipeline are the peptide sequences of the isoforms (FASTA), and/or the gene information file (GTF) and the cDNA sequences of the transcripts. The results of the six modules of APPRIS are used to annotate the splice isoforms and the final module selects the principal isoforms.*

More recently, the APPRIS WebServer and WebServices (Rodriguez *et al.*, 2015) were developed to provide access to the computational methods implemented in the APPRIS database and to allow the generation of annotations in a flexible, modular and automatic high throughput mode.

There are another number of databases that can annotate alternative transcripts with some of these features. ProSAS (Birzele *et al.*, 2008) provides a unified resource for analyzing effects of alternative splicing events in the context of human, mouse and rat protein

structures. AS-ALPS (Shionyu *et al.*, 2009) provides information useful for analyzing the effects of alternative splicing in human and mouse on protein structure, interactions with other biomolecules and protein interaction networks. ASPicDB (Martelli *et al.*, 2011) generates annotations for human protein variants through machine learning tools including protein type (globular and transmembrane), localization, presence of Pfam domains, signal peptides, GPIanchor propeptides, transmembrane and coiled-coil segments. Finally, tappAS (de La Fuente *et al.*, 2020) facilitates the analysis of alternative splicing and alternative UTR processing from a functional perspective. Most of the annotations of this framework are at the transcript level and for both coding and non-coding regions. Annotations at protein level integrate data from multiple databases and tools.

Most Alternative Exons Are Not Under Selective Pressure

APPRIS also classifies protein isoforms as either principal or non-principal (Rodriguez *et al.*, 2013) based on differences in cross-species conservation or biological features. Liu and Lin expanded on this by splitting coding regions into three distinct categories (Liu & Lin, 2015): principal isoform-specific (PI-specific) coding regions, non-principal isoform-specific (NPI-specific) regions, and overlapping regions (coding sequences that are shared by the principal and non-principal isoforms). Liu and Lin mapped the variants from the 1000 Genomes Project (Auton & Salcedo, 2015) onto coding regions and demonstrated that the NPI-specific coding regions are significantly enriched in amino acid-changing variants particularly those that have a strong impact on protein function, and have higher derived allele frequencies.

Previous studies have indicated that human alternatively spliced exons are subjected to relaxed selective pressure or positive selection (Ramensky *et al.*, 2008; Xing & Lee, 2005). Further investigations supported these results (Tress *et al.*, 2017). Exons from APPRIS principal isoforms have a substantially lower proportion of high-impact variants than exons from alternative isoforms (Figure 8). Although alternative sites represent only 5% of all data, they contribute 29% of the high-impact variants across all allele frequencies and 57% of high-impact variants for the most common allele frequencies.

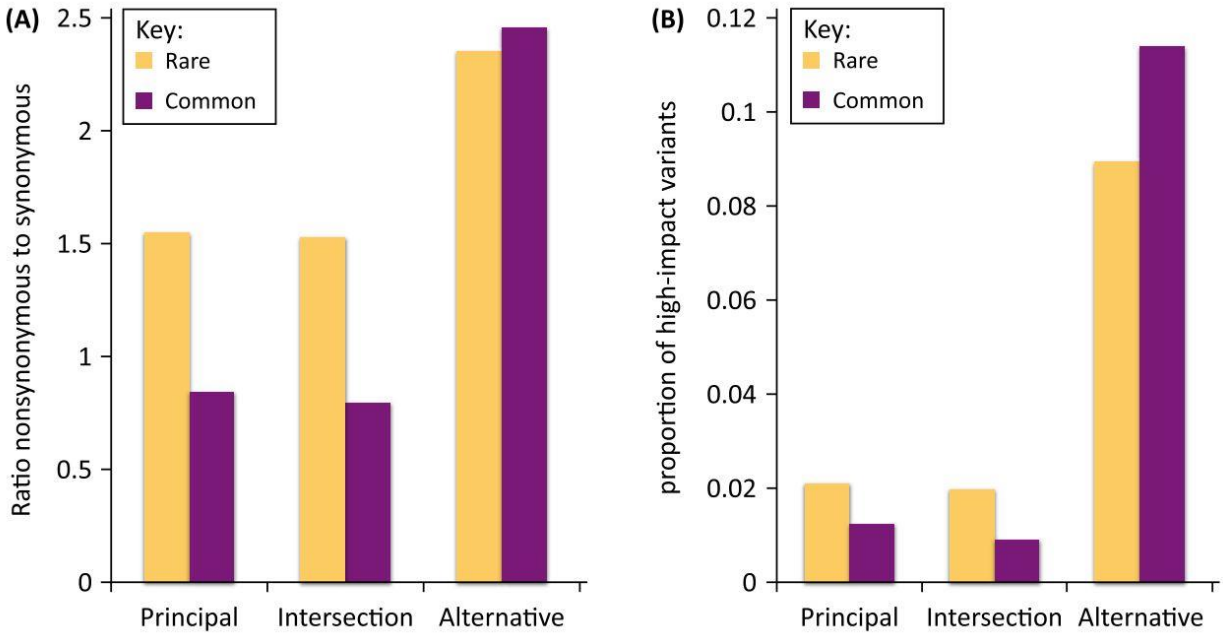


Figure 8. ***Genome-wide Distribution of Sequence Variants in Principal and Alternative Isoforms – figure from (Tress et al., 2017)****. (A) The ratio of non-synonymous to synonymous, and (B) the proportion of high-impact variants shown for three distinct categories using APPRIS (Rodriguez et al.,*

2018). Principal, those protein-coding sites from exons that code for the Principal isoform; Alternative, those protein-coding sites that fall inside exons belonging exclusively to alternative variants; and Intersection, those sites that fall inside exons that code for both principal and alternative isoforms. The variants were subdivided into rare (<0.5%), and common groups for each substitution dataset according to their derived allele frequencies. Each ratio was calculated for both rare and common allele frequencies identified from Phase 3 of the 1000 Genomes Project (Auton & Salcedo, 2015). High-impact variants defined by Variant Effect Predictor (Zerbino et al., 2018) were splice acceptor variants, splice donor variants, stop gains, stop losses, and frameshift variants.

These results indicate that alternative exons are under weaker purifying selection than the APPRIS principal isoforms, and suggest that most alternative exons evolve neutrally. The evidence for purifying selection highlights the importance of principal isoforms and the evidence that many alternative isoforms are evolving neutrally suggests that they have little or no functional applicability as proteins.

Detecting Alternatively Spliced Proteins

Some proteomics studies claim to have found substantially more cases of alternative splicing at the protein level than others. For example, an integrated analysis (Menon *et al.*, 2009) identified 420 distinct alternative isoforms for the mouse genome, of which 92 did not match any previously annotated mouse protein sequence. However, at the time, the mouse genome was not well annotated and the study did not require peptides to identify both constitutive and alternative splice isoforms. Other studies (Kim, M.-S. S. *et al.*, 2014; Wilhelm *et al.*, 2014) of the human genome found more evidence for alternative isoforms because they overestimated the number of reliable peptide identifications (Ezkurdia, Valencia, and Tress *et al.*, 2014). One analysis (Ly *et al.*, 2014) even chose to infer the expression of different isoforms based on peptide abundances in an analogous way to the protocols used for transcript level estimation in RNA-seq studies (Lahens *et al.*, 2014; Steijger *et al.*, 2013). This last form of identifying alternative protein isoforms is wholly inappropriate in proteomics studies because of the low peptide coverage typical of these experiments and because of the non-uniform distribution of the peptides detected.

Other studies generally found lower numbers of alternative protein isoforms, including experiments using human (Ezkurdia *et al.*, 2012; Tanner *et al.*, 2007), mouse (Brosch *et al.*, 2011), rat (Low *et al.*, 2013), *Drosophila* (Tress, Bodenmiller, *et al.*, 2008), *Arabidopsis* (Castellana *et al.*, 2008) and *Aspergillus flavus* (Chang *et al.*, 2010) tissues.

Although most annotated alternative isoforms are not supported by proteomics evidence, patterns emerge. A high proportion of alternative isoforms are generated by swapping one homologous exon for another (Abascal *et al.*, 2015; Ezkurdia *et al.*, 2012). Abascal *et al* compared human and mouse proteomics experiments and found that almost 60% of the orthologous splicing events found across both sets of experiments were homologous exons. The same study found that alternative isoforms generated from homologous exons were highly conserved, implying that they evolved in the ancestor of jawed vertebrates or earlier, at least 460 million years ago.

Analysis of the effect of splice events on Pfam functional domains (El-Gebali *et al.*, 2019) has shown that alternative splicing tends to not affect Pfam domain composition. Only 15% of the alternative splice events detected by Abascal *et al* (Abascal *et al.*, 2015) would damage or cause the loss of a Pfam domain, even though 68% of alternative splice events annotated in CDS regions would break or cause the loss of one or more Pfam domain.

These results suggest that the most damaging alternative splice events do not produce isoforms in quantities that are detectable in standard proteomics experiments. This strongly implies that there is some form of control at the level of translation, or post-translation, that protects the cell against protein isoforms with damaged domains.

Tissue Specific Alternative Splicing at the Protein Level

Many studies have noted tissue specificity at the transcript level. One study (Wang, E. T. *et al.*, 2008) identified over 22,000 tissue-specific alternative transcript events and showed that 47%-65% of alternative events were tissue specific depending on the type of splice event. Meanwhile, another analysis (Gonzàlez-Porta *et al.*, 2013) found that the major transcript varied according to conditions across more than 60% of coding genes.

The tissue-specific rewiring hypothesis is based on the tissue-specific expression of alternative transcripts, the loss of functional domains, and the prevalence of disordered protein regions in alternative isoforms (Colak *et al.*, 2013). In addition, the tissue-specific splice patterns are not always conserved across species. Merkin *et al* found that despite the abundant evidence for tissue specificity of alternative transcripts, patterns of tissue specific alternative splicing were only conserved in a few tissues between mammalian species and birds (Merkin *et al.*, 2012), while Reyes *et al* had similar results across six primate species (A. Reyes *et al.*, 2013). Both studies postulated that the different usage of exons was behind the tissue-specific “rewiring” of protein-protein interaction networks hypothesized by other groups (Buljan *et al.*, 2012; Ghadie *et al.*, 2017) that would be essential for morphological differences between different species.

Results from the large-scale GTEx consortium found that 84% of the variance between tissues was due to gene expression rather than alternative splicing (Melé *et al.*, 2015). A re-analysis of the GTEx data (Alejandro Reyes & Huber, 2018) found that 50% of genes had tissue-specific transcripts, but that most of the tissue-dependent splicing events would not affect proteome complexity of the cell since they involved untranslated exons.

Until now, there has been little research carried out into tissue specific alternative splicing at the protein level. The large-scale proteomics study of 30 human tissues and hematopoietic cells carried out by Kim *et al* (Kim, M.-S. S. *et al.*, 2014) remains the best source of tissue level proteomics data, in part, because it was carried out with replicates. The data from the Kim experiments has been re-analysed on a number of occasions (Kim, M.-S. S. *et al.*, 2014; Lau *et al.*, 2019; Wright, J. C. *et al.*, 2016). The original study highlighted distinct isoforms of *FYN* protein tyrosine kinase in brain and hematopoietic cells, while Wright *et al.* suggested that most tissue-specific alternative splicing was in testis without revealing details. The other two studies detailed evidence for tissue-specific alternative splicing in just a few genes mostly localized to the brain and heart tissues (Abascal *et al.*, 2015) or to heart and testis (Lau *et al.*, 2019).

OBJECTIVES

The human reference gene sets are curated by distinct teams of manual annotators and are in a certain state of flux with protein coding genes constantly added and reclassified. Currently, the human gene set is saturated with alternative splice variants, but the numbers of protein coding transcripts are constantly rising and long-read technologies are predicted to double the number of annotated coding transcripts.

For many of the genes and transcripts annotated as coding a functional role is unclear. Some genes annotated as coding may not actually code for proteins and although alternative splicing has the potential to expand the cellular functional repertoire, there is as yet little evidence to support this theory.

It would seem to be important to distinguish between those genes and transcripts that have functional roles and those that do not. Being able to distinguish which transcripts really do code for functional proteins will allow researchers to determine the real effect of mutations and concentrate on those isoforms of a gene that are predicted to have important cellular effects. Hence, the effort has been concentrated on the following objectives:

1. Improve the performance of each of the core modules in APPRIS principal isoform prediction pipeline.

2. Extend the APPRIS annotations to cover the largest possible number of model species and add new references gene databases, such as RefSeq and the UniProtKB proteome. In addition to updating the annotations for the Ensembl / GENCODE genes of each species stored in the APPRIS database.

3. Bring together the three main human reference databases: Ensembl/GENCODE, RefSeq, and the UniProtKB proteome, with the aim of improving the annotation of the overall gene set. Report the principal isoform for this union of reference sets.

4. Contribute data from APPRIS for the three human genome reference gene sets to help distinguish genes that genuinely code for proteins from potential noncoding genes.

5. Analyze and contrast large-scale proteomics and RNA-seq studies to determine the importance of tissue-specific alternative splicing at the protein level.

6. Determine to what extent alternative splicing is involved in tissue specific rewiring of protein-protein interaction networks, or whether it is responsible for species specific differences.

OBJETIVOS

El conjunto de genes de referencia en el genoma humano es seleccionado por distintos equipos de anotadores manuales y los genes codificantes a proteína se encuentran en un cierto estado de cambio, agregándose y reclasificándose constantemente. Actualmente, el conjunto de genes humanos se encuentra saturado con variantes de empalme alternativas, y se predice que las tecnologías de *long-read* duplicarán el número de transcripciones codificantes.

No está clara la función de muchos de los genes y transcritos anotados como codificantes. Es posible que algunos genes anotados como codificantes, no codifiquen a proteínas y, aunque el empalme alternativo tiene el potencial de expandir el repertorio funcional de la célula, todavía hay poca evidencia para apoyar esta teoría.

Es importante distinguir entre aquellos genes y transcritos que tienen roles funcionales y aquellos que no. Ser capaz de distinguir qué transcritos codifican realmente a proteínas funcionales permitirá a los investigadores determinar el efecto real de las mutaciones, y, por tanto, concentrarse en las isoformas de un gen que se prevé que tengan efectos importantes en la célula. De ahí que el esfuerzo se haya concentrado en los siguientes objetivos:

1. Mejorar el rendimiento de cada uno de los métodos del predictor automático de isoformas principales, APPRIS.

2. Extender las anotaciones de APPRIS para cubrir el mayor número de especies modelo y agregar nuevas referencias a bases de datos de genes, como RefSeq y los proteomas de UniProtKB. Además de actualizar las anotaciones para los genes de Ensembl/GENCODE de cada especie almacenada en la base de datos APPRIS.

3. Unificar las tres bases de datos de referencia humana: Ensembl/GENCODE, RefSeq y el proteoma UniProtKB, con el objetivo de mejorar la anotación de los genes codificantes. Reportar la isoforma principal de esta unión de bases de datos de referencias.

4. Contribuir con datos de APPRIS de las tres bases de datos de referencia del genoma humano para ayudar a distinguir los genes que realmente codifican a proteínas de los posibles genes no codificantes.

5. Analizar y contrastar estudios a gran escala de proteómica y de RNA-seq para determinar la importancia del empalme alternativo específico de tejido a nivel de proteína.

6. Determinar hasta qué punto el empalme alternativo está involucrado en el cableado específico de tejido entre interacción proteína-proteína, o si es responsable de diferencias específicas de especies.

(English) RESULTS AND MATERIAL & METHODS: First article

APPRIS 2017: principal isoforms for multiple gene sets.

The APPRIS Database (http://appris-tools.org) was developed to provide annotations of alternative splice variants (Rodriguez *et al.*, 2013) as part of the GENCODE Consortium (Harrow *et al.*, 2012). The first version of the APPRIS Database deployed a range of computational modules to annotate each isoform with protein structural and functional features, and with data from cross-species alignments.

The main task of APPRIS is to determine a principal splice isoform to represent each gene. Principal isoforms are the variants that maintain the most conserved protein features (Tress *et al.*, 2008). Recently we have demonstrated that these principal isoforms also almost certainly reflect the biological reality of the cell. Independent proteomics evidence demonstrates that most genes have a single main protein isoform and that this isoform is usually the APPRIS principal isoform. We found that the main proteomics isoforms and the APPRIS principal isoforms agreed over 97.8% of comparable genes (Ezkurdia *et al.*, 2015).

In the paper we presented the new developments and the updates since the last release of APPRIS. We altered the annotation pipeline so that it comprised just six modules: Matador3D, *firestar* (G. Lopez *et al.*, 2011), SPADE, CORSAIR, THUMP, and CRASH. Several new features were incorporated to improve the quality and coverage of the predictions. In particular, a second version of Matador3D was developed that makes use of the bit-scores from alignments with PDB (Rose *et al.*, 2017) sequences. Moreover, another version of SPADE that uses the bit-scores from alignments with Pfam (El-Gebali *et al.*, 2019) domains was also developed.

APPRIS was originally devised with the Ensembl/GENCODE (Frankish *et al.*, 2019) human genome in mind. Here, we extended the database to cover ten model species. In addition to the annotations for mouse, pig, rat, and zebra fish that were already in the database, chimpanzee, chicken, cow, and *Drosophila* and *C. elegans* were incorporated in this publication. As well as the Ensembl/GENCODE gene sets, APPRIS now also annotates RefSeq gene sets (O'Leary *et al.*, 2016) and the UniProtKB proteomes (The UniProt Consortium, 2018).

In addition, we created merged gene sets for vertebrate species by cross-referencing the Ensembl/GENCODE, RefSeq and UniProtKB data sets. The cross-reference sets were generated with the data-mining tool BioMart (Smedley *et al.*, 2015). APPRIS now produces a principal isoform for these common reference sets.

APPRIS selects a single CDS variant for each gene as the "PRINCIPAL" isoform based on the annotated protein features. Since this version of APPRIS, principal isoforms have been tagged with the numbers 1 to 5, with 1 being the most reliable. APPRIS determines a most reliable isoform for 75%-95% of annotated protein-coding genes depending on the gene set and the species.

Another novelty in the paper is that where the APPRIS core modules are unable to choose a clear principal variant, the database selects a principal isoform from among the "candidate" isoforms not rejected by the APPRIS core methods. Principal isoforms are selected first via the CCDS identifier (Pruitt *et al.*, 2009) and then on whether all splice junctions are supported by at least one non-suspect mRNA (TSL). CCDS variants are annotated only for the human and mouse genomes, and TSL only for human. Where CCDS and TSL evidence is not decisive or available, APPRIS selects the longest of the candidate isoforms and tags it as PRINCIPAL:5 (P5).

The "candidate" variants not chosen as principal are labeled as “ALTERNATIVE”. These alternative variants are also split into two types, those more likely to be functionally important because they are conserved in at least three tested non-primate species (ALTERNATIVE:1) and those that are not (ALTERNATIVE:2). Non-candidate transcripts are not flagged and are considered as "MINOR" transcripts.

In the human Ensembl/GENCODE gene set, APPRIS determined a PRINCIPAL:1 (P1) isoform for 76.8% of protein-coding genes and just 1.1% of principal isoforms were the longest of the candidate isoforms (P5). A total of 71.5% and 74.3% were tagged with a P1 isoform for the RefSeq genes and UniProtKB proteome, respectively. In the mouse Ensembl/GENCODE genome 82.6% of protein-coding genes were tagged with a P1 isoform. More than 90% of the genes in the Ensembl annotation of vertebrate species had P1 isoforms, though there are fewer genes with multiple coding transcripts in species outside of human.

There were a total of 22,207 protein-coding genes in the union of the Ensembl/GENCODE (release 24), RefSeq (release 107) and UniProtKB (version 201606) human reference sets. Just 5,132 (23.1%) of these genes have a single CDS variant, while APPRIS determined P1 principal isoforms for 9,204 (41.4%) of the remaining genes.

The APPRIS annotations are updated with each new stable Ensembl/GENCODE release, and also periodically for the RefSeq and UniProtKB data sets. The databases behind in each method (PDB, Pfam, non-redundant sequence database, etc.) are also updated to get the most correct annotations. Apart from the APPRIS WebServer (Rodriguez *et al.*, 2015), the annotations are available in the Ensembl web server (Zerbino *et al.*, 2018) and UCSC Genome Browsers (Haeussler *et al.*, 2019).

Michael Tress and Jose Manuel Rodríguez conceived of the presented idea. Jose Manuel Rodríguez developed the theory and performed the computations including the Ensembl/RefSeq/UniProtKB comparison. Juan Rodríguez-Rivas developed the second version of Matador3D. Michael Tress and Jose Manuel Rodríguez verified the analytical methods. Michael Tress encouraged Jose Manuel Rodríguez to investigate and supervised the findings of this work. Both Michael Tress and Jose Manuel Rodríguez authors contributed to the final version of the manuscript. Almost all authors commented on the manuscript.

(Español) RESULTADOS Y MATERIALES Y MÉTODOS: Primer artículo

APPRIS 2017: Isoformas principales en diversas bases de datos de genes.

La base de datos APPRIS (http://appris-tools.org) se desarrolló para proporcionar anotaciones de variantes de empalme alternativas (Rodríguez *et al.*, 2013) como parte del Consorcio GENCODE (Harrow *et al.*, 2012). La primera versión de la base de datos implementó una gama de módulos computacionales para anotar cada isoforma con características estructurales y funcionales de proteínas, y con datos de alineaciones entre especies.

La principal tarea de APPRIS es determinar una isoforma principal que representa cada gen. Las isoformas principales son las variantes que mantienen las características proteicas más conservadas (Tress *et al.*, 2008). Recientemente hemos demostrado que estas isoformas también reflejan, casi con certeza, la realidad biológica de la célula. Estudios independientes de proteómica demuestran que la mayoría de los genes tienen una única isoforma dominante y que esta isoforma suele ser la isoforma principal de APPRIS. Encontramos que las isoformas dominantes de proteómica y las isoformas principales de APPRIS coincidían en un 97,8% de los genes comparables (Ezkurdia *et al.*, 2015).

En el artículo presentamos los nuevos desarrollos y actualizaciones desde la última versión de APPRIS. Mejoramos los métodos de anotación: Matador3D, firestar (G. Lopez *et al.*, 2011), SPADE, CORSAIR, THUMP y CRASH. Se incorporaron varias funciones nuevas para mejorar la calidad y cobertura de las predicciones. En particular, se desarrolló una segunda versión de Matador3D que hace uso de los *bit-scores* del alineador de secuencias contra PDB (Rose *et al.*, 2017). Además, se añadió otra versión de SPADE que también utiliza los *bit-scores* de las alineaciones contra dominios Pfam (El-Gebali *et al.*, 2019).

APPRIS se diseñó originalmente para el proyecto del genoma humano Ensembl/GENCODE (Frankish *et al.*, 2019). Para esta última publicación, ampliamos la base de datos cubriendo diez especies modelo. Además de ratón, cerdo, rata y pez cebra que ya estaban en la base de datos, se incorporaron chimpancés, pollo, vaca y *Drosophila* y *C. elegans*. Además de los genes provenientes de Ensembl/GENCODE, ahora también se anotan los genes de RefSeq (O'Leary *et al.*, 2016) y los proteomas de UniProtKB (The UniProt Consortium, 2018).

Además, creamos una base de datos de genes combinados para especies de vertebrados mediante la referencia cruzada entre los genes de Ensembl/GENCODE, RefSeq y UniProtKB. Las referencias cruzadas se generaron con la herramienta de datos BioMart (Smedley *et al.*, 2015). APPRIS ahora produce una isoforma principal para estos genes de referencia comunes.

APPRIS selecciona una única variante de CDS para cada gen, "PRINCIPAL", basado en las características anotadas de las proteínas. En esta versión, las isoformas principales se han etiquetado con los números del 1 al 5, siendo 1 el más fiable. APPRIS determina la isoforma principal con una fiabilidad entre el 75% -95% de los genes codificantes según la base de datos de referencia y la especie.

Otra novedad es que, cuando los módulos centrales de APPRIS no pueden elegir una isoforma principal clara, la base de datos selecciona una isoforma de entre las isoformas "candidatas". Las isoformas principales se seleccionan primero a través del identificador CCDS (Pruitt *et al.*, 2009) y luego si todas las uniones del empalme alternativo están respaldadas por al menos un ARNm no sospechoso (TSL). Cuando la evidencia de CCDS y

TSL no es decisiva o no está disponible, APPRIS selecciona la más larga de las isoformas candidatas y la etiqueta como PRINCIPAL:5 (P5).

Las variantes "candidatas" que no han sido elegidas como principales, se etiquetan como "ALTERNATIVE". Estas variantes alternativas se dividen en dos tipos, las que tienen más probabilidades de ser funcionalmente importantes porque se conservan en al menos tres especies fuera de primates (ALTERNATIVE:1) y aquellas que no lo son (ALTERNATIVE:2). Las isoformas no candidatas se consideran "MINOR".

Para los genes de Ensembl/GENCODE en humano, APPRIS determinó una isoforma PRINCIPAL:1 (P1) para el 76,8% de los genes codificantes y solo el 1,1% de las isoformas principales eran determinado por las más largas (P5). Un total de 71,5% y 74,3% se etiquetaron con una isoforma P1 para los genes de RefSeq y el proteoma UniProtKB, respectivamente. En el genoma de ratón de Ensembl/GENCODE, el 82,6% de los genes se marcaron con una isoforma P1.

Un total de 22.207 genes humanos se produjo de la unión de Ensembl/GENCODE (versión 24), RefSeq (versión 107) y UniProtKB (versión 201606). Solo 5.132 (23,1%) de estos genes tienen una única variante de CDS, mientras que APPRIS determinó las isoformas principales P1 para 9.204 (41,4%) de los genes restantes.

Las anotaciones de APPRIS se actualizan con cada nueva versión estable de Ensembl/ GENCODE, y también periódicamente para las bases de datos RefSeq y UniProtKB. Las bases de datos detrás de cada método también se actualizan para obtener las anotaciones más correctas. Las anotaciones de APPRIS, a parte del servidor web (Rodríguez *et al.*, 2015), están disponibles en la web de Ensembl (Zerbino *et al.*, 2018) y en UCSC Genome Browsers (Haeussler *et al.*, 2019).

Michael Tress y Jose Manuel Rodríguez concibieron la idea presentada. JMR desarrolló la teoría y realizó los cálculos, incluida la comparación Ensembl/RefSeq/UniProtKB. Juan Rodríguez-Rivas desarrolló la segunda versión de Matador3D. MT y JMR verificaron los métodos analíticos. MT animó a JMR a investigar y supervisó los hallazgos de este trabajo. MT y JMR contribuyeron a la versión final del manuscrito. Casi todos los autores comentaron sobre el manuscrito.

Published online 23 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D213–D217***
doi: 10.1093/nar/gkx997

APPRIS 2017: principal isoforms for multiple gene sets

Jose Manuel Rodriguez[1,*], Juan Rodriguez-Rivas[2], Tomás Di Domenico[2], Jesús Vázquez[3,4], Alfonso Valencia[5,6] and Michael L. Tress[2,*]

[1]Spanish National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain, [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain, [3]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28029 Madrid, Spain, [4]CIBER de Enfermedades Cardiovasculares (CIBERCV), 28029 Madrid, Spain, [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona E-08010, Spain and [6]Life Sciences Department, Barcelona Supercomputing Centre (BSC-CNS), Barcelona E-08034, Spain

Received September 15, 2017; Revised October 10, 2017; Editorial Decision October 11, 2017; Accepted October 19, 2017

ABSTRACT

The APPRIS database (http://appris-tools.org) uses protein structural and functional features and information from cross-species conservation to annotate splice isoforms in protein-coding genes. APPRIS selects a single protein isoform, the 'principal' isoform, as the reference for each gene based on these annotations. A single main splice isoform reflects the biological reality for most protein coding genes and APPRIS principal isoforms are the best predictors of these main proteins isoforms. Here, we present the updates to the database, new developments that include the addition of three new species (chimpanzee, *Drosophila melangaster* and *Caenorhabditis elegans*), the expansion of APPRIS to cover the RefSeq gene set and the UniProtKB proteome for six species and refinements in the core methods that make up the annotation pipeline. In addition APPRIS now provides a measure of reliability for individual principal isoforms and updates with each release of the GENCODE/Ensembl and RefSeq reference sets. The individual GENCODE/Ensembl, RefSeq and UniProtKB reference gene sets for six organisms have been merged to produce common sets of splice variants.

INTRODUCTION

It has been estimated that 95% of multi-exon human genes produce alternatively spliced messenger RNA (1,2) transcripts. These alternative transcripts, if translated, would generate a range of alternative proteins that are often strikingly different from the constitutive gene product and that would add to the repertoire of cellular functions (3,4). However, the cellular role of alternative splicing is a controversial topic (5–7) and the functional importance of any potential alternative protein isoforms is an open question (7).

APPRIS (8,9) was developed within the GENCODE (10) consortium to cope with the challenge of annotating alternatively spliced protein-coding transcripts with functional information. The database employs a series of modules to map protein structure and functional features and cross-species conservation to all reference splice isoforms. Unlike the other maintained databases that annotate alternative splice isoforms with functional information (11,12), APPRIS concentrates only on the most reliably predicted features, including the presence of Pfam domains (13) and highly conserved functional residues (14).

Information from APPRIS is fed back to the GENCODE manual annotators to inform gene models. However, the main role of APPRIS is the annotation of a main (principal) isoform for individual coding genes (15). APPRIS selects principal isoforms based on the presence or absence of evolutionary evidence such as conserved functional and structural motifs. Principal isoforms are those with the most preserved structural and functional features and those with the greatest cross species conservation, while alternative isoforms often have non-conserved exons and structure or function features that are damaged or missing (15). APPRIS core modules almost always agree on the principal isoform.

Historically researchers and annotators have had to resort to choosing the longest annotated CDS as the reference

*To whom correspondence should be addressed. Tel: +34 91 732 80 00; Fax: +34 91 224 69 76; Email: mtress@cnio.es
Correspondence may also be addressed to Jose Manuel Rodriguez. Tel: +34 914 531 200; Fax: +34 914 531 265; Email: jmrodriguezc@cnic.es
Present addresses:
Jose Manuel Rodriguez, Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28029 Madrid, Spain.
Juan Rodriguez-Rivas, Barcelona Supercomputing Centre (BSC), Barcelona 08034, Spain.

variant for individual coding genes (16). We have shown that this simple solution is not ideal (17) and as databases expand and more alternative transcripts are annotated the fix will become less viable. Highest Connected Isoforms (18), from RNAseq data and from protein–protein interaction and structural information have been proposed as an answer, but we have shown that these isoforms agree with the main functional isoform less often than the longest isoforms (5). The APPRIS principal isoforms are the splice variants that best represent the gene (17).

Although APPRIS was developed for use with GENCODE/Ensembl annotations (10,19), there are other manually annotated reference sets, in particular RefSeq (RefSeqGene) (20) and UniProtKB (16). The RefSeq, GENCODE/Ensembl and UniProtKB annotations are not identical and many gene models or predicted proteins are present in one or more reference sets, but not in others. For that reason we have extended APPRIS to the RefSeqGene and UniProtKB annotations in the case of vertebrate genomes (human, mouse, zebra-fish, rat, pig and chimpanzee). In addition, we have made improvements to core methods in the APPRIS pipeline, implemented the UCSC Track Hub to enhance annotation access and created Docker images to help execute the annotation pipeline.

THE DATABASE

APPRIS annotates splice isoforms with protein structural and functional features, and data from cross-species alignments. The database uses these features to select a single reference isoform for each protein-coding gene, here termed the principal isoform. This principal isoform has the most conserved protein features and the most evidence of cross-species conservation. At the same time isoforms that have lost conserved protein features or do not have cross-species conservation are flagged as alternative.

Currently the APPRIS annotation pipeline comprises six modules (9). Matador3D detects similarity to structural homologs in the PDB (21); *firestar* (14) predicts functionally important amino acid residues; SPADE identifies Pfam functional domains via the PfamScan algorithm (13); CORSAIR carries out BLAST (22) searches against vertebrate protein sequences to determine the number of orthologs that align correctly and without gaps; THUMP makes unanimous predictions of trans-membrane helices from three predictors (23–25); and CRASH predicts the presence and location of signal peptides using the SignalP and TargetP programs (26,27). APPRIS maps protein features to all coding transcripts. The databases implied in each method (PDB, Pfam, non-redundant sequence database, etc.) are updated periodically to get most correct annotations.

Refinements to core methods

All modules in APPRIS are continually revised against the GENCODE annotation of the human reference gene set. As a result we have been able to improve the performance of each of the core modules in APPRIS. The gold standard set for principal isoforms are those genes with just one CCDS variant (consensus coding sequence, 28). Tests have shown that unique CCDS variants (transcripts annotated consistently by RefSeq and Ensembl/GENCODE manual annotators) and APPRIS principal isoforms are both highly reliable predictors of the dominant cellular isoform, so they should select the same reference isoform for the vast majority of genes.

Comparison between unique CCDS isoforms for each gene and those selected by the individual APPRIS modules shows that there is almost complete agreement between the two, both at the time of the initial database publication (8) and with the current version of APPRIS. The more recent APPRIS principal isoforms disagree with the unique CCDS isoforms less often and with the exception of CORSAIR (98.92%), all methods have more than 99% agreement with unique CCDS variants (Supplementary Figure S1).

Reliability scores

Many experiments require every studied gene to have a single representative, so APPRIS now automatically selects a principal isoform for every single coding gene. However, not all APPRIS principal isoforms are alike. Principal isoforms are tagged with a score from 1 to 5 depending on the reliability of the selection, with 1 being the most reliable. 'PRINCIPAL:1' isoforms are determined solely using information from the APPRIS core modules. For those genes where the modules cannot make a unique selection, APPRIS uses external data such as the CCDS annotation and the GENCODE Consortium Transcript Support Level (29). Where all else fails, the longest not previously rejected isoform is selected as the principal ('PRINCIPAL:5'). Splice variants rejected as principal isoforms by the APPRIS core modules are labeled as 'MINOR', while those variants not rejected by the core modules, but rejected using external information are labeled as 'ALTERNATIVE' (for more details on the reliability scores see the Supplementary Data).

Additional features

Annotations are stored in a MySQL relational database and these can be downloaded via the APPRIS web site. The human and mouse annotations are available through GENCODE, and Ensembl exports APPRIS principal isoforms of human, mouse, zebra-fish, rat and pig within its website, BioMart data-mining tool and API. Furthermore, APPRIS annotations can be visualized in the UCSC Genome Browser (30) from its own Public Track Hub. In addition, users can extract APPRIS annotations for specific reference sets (Ensembl, RefSeqGene, UniProt) via the APPRIS WebServer and WebServices (9). All the APPRIS source code is available in a GitHub public-repository (https://github.com/appris/appris/) offering a distributed version control.

The APPRIS pipeline is executed on the Linux (Ubuntu) system but it can be run on Windows, Mac OS X or Unix-based systems using the Docker image (appris/core) provided by the software container platform, Docker (http://www.docker.com). The APPRIS-Docker image is stored in the public Docker Hub (https://hub.docker.com/).

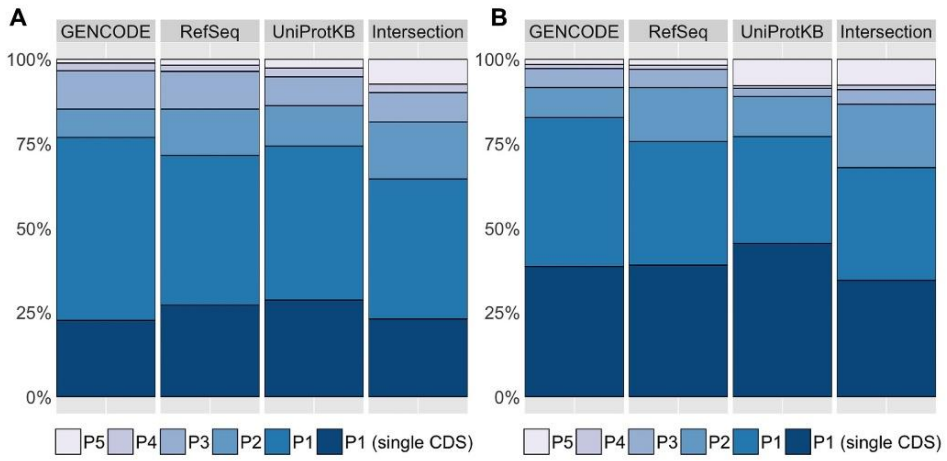


Figure 1. Bar-plots with the percentage of genes identified with the final annotations of APPRIS for the human (**A**) and mouse (**B**) species house in database. APPRIS identifies a principal isoform (Pn) for each gene that are tagged with numbers from 1 to 5, with 1 being the most reliable. Isoforms in genes with a unique protein representative (single CDS) are automatically categorized as P1. The APPRIS Database annotates the protein-coding genes in all public sets GENCODE, RefSeq and UniProtKB. In addition, we established a common gene set (Intersection) with the GENCODE, RefSeq, and UniProtKB reference sets.

A

Ensembl Id.	RefSeq Id.	UniProt Id.	Length (aa)	No. Functional Residues	3D Structure Score	Domain Score	Conservation Score
-	NM_001291784	-	148	8	93.6	72.6	0
-	XM_011515405	-	208	8	93.4	72.6	0
ENST00000614228	-	A4D1Z8	144	8	109.1	86.8	5.8

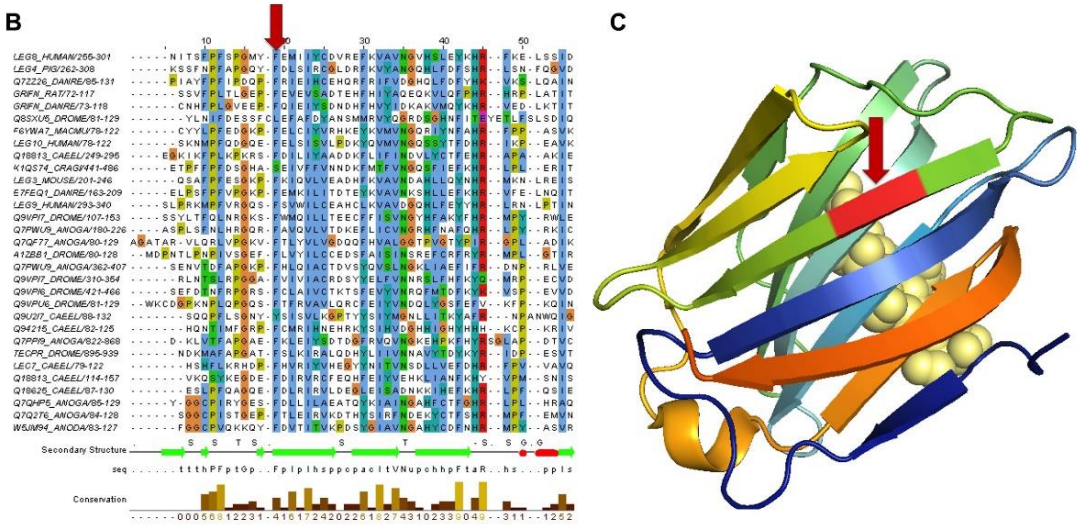


Figure 2. APPRIS annotations for gene *GRIFIN*. (**A**) APPRIS results for the three protein-coding variants from the gene reference sets, GENCODE (Ensembl), RefSeq and UniProtKB. APPRIS chooses isoform ENST00000614228+A4D1Z8 as the principal isoform (highlighted in green), which belongs to Ensembl and UniProtKB. A selection based on the 3D structure, the functional domains and the conservation in related species. (**B**) Alignment for a section of the Pfam galectin family of proteins. The red arrow shows where 8 extra residues in the RefSeq variants would disrupt a region of the galectin functional domain of *GRIFIN*. (**C**) The 3D structure of 4LBJ (human galectin-3 CRD) that has 29% identity with variants ENST00000614228+A4D1Z8. The galectins are a family of proteins defined by their binding specificity for β-galactoside sugars (displayed in light yellow spheres). The red arrow shows where the 8 extra residues would have to insert into the structure, breaking a β-sheet.

New APPRIS annotations

The APPRIS database has increased in size to cover six different vertebrate genomes (human, mouse, rat, pig, zebra-fish, chimpanzee), and two invertebrate genomes (*Drosophila* and *Caenorhabditis elegans*). The human GENCODE gene set (release 24) recognizes 20 250 protein-coding genes and APPRIS determines a P1 isoform for 76.8% of these genes (see Figure 1A). The mouse GENCODE gene set (release M12) has 22 538 protein-coding genes and 82.6% are tagged with a P1 isoform (see Figure 1B). More than 90% of the genes in the Ensembl annotation of three vertebrate species (rat, pig, chimpanzee) have P1 isoforms. The number is higher for these species because the majority of genes have a unique CDS (see Supplementary Figure S2).

RefSeqGene and UniProtKB annotations

APPRIS has now been extended to the other main public genome annotation, RefSeqGene and to the UniProtKB proteome. RefSeqGene human (release 107) currently houses 20 066 protein-coding genes, while the UniProtKB human proteome (release 2016_06) has 21 608 genes. APPRIS identifies a P1 principal isoform for 71.5% of genes in the RefSeqGene set, and 74.3% of genes in the UniProtKB proteome (see Figure 2A).

The pipeline of annotations in the RefSeqGene set is identical to that of GENCODE/Ensembl, but two of the modules used in the pipeline (Matador3D, and CORSAIR) have had to be modified for use with the UniProtKB and Intersection (see below) gene sets because the original versions of these modules made use of genomic coordinates.

Intersection gene sets

We have also created merged gene sets for vertebrate species by cross-referencing the GENCODE/Ensembl, RefSeqGene and UniProtKB reference sets. For the human genome we established a common gene set (Intersection) with the GENCODE (release 24), RefSeqGene (release 107) and UniProtKB (version 2016_06) reference sets. The initial cross-reference was generated with the data-mining tool, BioMart (31) and from there we re-annotated the cross-database relationships manually. For the remaining species we generated common gene sets with the BioMart tool, although these relationships are not yet manually annotated. The version and the number of genes for each reference set are shown in Supplementary Table S1.

There were a total of 22 207 protein-coding genes in the human intersection reference set composed of GENCODE (release 24), RefSeqGene (release 107) and UniProtKB (version 2016_06) genes. Just 5132 (23.1%) of these genes have a single CDS variant, while APPRIS determined P1 principal isoforms for 9204 (41.4%) of the genes (see Figure 1A).

The merged Intersection gene set allows us to identify principal isoforms missing in the individual gene sets. For example the principal isoform from the merged set for *GRIFIN* is annotated in GENCODE (ENST00000614228) and UniProtKB (A4D1Z8), but not in RefSeqGene (See Figure 2). This principal isoform is chosen because it maps better to known 3D structures, has an unbroken Pfam domain and has orthologous sequences in vertebrate species. In contrast, the domain in the RefSeqGene isoforms is broken and neither isoform has cross-species conservation. The 8-residue insertion in the two RefSeqGene variants breaks a Pfam functional domain (Figure 2B) and 3D structure (Figure 2C). The C-terminal extension in the GENCODE/Ensembl/UniProtKB principal isoform (but not in the RefSeqGene variants) is also established in mammals (see Supplementary Figure S3).

DISCUSSION

APPRIS annotate alternatively spliced protein isoforms with protein structural and functional information and cross-species conservation using a range of computational prediction methods. It also selects one of these isoforms to be the representative protein sequence for each coding gene.

We have shown that a single representative protein reflects the biological reality of the cell: most coding genes have a single dominant protein isoform (5,7,17) and this seems to be true regardless of cell type (17). This dominant protein isoform is almost always the APPRIS principal isoform: APPRIS principal isoforms overwhelmingly coincide with the manually annotated unique CCDS variants and with the main isoforms detected in large-scale proteomics experiments (17). In fact where dominant isoforms could be determined for all three methods, the agreement was 99.5% (17). Further corroboration of the importance of APPRIS principal isoforms comes from large-scale genetic variation studies, which show that exons from principal isoforms are under purifying selection. By way of contrast alternative exons are under neutral selection (5,32).

APPRIS principal isoforms have a wide range of uses. Designating a single alternative splice variant as principal is an important technical issue and is a critical first step for any genome-wide analysis. Large-scale analyses are highly dependent on the quality of input data; so principal isoforms should improve the reliability of these experiments. Determining whether an exon belongs to a principal or alternative variant is key in biomedical studies. APPRIS principal isoforms can also be useful when working with individual genes; since it is not always clear which splice isoform (or isoforms) is functionally important.

APPRIS principal isoforms and annotations are freely accessible to all via the APPRIS web page, via the APPRIS WebServices (9), and the Ensembl reference annotations for individual species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health [U41 HG007234, 2U41 HG007234]; Spanish Ministry of Economics and Competitiveness [BIO2015-67580-P]; Spanish National Institute of Bioinformatics (www.inab.org) [INB-ISCIII, PRB2 to J.M.R.]; ProteoRed [IPT13/0001-ISCIII-SGEFI/FEDER to J.V.]; Joint BSC-IRB-CRG Program in Computational

Biology and Award Severo Ochoa [SEV 2015-0493 to A.V.]. Funding for open access charge: U.S. Department of Health and Human Services; National Institutes of Health; National Human Genome Research Institute [2U41 HG007234].
Conflict of interest statement. None declared.

REFERENCES

1. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
2. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
3. Smith,C.W. and Valcárcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
4. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J-J., Yeats,C., Olason,P.I., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 5495–5500.
5. Tress,M.L., Abascal,F. and Valencia,A. (2017) Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.*, **42**, 98–110.
6. Blencowe,B.J. (2017) The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.*, **42**, 407–408.
7. Tress,M.L., Abascal,F. and Valencia,A. (2017) Most alternative isoforms are not functionally important. *Trends Biochem. Sci.*, **42**, 408–410.
8. Rodriguez,J.M., Maietta,P., Ezkurdia,I., Pietrelli,A., Wesselink,J.J., Lopez,G., Valencia,A. and Tress,M.L. (2013) APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.*, **41**, D110–D117.
9. Rodriguez,J.M., Carro,A., Valencia,A. and Tress,M.L. (2015) APPRIS WebServer and WebServices. *Nucleic Acids Res.*, **43**, W455–W459.
10. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinsk,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
11. Shionyu,M., Yamaguchi,A., Shinoda,K., Takahashi,K. and Go,M. (2009) AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res.*, **37**, D305–D309.
12. Martelli,P.L., D'Antonio,M., Bonizzoni,P., Castrignanò,T., D'Erchia,A.M., D'Onorio De Meo,P., Fariselli,P., Finelli,M., Licciulli,F. *et al.* (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res.*, **39**, D80–D85.
13. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
14. Lopez,G., Maietta,P., Rodriguez,J.M., Valencia,A. and Tress,M.L. (2011) firestar–advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
15. Tress,M.L., Wesselink,J-J., Frankish,A., López,G., Goldman,N., Löytynoja,A., Massinghamm,T., Pardi,F., Whelan,S., Harrow,J. and Valencia,A. (2008) Determination and validation of principal gene products. *Bioinformatics*, **24**, 11–17.
16. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
17. Ezkurdia,I., Rodriguez,J.M., Carrillo-de Santa Pau,E., Vázquez,J., Valencia,A. and Tress,M.L. (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, **14**, 1880–1887.
18. Li,H.D., Menon,R., Govindarajoo,B., Panwar,B., Zhang,Y., Omenn,G.S. and Guan,Y. (2015) Functional networks of highest-connected splice isoforms: From the chromosome 17 human proteome project. *J. Proteome Res.*, **14**, 3484–3491.
19. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., García Girón,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database (Oxford)*, **2016**, baw093.
20. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
21. Rose,P.W., Prlić,A., Altunkaya,A., Bi,C., Bradley,A.R., Christie,C.H., Costanzo,L.D., Duarte,J.M., Dutta,S., Feng,Z. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.
22. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.
24. Käll,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
25. Viklund,H and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.
26. Petersen,T.N., Brunak,S., von Heijne,G. and Nielsen,H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.
27. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
28. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
29. Ezkurdia,I., Juan,D., Rodriguez,J.M., Frankish,A., Diekhans,M., Harrow,J., Vazquez,J., Valencia,A. and Tress,M.L. (2014) Multiple evidence strands suggest that there may be as few as 19, 000 human protein-coding genes. *Hum. Mol. Genet.*, **23**, 5866–5878.
30. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D, Gonzalez,J.N., Guruvadoo,L *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
31. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
32. Liu,T. and Lin,K. (2015) The distribution pattern of genetic variation in the transcript isoforms of the alternatively spliced protein-coding genes in the human genome. *Mol. Biosyst.*, **11**, 1378–1388.

(English) RESULTS AND MATERIAL & METHODS: Second article - collaboration

Loose ends: almost one in five human genes still have unresolved coding status

Here, we expanded on a previous analysis (Ezkurdia *et al.*, 2014) to evaluate protein-coding genes from the three main human reference gene sets, Ensembl/GENCODE, RefSeq and UniProtKB. In the versions used in the analysis, RefSeq (v107) annotated 20,450 coding genes, and Ensembl/GENCODE (v24/e83) contained 20,266 coding genes. The UniProtKB proteome (June 2016), which is based around proteins rather than genes, mapped to 21,212 coding genes. There were 19,446 genes annotated as coding in the intersection of the three sets. Beyond the intersection of the three reference databases, 851 genes were supported by two of the three reference sets and 1,903 genes were annotated in just one of the three reference sets. In total, the manual curation of the three reference sets found 22,210 protein-coding genes.

We defined a set of 16 features that we had previously shown were good indicators of misannotated coding genes and used these to separate “potential non-coding” (PNC) genes from likely coding (LC) genes. 4,234 of the 22,210 coding genes annotated in the union of the three sets were tagged with at least one of the potential non-coding features, including almost all genes outside the intersection of the three reference sets and including 2,278 (11%) of Ensembl/GENCODE coding genes. Even within the intersection of the three sets 1,471 genes, 7.6% of genes agreed on by all three reference sets, were tagged as PNC.

To test the hypothesis that many of the genes outside of the intersection may not code for proteins under normal cellular conditions, we analyzed the experimental expression of potential non-coding genes using available experimental transcriptomics, proteomic and antibody binding data and compared this to the data for likely coding genes.

We downloaded RNA expression data from the Human Protein Atlas (Uhlén et al, 2015). The Human Protein Atlas details RNA-seq experiments carried out on 36 tissues using Ensembl/GENCODE. We binned genes by maximum expression and by number of tissues and compared the tissue distributions of LC genes and PNC genes. We found considerably more evidence for the transcript expression of LC genes. In fact, while 73.5% of LC genes had a maximum TPM (transcript per million) of 20 or more, just 24.3% of PNC genes reached this level of transcription.

For protein expression, we collected peptides to identify gene products from the human PeptideAtlas (Desiere, 2006) proteomics database. In addition, we downloaded antibody-specific information to validate tissue-specific protein expression from the Human Protein Atlas (Uhlén *et al.*, 2015). For genes annotated by GENCODE we detected a higher proportion of peptides for the genes in the likely coding set (13,360 of 17,988, 74.3%). PNC genes had little peptide support (just 142 of 2,278, 6.2%). Additionally, antibodies produced similar results. LC genes were detected in higher numbers with antibodies (9,896 of 17,988 genes, 55%) than genes in the potential non-coding set (79 of the 2,278 genes, 3.5%).

We also evaluated detected peptides for those genes that had RNA-seq expression in at least 10 tissues with a minimum of 10 TPM. Considering just those genes with this level of RNA-seq expression, we detected peptides for 85.6% of LC genes. Even for these well expressed genes, we detected peptides for just 19.4% of those PNC genes annotated in all three sets, and just 6.1% of PNC genes annotated in two or fewer sets.

Moreover, we compared rates of genetic variation for genes with potential non-coding features and for LC genes using data from the 1,000 Genomes Project (Altshuler *et al.*, 2012). We calculated the percentage of high-impact variants and the ratio of non-synonymous to synonymous variants for rare and common allele frequencies. Whole genome copy number variation (CNV) maps were also downloaded from five different publications (Abyzov *et al.*, 2015; Handsaker *et al.*, 2015; Sudmant, Mallick, *et al.*, 2015; Sudmant, Rausch, *et al.*, 2015; Zarrei *et al.*, 2015).

For LC genes, the percentage of high impact variants was 1.88 at rare allele frequencies against 0.61 for common alleles. PNC genes had proportionally more high impact variants; for those PNC genes annotated in all three sets 3.72% of mutations at rare allele frequencies and 2.16% of those at common allele frequencies were high impact mutations. PNC genes also had higher synonymous to non-synonymous ratios with little difference between common and rare allele frequencies, while LC genes had lower synonymous to non-synonymous at common allele frequencies than at rare allele frequencies, as would be expected if they were under selective pressure.

All these results suggest that many of the 4,234 genes that are annotated as coding in at least one of the reference databases but tagged as potential non-coding in our study. In fact, these annotated coding genes may be pseudogenes, non-coding genes or other artifacts rather than code for functional proteins.

Michael Tress and Federico Abascal conceived and planned the presented idea. Federico Abascal generated the human genetic variation data. David Juan provided the gene family data. Irwin Jungreis provided the PhyloCSF data. Laura Martinez generated the potential non-coding gene pipeline. Maria Rigau provided the CNV data. Jose Manuel Rodriguez generated the combined RefSeq, UniProtKB and Ensembl/GENCODE reference set and provided the data from APPRIS. Michael Tress conceived, designed, analysed and interpreted the data. MT and FA were involved in drafting and revising the manuscript. DJ, IJ, JMR and MR were involved in revising the manuscript.

(Español) RESULTADOS Y MATERIALES Y MÉTODOS: Segundo artículo - colaboración

Extremos sueltos: casi uno de cada cinco genes humanos aún tiene un estado de codificación no resuelto

En este estudio, ampliamos un análisis anterior (Ezkurdia *et al.*, 2014) para evaluar que genes en humano codifican a proteínas basado en la unión de las bases de datos de referencia, Ensembl/GENCODE, RefSeq y el proteoma de UniProtKB. En las versiones usadas en este análisis, RefSeq (v107) anotó 20.450 genes y Ensembl/GENCODE (v24/e83) contenía 20.266 genes. El proteoma UniProtKB (junio 2016), que se basa en proteínas en lugar de genes, mapeaba a 21.212 genes. Había 19.446 genes anotados como codificantes en la intersección de los tres conjuntos. Más allá de la intersección, 851 genes estaban respaldados por dos de los tres conjuntos y 1.903 estaban anotados en uno solo de los conjuntos. En total, y después de una validación manual, se encontró 22.210 genes codificantes a proteínas.

Definimos un conjunto de 16 características que previamente habíamos demostrado que eran buenos indicadores de genes codificantes mal anotados, y los usamos para separar los genes “potencialmente no codificantes" (PNC) de los genes “probablemente codificantes” (LC, *likely coding*). 4.234 de los 22.210 genes se etiquetaron con al menos una de las posibles características no codificantes y 2.278 (11%) genes provenientes de Ensembl/GENCODE. Incluso dentro de la intersección de las tres bases de datos, se etiquetaron 1.471 genes como PNC. Esto es el 7,6% de los genes acordados por los tres conjuntos de referencia.

Para probar la hipótesis de que muchos de los genes fuera de la intersección pueden no codificar proteínas en condiciones celulares normales, analizamos la expresión de estos genes utilizando datos de experimentos proteómicos, de anticuerpos y transcriptómicos, y lo comparamos con los genes codificantes probables.

Descargamos datos de expresión de ARN del *Human Proteome Atlas* (Uhlén et al, 2015). Estos datos provenían de experimentos *RNA-seq* llevados a cabo en 36 tejidos. Agrupamos los genes por expresión máxima y por número de tejidos, y comparamos las distribuciones de los tejidos entre genes LC y genes PNC. Encontramos una evidencia considerablemente mayor de la expresión en genes LC. De hecho, el 73,5% de los genes probablemente codificantes tenían un TPM máximo (transcripción por millón) de 20 o más, frente a solo el 24,3% de los genes PNC.

Para la expresión de proteínas, recolectamos péptidos provenientes de la base de datos de proteómica, PeptideAtlas (Desiere, 2006). Además, descargamos del *Human Proteome Atlas* (Uhlén *et al.*, 2015) información específica de anticuerpos para validar la expresión de proteínas específicas en tejido. Para genes anotados por Ensembl/GENCODE detectamos una mayor proporción de péptidos para los genes LC (13.360 de 17.988, 74,3%). Los genes PNC tenían poco soporte de péptidos (solo 142 de 2.278, 6.2%). Por otro lado, los anticuerpos produjeron resultados similares. Se detectaron mayor número de anticuerpos en los genes LC (9.896 de 17.988 genes, 55%) que los genes PNC (79 de los 2.278 genes, 3,5%).

También detectamos péptidos para aquellos genes que tienen expresión de RNA-seq en al menos 10 tejidos con un mínimo de 10 TPM. Considerando sólo aquellos genes con este nivel de expresión en RNA-seq, detectamos péptidos para el 85,6% de los genes LC. Sin embargo, detectamos péptidos para solo el 19,4% de los genes PNC anotados en los tres

conjuntos, y para el 6,1% de los genes PNC anotados en dos o menos conjuntos de referencia.

Además, comparamos las tasas de variación genética de genes PNC con las tasas de variación genética provenientes del Proyecto 1,000 Genomas (Altshuler *et al.*, 2012). Calculamos el porcentaje de variantes de alto impacto y la proporción de variantes no sinónimas y sinónimas en frecuencias alélicas raras y comunes. También se descargaron mapas de variación del número de copias (CNV) del genoma completo de cinco publicaciones diferentes (Abyzov *et al.*, 2015; Handsaker *et al.*, 2015; Sudmant, Mallick, *et al.*, 2015; Sudmant, Rausch, *et al.*, 2015 ; Zarrei *et al.*, 2015).

Para los genes LC, el porcentaje de variantes de alto impacto fue 1,88 en frecuencias de alelos raros contra 0,61 para los alelos comunes. Los genes PNC tenían proporcionalmente más variantes de alto impacto (anotados en las tres bases de datos), el 3,72% de las mutaciones en frecuencias alélicas raras y el 2,16% de aquellas en frecuencias alélicas comunes. Los genes PNC también tenían proporciones de sinónimos y no sinónimos más altas con poca diferencia entre las frecuencias alélicas comunes y raras, mientras que los genes LC tenían menos sinónimos o no sinónimos en las frecuencias de los alelos comunes que en las frecuencias de los alelos raros, como sería de esperar si estuvieran bajo selectividad presión.

Todos estos resultados sugieren que muchos de los 4.234 genes que están anotados como codificantes a proteína, en al menos una de las bases de datos de referencia, las llegamos a etiquetar como potenciales no codificantes. De hecho, el 19,1% de todos los genes codificantes anotados pueden ser pseudogenes, genes no codificantes o artefactos en lugar de codificar proteínas funcionales.

Michael Tress y Federico Abascal concibieron y planificaron la idea presentada. Federico Abascal generó los datos de variación genética humana. David Juan proporcionó los datos de la familia genética. Irwin Jungreis proporcionó los datos de PhyloCSF. Laura Martínez generó el flujo de trabajo para obtener los genes PNC. Maria Rigau proporcionó los datos de la CNV. José Manuel Rodríguez generó la combinación/unión de las bases de datos de referencia RefSeq, UniProtKB y Ensembl/GENCODE y proporcionó las anotaciones de APPRIS. Michael Tress concibió, diseñó, analizó e interpretó los datos. MT y FA participaron en la redacción y revisión del manuscrito. DJ, IJ, JMR y MR participaron en la revisión del manuscrito.

7070–7084 *Nucleic Acids Research, 2018, Vol. 46, No. 14* *Published online 30 June 2018*
doi: 10.1093/nar/gky587

Loose ends: almost one in five human genes still have unresolved coding status

Federico Abascal[1], David Juan[2], Irwin Jungreis[3], Laura Martinez[4], Maria Rigau[5], Jose Manuel Rodriguez[6], Jesus Vazquez[6] and Michael L. Tress[4,*]

[1]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, UK, [2]Comparative Genomics Lab, Instituto de Biologica Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain, [3]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA and Broad Institute of MIT and Harvard, Cambridge, MA, USA, [4]Bioinformatics Unit, Spanish National Cancer Research Centre, Madrid, Spain, [5]Computational Biology Life Sciences Group, Barcelona Supercomputing Center, Barcelona, Spain and [6]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain

Received May 15, 2018; Revised June 12, 2018; Editorial Decision June 13, 2018; Accepted June 18, 2018

ABSTRACT

Seventeen years after the sequencing of the human genome, the human proteome is still under revision. One in eight of the 22 210 coding genes listed by the Ensembl/GENCODE, RefSeq and UniProtKB reference databases are annotated differently across the three sets. We have carried out an in-depth investigation on the 2764 genes classified as coding by one or more sets of manual curators and not coding by others. Data from large-scale genetic variation analyses suggests that most are not under protein-like purifying selection and so are unlikely to code for functional proteins. A further 1470 genes annotated as coding in all three reference sets have characteristics that are typical of non-coding genes or pseudogenes. These potential non-coding genes also appear to be undergoing neutral evolution and have considerably less supporting transcript and protein evidence than other coding genes. We believe that the three reference databases currently overestimate the number of human coding genes by at least 2000, complicating and adding noise to large-scale biomedical experiments. Determining which potential non-coding genes do not code for proteins is a difficult but vitally important task since the human reference proteome is a fundamental pillar of most basic research and supports almost all large-scale biomedical projects.

INTRODUCTION

Before the human genome was sequenced, most researchers estimated that human protein coding gene numbers would be between 25 000 and 40 000 (1), with some estimates closer to 100 000 genes (2,3). However, the accumulation of experimental data has progressively brought this estimate down. The 'finished' version of the human genome revised the estimates to between 20 000 and 25 000 coding genes (4).

The gradual downward trend of the human protein gene count has been mirrored in the reference human gene set. The annotation of human coding genes began with the Ensembl project (5) and the initial release included more than 24 000 coding genes. This number soon decreased to 22 000 as the genome assembly improved and automatic predictions were refined (6). Until recently there were still gene loci in the reference set defined as coding based on the initial automatic predictions, and a number of these had little support as coding genes beyond their initial prediction. After the merge with the GENCODE manual annotations (7) in 2009, 1004 poorly supported automatic annotation models were removed from the Ensembl annotation set.

These refinements and intensive manual annotation have brought the number of annotated protein coding genes down to slightly over 20 000 genes in the Ensembl/GENCODE (7,8) reference, and indeed the three maintained manual reference databases, Ensembl/GENCODE, RefSeq (9) and UniProtKB (10), have converged on similar numbers of protein coding genes [f1000research: doi: 10.12688/f1000research.11119.1], a number that is in line with the prediction by Clamp *et al.* using evolutionary comparisons (11).

However, the human gene sets are in certain state of flux with coding genes being added and reclassified with each new release, and it is important to note that these 20 000 plus coding genes are not the same in each database. Indeed, as we show in this paper, the number of annotated coding genes in the union of the three reference sets exceeds 22 000.

The task of manually inspecting >20 000 annotated coding genes is enormous and the process has taken many years (7). Manual annotators have to accomplish two dif-

*To whom correspondence should be addressed. Tel: +34 917 328 059; Fax: + 34 912 246 976; Email: mtress@cnio.es

ficult tasks, detecting the remaining hard-to-find coding genes, and separating *bona fide* coding genes from misannotated pseudogenes and non-coding genes. Curators determine the status of the gene models based on transcript (ESTs and mRNAs) and protein data (from the main protein databases) available for each gene (12). Protein-coding potential depends first on whether an open reading frame (ORF) can be defined. However, the definition of ORFs is complicated by the fact that many noncoding transcripts may contain long ORFs by chance, particularly in GC-rich regions (11). In order to get round this problem, annotators also require some sort of protein evidence, such as whether the locus has sequence similarity to orthologues from other species, whether the resulting gene product contains Pfam functional domains (13), or whether experimental data is available from published papers, large-scale interaction studies (14) or mass spectrometry experiments (15).

Genes and transcripts may change their status between releases as annotators adjust the annotation to the available evidence. A gene's status is updated based on the available evidence and this evidence can change over time. For example GENCODE manual annotators recently decided to reclassify as non-coding approximately 200 'orphan' protein coding genes [GENCODE blog, https://gencodegenes.wordpress.com, April 2018]. Most of these genes were early *in silico* predictions.

A number of studies have put an estimation on the number of human coding genes, including several that have estimated the number to be close to or below 20 000 (11,16–18). Two of the more comprehensive studies into the coding complement of the human genome, Clamp *et al.* (11) and Church *et al.* (17), were carried out before GENCODE and other groups began the systematic manual reannotation of the genes in the human gene set. Both analyses assumed that most novel genes, defined as genes that arose from scratch in the primate lineage, are not protein coding. According to the Clamp analysis, the vast majority of novel ORFs did not have evolutionary conservation and had features that resembled non-coding RNA rather than coding genes. After discarding these orphan DNA sequences, as well as genes that appeared to be transposons, pseudogenes, and other miscellaneous artefacts, the authors ended up with a gene count of 20 500, roughly 4000 fewer than were annotated at that time. Church *et al.* carried out a comparison between the human and the mouse genomes and found that there were very few truly novel human genes, and that almost all protein-coding genes gained in the mammalian lineage were generated from whole gene duplications. They estimated that the number of protein coding genes was <20 000.

Many of the genes tagged as non-coding in these two analyses have since been removed from the reference set after manual annotation, though a number of genes identified in both studies as orphans or pseudogenes are nevertheless still annotated as protein coding, including the predicted pseudogenes *DHFRL1*, which has experimental evidence for a protein, and *HIGD2B*, which does not.

In 2014, we predicted that the human genome was likely to have just 19 000 protein coding genes based on the identification of 2001 'potential non-coding' genes (18). GENCODE manual annotators have since withdrawn or reclassified almost half of these genes from the human reference set. Most recently Southan [f1000research: doi: 10.12688/f1000research.11119.1] contrasted gene numbers in the three manually annotated reference sets with those of the HUGO Gene Nomenclature Committee [HGNC, (19)], noting the differences in coding gene counts and showing that UniProtKB proteins missing in RefSeq and Ensembl were enriched for elements classified by HGNC as endogenous retrovirus, long non-coding RNA or pseudogene.

Here, we expand our previous analysis to incorporate an analysis of the RefSeq and UniProtKB proteomes. We find that these two references databases and Ensembl/GENCODE annotate 22 210 genes as coding but only agree on 86% of the genes they annotate. In order to determine whether all 22 210 genes will code for proteins we contrasted the experimental evidence for genes annotated as coding in all three reference sets with those that are classified differently.

MATERIALS AND METHODS

Comparison of Ensembl/GENCODE, RefSeq and UniProtKB gene sets

We merged the coding genes in the three main versions of the reference human proteome, the Ensembl/GENCODE reference set (GENCODE v24, which is the equivalent of Ensembl 83), the RefSeq gene set (RefSeq 107) and the UniProtKB proteome (UniProtKB June 2016).

The UniProtKB reference proteome contained more than 70 612 SwissProt (reviewed) and TrEMBL (non-reviewed) entries. In order to compare UniProtKB with RefSeq and Ensembl/GENCODE, we merged these entries where possible by gene name. In UniProtKB genes can have more than one entry and UniProtKB entries may have more than one gene. After the initial merge the many orphan transcripts were merged first by their associated Ensembl identifier and then by hand where possible. This set of UniProtKB genes were then merged with the RefSeq and Ensembl genes using Ensembl's BioMart, UniProtKB's mapping tools and the HGNC gene names provided by the three reference sets. We carried out a painstaking manual reannotation of the more than 2700 genes where HGNC gene names, BioMart and UniProtKB correspondences did not agree.

Finally, for the 2764 genes not classified as coding in all the three reference databases we manually cross-referenced their status in the reference sets in which they were not annotated as coding.

Possible non-coding features

We have shown that a number of protein features, such as gene family age and cross-species conservation, are correlated with the detection of peptides in mass spectrometry experiments (18). These features can also be used to predict whether peptides will be detected in proteomics experiments and to flag protein-coding genes as potentially non-coding. The features are listed below.

UniProtKB uncertain, predicted, homology and missing evidence codes

Protein evidence codes are taken from the UniProtKB database. UniProtKB carries out manual annotation of proteins and human proteins in particular are well annotated and a large majority are annotated with the highest evidence score 'protein evidence'. The other four evidence codes in decreasing order are: 'Transcript evidence', 'Homology', 'Predicted' and 'Uncertain'.

Where there was more than one UniProtKB entry associated to an Ensembl/GENCODE gene we chose the UniProtKB entry with the highest ranked evidence to represent the gene. Genes annotated with 'Homology', 'Predicted' or 'Uncertain' evidence, and those genes for which we could not detect any evidence code at all, had very little evidence of protein expression; the four features between them covered 1599 genes and we found peptide evidence for 52.

UniProtKB cautions

UniProtKB appends cautions to many of their protein entries. Several of these cast doubt on whether they are expressed as proteins. We did not select all UniProtKB cautions, just those that suggested that the gene might be non-coding, non-functional or a pseudogene. The two most common cautions were: 'Product of a dubious gene prediction', 'Could be the product of a pseudogene'. There were 86 genes tagged with these cautions. We found peptide evidence for just three of these genes.

GENCODE

We took the translated GENCODE sequences as the coding gene set. The 20,266 genes in this set included not just protein coding genes, but also immunoglobulin receptors, nonsense mediated decay (NMD) transcripts and polymorphic pseudogenes. 13 148 of the coding genes are also annotated with non-coding transcripts, but these were not analysed.

Polymorphic pseudogenes

Polymorphic pseudogenes are loci that are pseudogenes in the reference genome that are intact in other individuals, and may represent coding genes that are undergoing a process of pseudogenization. There are 58 polymorphic pseudogenes in the reference gene set, of which 43 are olfactory receptors. It is particularly difficult to determine whether olfactory receptors are pseudogenes or code for functional proteins (20). We find peptide evidence for two of these polymorphic pseudogenes, *GBA3* and *PNLIPRP2*. Unlike most genes annotated with the polymorphic pseudogene tag, these two genes were annotated with both coding and polymorphic pseudogene transcripts.

Nonsense-mediated decay genes

A number of genes in the reference gene set only have NMD and non-coding transcripts. There were 204 genes annotated just with NMD and/or non-coding transcripts in the GENCODE v24 reference set. As might be expected, we did not find peptides for any of these genes.

Read-through transcripts

Read-through genes are genes in which all coding or NMD transcripts are tagged as read-through transcripts. There are also genes that have a mix of read-through and coding transcripts, though these are gradually being cleaned up. Read-through transcripts usually occur when a transcript skips the 3′ exon and reads through to exons from the neighbouring gene (which is usually coding but may be non-coding or pseudogene too). If translated, read-through transcripts would produce fusion proteins.

Read-through variants are annotated as part of the human coding gene set for technical reasons. While it is possible that the splicing together of two neighbouring genes is one way for proteins to gain new domains (21), it appears that very few of these read-through transcripts produce proteins at detectable levels. While we found peptide evidence for one of these genes (*IQCJ-SCHIP1*), there is enough evidence to suggest that it may actually be a single gene rather than two separate genes with read-through transcripts

Because read-through transcripts and proteins overlap with transcripts and proteins from known coding genes, these transcripts introduce a number of technical problems to genome-scale analysis. For example we had to map the spectra from the MS analyses to the GENCODE v24 database twice, once including the read-through proteins and once excluding them.

The numbers of read-through genes in the coding gene set is ever increasing. There were 470 read-through genes annotated either by GENCODE or in the Ensembl description.

Ensembl

Pseudogenes, non-functional genes, non-coding genes, antisense/opposite strand genes, miscellaneous RNA. We manually curated genes with tags from the Ensembl gene descriptions. Genes that were annotated as 'pseudogene', 'read-through', 'non-coding', 'non-functional', 'antisense', 'opposite strand' and 'long non-coding RNA' were tagged as potentially non-coding. There were 131 genes described as pseudogenes by Ensembl, 70 were olfactory receptors. We found peptide evidence for 4 of these genes. Another 93 genes were described as 'non-functional', 'antisense' or 'opposite strand'. We found peptide evidence for 6 of these genes. Finally 6 genes were described as 'non-coding' or 'long non-coding RNA'. We found peptides for three of these genes.

Primate gene family. These were genes from families that evolved in the primate lineage according to our analysis of data from Ensembl Compara (22). The primate lineage was here defined as all strata more recent than the boroeutheria class. Gene birth dating was carried out used the phylogenetic reconstructions of Ensembl Compara v84. We estimated a gene family age and an individual gene age for all coding genes annotated in GENCODE v24. The analysis was identical to that carried out in the previous paper (18), which itself was based on earlier study of gene ages (23) and is detailed below. Ensembl Compara v84 is constructed from genes from 70 different species; here we focused on phylostrata that represented the last common

ancestors of *Homo sapiens* and that had at least 5× coverage. Inconsistencies between gene trees and species phylogeny have been described for the *Euarchontoglires phylostratum* (24,25), so this was collapsed into the Eutherian level. Human coding genes were classified in the following age classes: Fungi/Metazoa, Bilateria, Chordata, Vertebrata, Euteleostomi, Sarcopterygii, Tetrapoda, Amniota, Mammalia, Theria, Eutheria, Boreoeutheria, Primates, Simiiformes, Catarrhini, Hominoidea, Hominidae, HomoPanGorilla and *H. sapiens*. In the analysis, all classes from Boreoeutheria to *H. sapiens* formed the 'Primate' class. The Sarcopterygii class was later clustered with Euteleostomi class because it contained few genes.

Compara classifies speciation and duplication nodes in family trees by the phylogenetic level in which the event took place (26) and our pipeline uses this information to define the *gene family age* and the *gene age* of each coding gene. Gene family age is the phylostratum at the root of the family tree (the earliest common ancestor that has a member of the gene family) while gene age is the phylostratum in which the genomic event leading to an extant gene takes place. For singleton genes the family gene age is always the same as the gene age, for duplicated genes the gene age represents the species in which the last duplication took place. Only duplication events with a consistency score (27) >0.3 were considered in the gene age analysis. Nodes with zero scores were trimmed out of the analysis. Duplication nodes with consistencies between 0 and 0.3 were labelled as 'unclear' and gene age was not assigned.

To our surprise we found more primate family genes in this study (700) than in our previous study (563). We found protein evidence for just 27 primate family genes.

Curiously there are sixteen coding genes that Compara tags as novel (non-duplicated) human genes in GENCODE v24. All are single exon genes predicted by Ensembl automatic prediction programs (e.g. see Supplementary Figure S1). None of these novel human genes have their coding status supported in any other reference set or any by peptide or antibody evidence.

PhyloCSF Score. We used exon-based PhyloCSF scores (27) to represent a measure of conservation for each gene. PhyloCSF was run using the 58 mammals parameters and the 'mle' and 'bls' option on the coding portion of each exon, trimmed to codon boundaries and excluding the final stop codon. Alignments were extracted from the 100-vertebrate MULTIZ hg38 alignment, with species restricted to the 58 placental mammals.

The conservation score was the PhyloCSF score of the highest scoring exon, counting only exons at least 42 bases in length and for which the relative branch length of the local alignment reported by PhyloCSF's 'bls' option was at least 0.1, since PhyloCSF scores are unreliable if there is insufficient branch length. Genes having no exons satisfying these conditions were flagged as having exons that were too short or with too few relatives to return a PhyloCSF score.

Genes with a maximum PhyloCSF exon score of less than −16 or genes that had a relative branch length of less than 0.1 were flagged as having a poor PhyloCSF score. We found peptide evidence in PeptideAtlas for 28 of the 453 genes with poor branch length and 2 of the 132 genes with a maximum PhyloCSF exon score of less than −16.

APPRIS. All Ensembl genes are annotated with protein data in the APPRIS database (28). APPRIS annotates the following protein-based features: homology to proteins with known structure is mapped onto variants using HH-search (29); functionally important residues and protein functional domain mapping comes from *firestar* (30) and pfamscan (31); trans-membrane helices are mapped using three separate trans-membrane predictors (32–34) and signal peptides are predicted by SignalP (35). A module of APPRIS calculates a measure of conservation by mapping vertebrate orthologues present in the protein databases. While APPRIS calculates features for all annotated coding variants, we took the mapping from the principal isoforms for each gene.

Protein features were calculated for all genes. Genes that did not have functional information, structural information or conservation in formation were tagged as potential non-coding when they had a PhyloCSF score below 2. There were just 17 genes with no protein information but with peptide evidence in our analysis.

Transcript expression from Human Protein Atlas. We downloaded data from the RNAseq experiments carried out for the Human Protein Atlas (36). The Human Protein Atlas RNAseq experiments were carried out on 36 tissues using Ensembl v83 (equivalent to GENCODE v24). For each gene, we counted the number of tissues in which the expression level was measured to be at least 1 transcript per million (TPM). Genes were binned by the number of tissues in which they were detected with at least 1 TPM.

Peptide data from PeptideAtlas. We downloaded all peptides identified in the January 2016 build of the human PeptideAtlas (15), in total 1 166 164 peptides. 880 101 peptides (75.5%) were semi-tryptic with respect to the GENCODE v24 human reference set, even though trypsin is used to cleave the proteins in the vast majority of proteomics experiments. We have previously found that semi-tryptic peptides are considerably less reliable than tryptic peptides (18), though most of these peptides were by-products of wholly tryptic peptides.

Including semi-tryptic peptides would have identified 711 more genes, 13.5% of which would have been potential non-coding genes. Less than 1% of the genes identified with tryptic peptides were potential non-coding genes. There is no reason why semi-tryptic peptides should identify 10 times as many potential non-coding genes than tryptic peptides, so semi-tryptic peptides were excluded on the grounds of accuracy.

We also eliminated peptides shorter than nine residues and peptides that mapped to more than one gene. Finally, we eliminated nested peptides; where two peptides had the same sequence but one was shorter than the other, we eliminated the shorter peptide. We mapped the remaining 153 913 peptides to the genes in GENCODE v24. At least two peptides had to map to each gene in order to identify it.

Obtaining and filtering of CNV maps. Whole genome copy number variation (CNV) maps were downloaded from

five different publications (37–41). In order to homogenize the different maps, we selected autosomal and not private CNVs. Additionally, we removed CNVs marked as low quality from Handsaker *et al.* (40) and all the variants from two of the individuals (NA07346 and NA11918) because we were not confident about their genotype. From the maps in Zarrei *et al.*, (39) we selected the stringent map that considered CNVs that appeared in at least two individuals and in two studies. Homozygous whole gene losses were calculated for all maps except for Abyzov *et al.*, (41) which did not specify the copy number of the deletions.

Genetic variation. We compared rates of genetic variation for genes with potential non-coding features against the genetic variation rates for likely coding genes using data from 2504 individuals in phase 3 of the 1000 Genomes Project (42). We remapped these variants from GRCh37 to GRCh38 using dbSNP v149 (43). Most of the variants could be mapped from GRCh37 to 38 by using dbSNP identifiers (rsIDs). The exceptions were 186 854 variants with no rsID in dbSNP v149, and 256,769 variants for which the reference base has changed between GRCh37 and GRCh38. The rest of the variants (99.47%; 84 358 257/84 801 880) were successfully mapped. When available, ancestral allele information from the 1000 Genomes Project was used to translate allele frequencies into derived allele frequencies.

We ran VEP (variant_effect_predictor.pl, (44)) v84 using either the Ensembl v84 cache (for Ensembl/GENCODE) or a cache built locally using gene annotations from RefSeq v107 to predict the effects of variants. We calculated the percentage of high-impact variants and the ratio of non-synonymous to synonymous variants for rare and common allele frequencies. High impact variants were splice acceptor, splice donor, stop gain and stop loss variants. Common alleles were those with an allele frequency higher than 0.005 (equivalent to >25 allele counts in autosomes), while rare alleles were those with an allele frequency <0.005.

Only variant effects corresponding to the APPRIS principal isoform (28) of each coding gene were considered. Variants were considered only for strictly defined protein coding genes, not for the immunoglobulin and t-cell receptor fragment genes to exclude the possibility of positive selection.

RESULTS

Coding genes in the three main reference sets

We compared the coding genes in the three main versions of the human proteome, the merged Ensembl/GENCODE reference set, the RefSeq gene set and the UniProtKB proteome. The comparison was based on GENCODE v24 (Ensembl 83), UniProtKB June 2016, and RefSeq 107. RefSeq 107 annotates 20 450 coding genes, and the Ensembl/GENCODE merge contains 20,266 coding genes. The UniProtKB proteome is based around proteins rather than genes. UniProtKB June 2016 proteins mapped to 21 212 coding genes.

In total the three reference sets annotate 22 210 protein-coding genes. There are a maximum of 19 446 genes annotated as coding in the intersection of the three sets (Figure 1). This is a maximum because boundaries are disputed for a small number of genes. There are eight

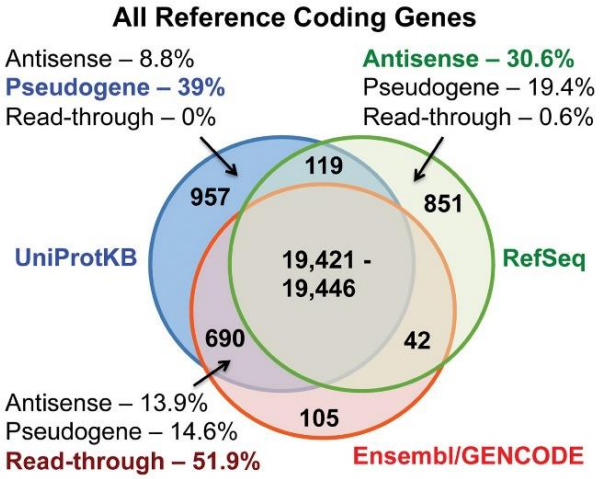


Figure 1. The overlap between Ensembl/GENCODE, RefSeq and UniProtKB genes. The number of genes classified as coding in each of the three reference databases and the intersection between them. The number of genes in the intersection of A is variable because RefSeq and Ensembl/GENCODE disagree on gene boundaries for a number of genes. For three subsets of genes, we show the percentage of coding genes annotated as antisense, pseudogene or read-through in another database.

cases where Ensembl/GENCODE has two genes but RefSeq annotates one gene and sixteen cases where single Ensembl/GENCODE genes are annotated as multiple genes in RefSeq (*PTPRQ* is three RefSeq coding genes and only one in Ensembl/GENCODE). If all 24 genes were single genes rather than being split, there would be 19 421 coding genes common to the three reference sets. Beyond the intersection of the three reference databases, 851 genes are supported by two of the three reference sets and 1903 genes are annotated in just one of the three reference sets.

Ensembl/GENCODE has the fewest unique coding genes (105). This is for technical reasons. Most genes annotated by Ensembl/GENCODE are automatically included in UniProtKB. Given the near automatic transmission of coding genes between Ensembl/GENCODE and the UniProtKB proteome, the 690 genes annotated as coding by Ensembl/GENCODE and UniProtKB might also be regarded as singleton coding genes.

Almost a quarter of coding genes not present in all three reference sets are annotated as pseudogenes by manual annotators from other databases (Supplementary Table S1) and this rises to 39% of coding genes annotated in UniProtKB only (Figure 1). Potential 'antisense' genes, non-coding genes on the opposite strand to protein-coding loci, form the second largest group of differently annotated genes; 17% of coding genes not annotated in all three sets and 31% of genes classified as coding in RefSeq only are antisense. More than 50% of genes that are coding in Ensembl/GENCODE and UniProtKB but not in RefSeq are read-through genes. Read-through genes (genes made up entirely of transcripts that skip the last exon of one coding gene to read through to exons from the neighbouring gene or pseudogene) are currently annotated as coding by both the RefSeq and

Ensembl/GENCODE annotations even though there is little indication that they code for proteins.

Each reference set has its own biases and idiosyncrasies. UniProtKB annotates 26 retroviral genes and a large number of T-cell receptor and immunoglobulin genes as part of the human reference proteome and include 84 genes that are part of alternative loci in the haploid assembly. RefSeq annotates 44 genes as sense overlapping (i.e. the locus of the gene overlaps with a known protein coding gene in the same sense), while Ensembl/GENCODE has 41 genes that are exact duplicates of annotated coding genes due to technical problems with the merge between GENCODE and Ensembl (Supplementary Table S1).

Are there 22 210 coding genes in the human genome?

There is a remarkable discrepancy between the number of genes classified as coding by all three reference sets and the number of genes classified as coding by at least one of the individual reference sets; 14.4% more genes are classified as coding in the union of the three reference sets than in the intersection. How many of these 2764 extra genes annotated by just one or two of the reference databases are protein coding?

UniProtKB annotation

Genes classified as coding solely by UniProtKB are unique in that they do not come with reference coordinates. Indeed many UniProtKB proteins are annotated as unplaced because the annotators do not know where in the genome the gene is found. However, the UniProtKB database provides an evidence scale for their manual annotations, ranging from the most reliable ('supported by protein evidence') to the most dubious ('uncertain'). We used these classifications to compare genes classified as coding by UniProtKB. For each gene, we took the protein with the most reliable evidence as the representative.

The evidence codes of genes classified as coding in the coding gene subsets (UniProtKB and RefSeq, UniProtKB and Ensembl/GENCODE and solely UniProtKB) are clearly distinct from those classified as coding in all three reference databases (Supplementary Figure S2). More than 80% of the genes classified as coding across all three reference databases are annotated with the highest UniProtKB evidence score, 'supported by protein evidence'. Outside of this intersection the proportion of genes supported by protein evidence is much smaller; those genes annotated by UniProtKB only have the next highest level of confirmation with just 19% of proteins supported by protein evidence, and three quarters of these are immunoglobulin genes, T-cell receptors, viral proteins and proteins from alternative loci that are not in the reference genome-based databases. By contrast >50% of the coding genes unique to UniProtKB are supported by the 'uncertain' evidence code, while over half the genes classified as coding by UniProtKB and RefSeq are supported by transcript evidence alone, and more than two thirds of genes that are classified as coding by UniProtKB and Ensembl/GENCODE are annotated as being supported by 'predicted' evidence. Genes annotated as coding in just one or two reference databases clearly have much weaker evidence in UniProtKB.

Table 1. The 16 potential non-coding features used to select the 2278 potential non-coding genes

Features	Genes G24	No. Detected in MS
No protein features [A]	586	17
Primate gene [C]	700	27
Pseudogene [E]	131	4
Non-functional [E]	74	6
Antisense/Opposite Strand [E]	19	3
Non-coding [E]	6	3
Read-through gene [G]	467	1
Nonsense mediated decay [G]	204	0
Polymorphic pseudogene [G]	56	2
PhyloCSF branch length [M]	453	28
PhyloCSF maximum [M]	132	2
Predicted evidence [U]	853	12
Homology evidence [U]	613	39
No evidence code [U]	101	0
Caution note [U]	86	3
Uncertain evidence [U]	32	1

The abbreviations show the source of each annotation: A – APPRIS, C-Ensembl Compara, E – Ensembl annotations, G – GENCODE annotations, M - MIT, U – UniProtKB annotations.

Potential non-coding features

In a previous work we flagged 2001 coding genes from the GENCODE v12 gene set as potentially non-coding (18) based on a set of features that were more typical of non-coding genes than coding genes (potential non-coding features). These features were all associated with extremely poor detection rates in mass spectrometry analyses. Manual annotators have since reclassified 908 of these genes as pseudogenes or non-coding RNA. Since genes annotated as coding in just one or two reference sets have less evidence in UniProtKB, it seems logical that many of these genes will also be enriched potential non-coding features.

Using the Ensembl/GENCODE coding genes we defined a set of potential non-coding features. The features included the weakest three UniProtKB evidence codes and manually added caution notes from the UniProtKB manual annotators, read-through, nonsense mediated decay and polymorphic pseudogenes tags from the GENCODE manual annotation, labels indicating pseudogene or non-coding gene from the Ensembl database and four measures of conservation, poor PhyloCSF (21) maximum score and relative branch length (which indicates that evolutionary coding potential within placental mammals is low), absence of conserved protein structure, function or conservation according to the APPRIS (28) database and those genes that have evolved within the primate clade according to Ensembl Compara (22).

The 16 potential non-coding features, the numbers of genes that were tagged with each feature and the number of these genes that had peptide evidence from large-scale proteomics analyses are listed in Table 1 and the features themselves are detailed in the Materials and Methods section.

A total of 2278 Ensembl/GENCODE coding genes were tagged with at least one of the 16 potential non-coding features. These genes were labelled as 'potential non-coding genes'. The remaining 17 988 coding genes are referred to

as the 'likely coding gene' set in this analysis. The correspondence between the potential coding genes tagged in Ensembl/GENCODE and in GENCODE v12 is shown in Supplementary Figure S3.

Potential non-coding genes are not distributed evenly between the intersection of three reference sets and the Ensembl/GENCODE gene subsets (Ensembl/GENCODE and UniProtKB, Ensembl/GENCODE and RefSeq, and Ensembl/GENCODE alone). While there were 1471 potential non-coding genes in the intersection of the three sets, this was just 7.6% of the genes. By contrast potential non-coding genes made up 96.5% of the genes (808 of 837) in the Ensembl/GENCODE gene subsets (Supplementary Figure S4).

The fact that almost all the genes outside the intersection of the three reference sets have potential non-coding features suggests that many of them may not code for proteins under normal cellular conditions. As a first step to testing this hypothesis we analyzed the experimental expression of potential non-coding genes using available experimental transcriptomics, proteomic and antibody binding data and compared this to likely coding genes.

Transcript evidence

We downloaded RNA expression data from the Human Protein Atlas. The Human Protein Atlas details RNAseq experiments carried out on 36 tissues using Ensembl83 (equivalent to GENCODE v24). For each gene we looked at the maximum expression in any one tissue and counted the number of tissues in which expression was at least 1 transcript per million (TPM). We binned genes by maximum expression and by number of tissues and compared the tissue distributions of likely coding genes and potential non-coding genes in both the intersection and subsets of coding genes annotated in Ensembl/GENCODE, but not in both other reference sets.

There was considerably more evidence for the expression of likely coding genes: 73.5% of likely coding genes had a maximum TPM of 20 or more against just 24.3% of potential non-coding genes (Figure 2). In fact 52.9% of potential non-coding genes had a maximum TPM of fewer than 5. The median expression level for potential non-coding genes was just 4.1 TPM, compared to 43.4 TPM in the likely coding set. Potential non-coding genes from the Ensembl/GENCODE coding subsets have a very similar distribution to potential non-coding genes from the intersection of the three sets. There were too few likely coding genes in the Ensembl/GENCODE coding subsets (29) to show in the graphic.

Likely coding genes also have entirely different tissue-specific characteristics from potential non-coding genes. While likely coding genes tend to be expressed in detectable quantities over most tissues (62.7% of these genes are detected in at least 30 tissues), the majority of potential non-coding genes are found in few tissues (Supplementary Figure S5). More than two thirds of potential non-coding genes (66.5%) have detectable expression in five or fewer tissues.

The skewed tissue-distribution of both sets of possible non-coding genes (Supplementary Figure S5) might suggest that these genes are more tissue-specific, and it is true that a higher proportion of potential non-coding genes are olfactory receptors and would be expected to be expressed in limited tissues. However, potential non-coding genes still have much lower expression levels even when olfactory receptors are removed (Supplementary Figure S6). Most potential non-coding genes had a maximum expression of fewer than 5 TPM, so differences in tissue expression might also be a reflection of generally low expression levels in which the 1 TPM threshold is crossed only in few tissues.

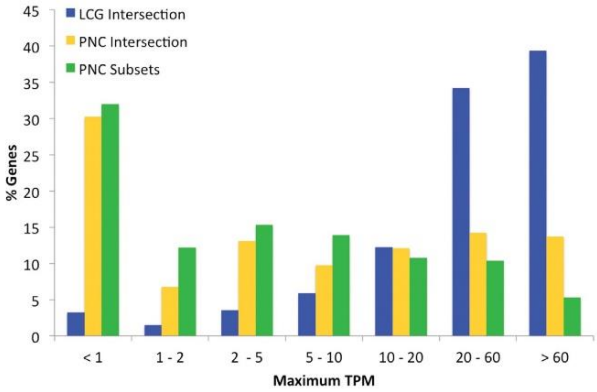


Figure 2. Maximum transcript expression of potential non-coding genes and likely coding genes. The percentage of genes in seven different maximum TPM bins. Maximum TPM comes from the 36 tissues of the Human Protein Atlas RNAseq experiments. Tissue distribution shown for the likely coding genes (LCG Intersection) as well as potential non-coding genes annotated by all three reference sets (PNC Intersection) and by just one or two sets of annotators (PNC Subsets).

Protein expression

We carried out two analyses to identify gene products, an analysis of the collected peptides from the PeptideAtlas proteomics database and an investigation of the antibody information housed in the Human Protein Atlas.

We culled peptides from the PeptideAtlas database (January 2016), which contains 238 402 discriminating tryptic peptides. We required protein detection to be supported by two or more distinct uniquely-mapping, non-nested peptide sequences of at least 9 amino acids as suggested by Human Proteome Project consortium (45).

We detected at least two non-nested peptides for 13 360 of the 17 988 likely coding genes (74.3%). By way of contrast genes with potential non-coding features had extremely low levels of peptide detection [Table 1]. In total we detected peptides for just 142 of the 2278 potential non-coding genes (6.2%). Less than 1% of the genes identified by PeptideAtlas were potential non-coding genes.

Human Protein Atlas antibodies

The Human Protein Atlas has been developed to validate tissue-specific protein expression. We downloaded antibody-specific protein expression information from normal tissues from the Human Protein Atlas (Version 16, January 2016). We excluded expression data for antibodies that

identified more than one gene and identifications tagged as 'uncertain'.

The remaining antibodies detected a higher proportion of protein expression for the genes in the likely coding set (9896 of 17 988 genes, 55%) than for the genes in the potential non-coding set (just 79 of the 2278 genes, 3.5%). Potential non-coding genes that were validated by Human Protein Atlas antibodies included primate genes *STATH* (statherin), *HTN3* (histatin-3) and *SCT* (secretin), all of which code for secreted proteins.

Genes detected by PeptideAtlas peptides and Human Protein Atlas antibodies are shown in Supplementary Figure S7. 8794 genes were detected in both analyses, only 46 of which were potential non-coding genes (0.52%). At the same time 2101 of the 5681 genes not detected in either analysis (37%) were potential non-coding genes.

There is quite clearly less evidence for the expression of potential non-coding genes both at the transcript and protein level. Chi-squared tests show that expression patterns of potential non-coding genes are significantly different from those of likely coding genes in all three sets of experimental observations.

Potential non-coding genes even have even less protein evidence than one would expect from the RNAseq levels. Peptides can be found in PeptideAtlas for 92% of likely coding genes that have RNAseq expression of at least 1TPM in all 36 Human Protein Atlas tissues (Figure 3), but the peptide support falls to just 25% for potential non-coding genes. A similar pattern can be seen when genes are binned by maximum TPM across all 36 Human Protein Atlas tissues. Proportionally we found 5–10 times more likely coding genes than potential non-coding genes (Figure 3) in each bin. Even in the most widely expressed genes, which we defined those genes that are expressed in at least 10 tissues with a minimum of 10 TPM, there is still much more peptide evidence for likely coding genes than potential non-coding genes. We detected peptides for 85.6% of likely coding genes, 19.4% of potential non-coding genes annotated by all three reference sets and just 6.1% of potential non-coding genes annotated in two or fewer sets (Figure 3).

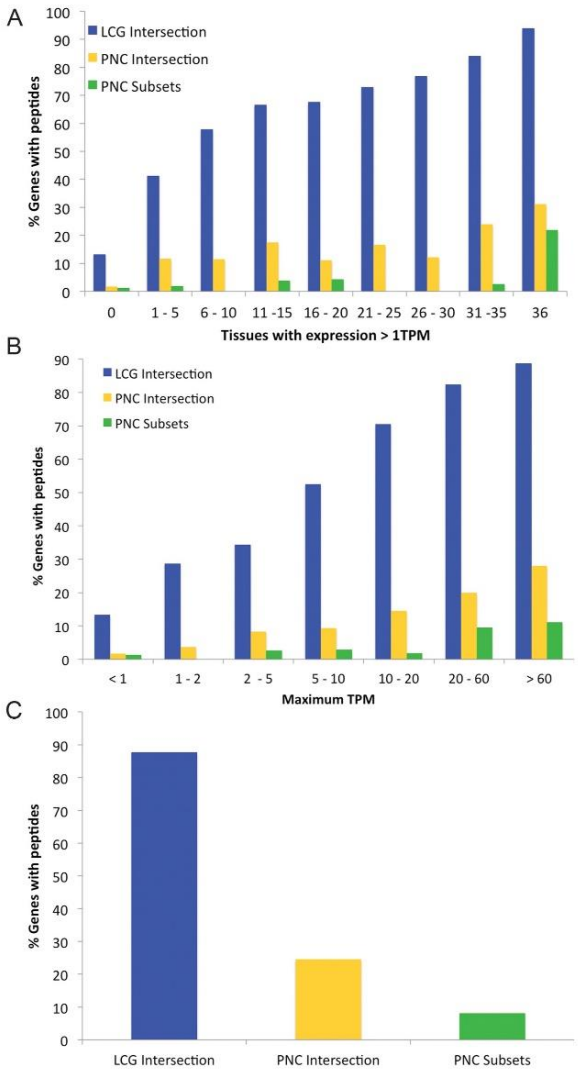


Figure 3. The relation between peptides in proteomics experiments and transcript expression. (**A**) The percentage of genes for which peptides are detected in PeptideAtlas across nine different bins. The bins are based on the number of tissues in which the transcripts are detected with a TPM of >1 in the Human Protein Atlas RNAseq experiments. (**B**) The percentage of genes for which peptides are detected in PeptideAtlas divided across seven bins of maximum TPM for each gene taken from the 36 tissues of the Human Protein Atlas RNAseq experiments. (**C**) The percentage of genes for which peptides are detected for those genes that have RNAseq expression in at least 10 tissues with a TPM of 10 or more. In each case the percentage of genes for the likely coding genes (LCG Intersection, blue bars) as well as potential non-coding genes annotated by all three reference sets (PNC Intersection, yellow) and by just one or two sets of annotators (PNC Subsets, green).

Genetic variation

Human genetic variation can be used to shed light on whether or not potential non-coding genes code for proteins. The rate of copy number variation and the proportion of damaging high impact variants can provide clues to the functional relevance of coding (or non-coding) genes. Because of the effects of purifying selection coding genes should have substantially lower non-synonymous to synonymous variant ratios than non-coding genes that are misannotated as coding.

Copy number variation

We downloaded genome copy number variations (CNV) maps from five different publications (37–41). The CNVs were mapped to the GRCh37 build of the human genome, so we compared rates of gene gain and loss in the subset of Ensembl/GENCODE genes that were also annotated GENCODE v12. We also looked at CNVs in those GENCODE v12 genes that have since been removed from the coding reference set and reclassified as non-coding, pseudogene or artefact.

The rate of gene loss and homozygous gene loss through CNVs for each set are shown in Figure 4. Potential non-coding genes from Ensembl/GENCODE have more than five times as much gene gain as likely coding genes and

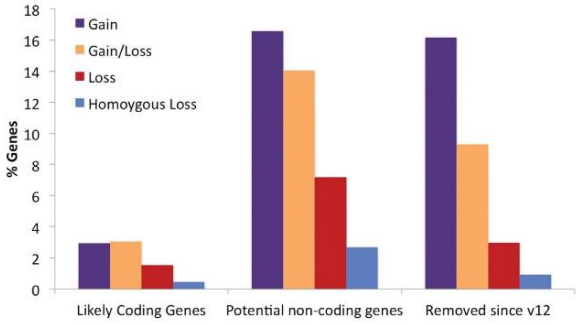


Figure 4. Whole gene gains and losses for likely coding and potential non-coding gene in GENCODE v12. The percentage of genes that have undergone gene gain/loss (purple), whole gene gain (orange), whole gene loss (red) or homozygous gene loss (blue) in at least one of the five different analyses. Potential non-coding genes present in both GENCODE v12 and v24 undergo a similar proportion of gene gain and loss to GENCODE v12 genes that have since been reclassified as not coding.

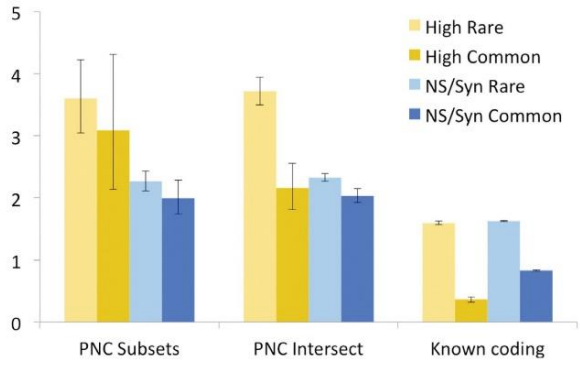


Figure 5. Genomic variation in likely coding genes and possible non-coding genes. Percentage high impact variants (yellow) and non-synonymous/synonymous ratios (blue) for known coding genes (likely coding genes with peptide evidence, see text) and for possible non-coding genes (PNC) from the intersection of the three sets (Intersect) or annotated by two or fewer reference sets (Subsets). Read-through genes were removed when calculating variants because they always overlap known coding genes. The darker colours show the values for common variants and the lighter shades show the values for rare variants. 95% confidence intervals are shown.

almost five times as much genes loss. The distribution of CNVs in potential non-coding genes is similar to that of GENCODE v12 genes that are no longer classified as coding, though potential non-coding genes have even more evidence of gene loss.

Genetic variation within the human population

The patterns from the CNV study suggest that potential non-coding genes are under weaker selection than likely coding genes. To further characterize the strength of selection we analysed the patterns of genetic variation in the human population using data from 2504 individuals in phase 3 of the 1000 Genomes Project (42). For the calculation we separated variants by allele frequency: common alleles were those with an allele count of more than 25, equivalent to an allele frequency of 0.005, while rare alleles were those with an allele count of fewer than 25. Variant effects were determined using the main protein isoform to represent each GENCODE v24 coding gene (28).

The percentage of high-impact variants and the ratio of non-synonymous to synonymous variants for rare and common allele frequencies were calculated using the results from VEP (44). For the large-scale comparison high impact variants included splice acceptor, splice donor, stop gain, stop loss, but not indel variants. This is because indels are generally validated only with higher allele counts and are therefore almost always overrepresented in common alleles.

If purifying selection is preventing high impact or missense substitutions, these variants should be depleted from higher allele frequencies. Hence, differences in the patterns of high-impact and missense substitutions between rare and common alleles can be used to determine whether there is purifying selection or neutral evolution in large sets of protein coding genes. We have used this method previously to show that the majority of alternative exons are not undergoing purifying selection (46,47).

The percentage of high impact variants at rare allele frequencies is 1.88% for likely coding genes and this drops to 0.61% for common alleles. Within the likely coding gene set there are genes undergoing positive selection and there may even be genes that are not functionally important within this set. When we filter out immune system genes from the likely coding gene set and calculate high impact variants just for those genes that we detect peptides for, the difference is even starker: 1.6% for rare allele frequencies and just 0.36% for common allele frequencies (Figure 5). Likely coding genes with peptide support also have a much lower non-synonymous to synonymous ratio in common alleles, as would be expected for protein-coding genes evolving under negative selection.

By way of contrast potential non-coding genes annotated in all three sets have proportionally more high impact variants (3.72% at rare allele frequencies and 2.16% at common allele frequencies) and non-synonymous to synonymous ratios (2.33 for rare allele frequencies and 2.03 for common allele frequencies), and the results for potential non-coding genes annotated as coding in just one or two sets are similar (Figure 5). The fact that potential non-coding genes have a much higher proportion of high impact variants and greater non-synonymous to synonymous ratios than likely coding genes, suggests that many potential non-coding genes are unlikely to code for functionally important proteins.

With the genes annotated in Ensembl/GENCODE and RefSeq it is possible to generate human population data for all subsets of genes in Figure 1 with the exception of those genes annotated as coding only by UniProtKB. The percentage of high-impact variants and the ratio of non-synonymous to synonymous variants for these subsets are shown in Supplementary Figure S8. The contrast between genes classified as coding in all three reference databases and those in two or fewer sets is clear. Genes classified as coding in just one or two sets have much higher rates of high impact variants than genes classified as coding across all three databases. There are also no significant differences in non-synonymous to synonymous ratios between rare and

common allele frequencies in any set of genes that are classified differently across the three reference sets.

The genetic variation for individual potential non-coding features sheds some light on which of the potential non-coding genes are more likely to code for functional proteins. With the exception of read-through genes (most read-through genes are two known coding genes joined together) all features have genetic variant distributions that are very different from likely coding genes (Supplementary Figure S9). Primate genes, genes with 'predicted' UniProtKB evidence and genes with poor PhyloCSF scores have much higher non-synonymous to synonymous ratios and percentages of high impact variants than likely coding genes. However, the non-synonymous to synonymous ratios are lower for common allele frequencies and the differences between rare and common allele frequencies are significant. This suggests that a certain number of genes in these three categories may be functionally important protein-coding genes.

By contrast there are no significant differences in non-synonymous to synonymous ratios between rare and common allele frequencies for genes tagged with the potential non-coding features 'pseudogene', 'uncertain' UniProtKB evidence and 'UniProtKB caution', which suggests that a large majority of these genes are undergoing neutral evolution and are not functionally important.

Another subset of genes with high rates of damaging mutations and little differences between rare and common allele frequency non-synonymous to synonymous ratios are those genes populated entirely by automatically predicted transcript models. There were more than 800 genes predicted automatically in RefSeq and more than 200 in Ensembl/GENCODE. In sets of automatically predicted genes non-synonymous to synonymous ratios are practically identical for rare and common allele frequencies (Supplementary Figure S10), suggesting that most of these genes are also subject to neutral evolution.

Genes with high rates of missense variants

Genetic variation data is useful for pinpointing probable neutral evolution in large cohorts of genes, but the sparseness of the variants means that it is difficult to make conclusions about most individual genes. A number of coding genes do have remarkably high rates of missense and damaging variants though. We looked at the 15 genes with the highest proportion of non-synonymous variants (minimum 30 common allele variants). Nine were HLA histocompatibility antigens (Figure 6), which is not surprising since these genes are known to have many missense variants. Two of the other six genes might also be expected to have higher levels of missense variants because of their likely function. *MICA* (MHC class I polypeptide-related sequence A) a self-recognising antigen from the major histocompatibility complex class I locus and has more than 50 known alleles, several of which are truncating. Similarly, *BTNL2* (Butyrophilin-like protein 2) is a known polymorphic locus bordering the major histocompatibility complex class II and class III regions.

The remaining four genes (Figure 6) are *CRIPAK, PRAMEF2, PRR21* and *OR2T8*.

PRAMEF2 and *OR2T8* are likely to be pseudogenes; olfactory receptors are highly duplicated and many of these duplications may be pseudogenes, while *PRAMEF2* has 22 almost identical paralogues, none of which is supported by protein evidence. *PRR21 (*Putative proline-rich protein 21) is a single exon gene, which was annotated as 'uncertain' by UniProtKB but has since been removed from the UniProtKB proteome. It has an orthologue in chimpanzee, but little other supporting evidence and no evidence of transcript expression. *CRIPAK* (Cysteine-rich PAK1 inhibitor) was described in a 2006 paper (48) in which *CRIPAK* constructs appeared to bind and block *PAK1* activity. It was described as having 13 zinc finger domains, but the zinc finger domains are not real domains, merely degenerate cysteine-rich repeats (Supplementary Figure S11). Meanwhile *CRIPAK* is primate-specific and has practically no cross-species conservation at all, as can be seen from the partial alignment of the few orthologous sequences that can be found in UniProtKB (Figure 6). Although transcript expression is ubiquitous, there is no evidence for its expression as protein. Curiously it has the same expression pattern as the upstream coding gene, *UVSSA* (Supplementary Figure S11). *CRIPAK* is highly unlikely to be a coding gene and has been reclassified by GENCODE annotators.

Annotation of coding genes based on conflicting evidence

Manual annotators determine the status of genes based on the balance of the available evidence. For most genes, the available evidence is in agreement and the designation of coding or non-coding status is fairly straightforward. However for those genes that might be considered edge cases at the boundary between coding and non-coding, the evidence can often be contradictory.

There are a number of genes in the potential non-coding gene set that are supported by published studies, but that have little other evidence to support their translation to protein in normal tissues. One example is *CRIPAK* (see above), annotated as coding based on a single published study. At the other end of the spectrum is *ARMS2*, a gene that evolved in the primate clade from an L2 transposon. Since *ARMS2* has been linked to macular degeneration, it has >200 publications, many of which are association studies. The exact role of *ARMS2* in macular degeneration is not clear. Experiments carried out with a plasmid-induced protein show that if *ARMS2* were expressed in retinal cells, it would be secreted via an unconventional route (49).

Ensembl/GENCODE and UniProtKB annotate *DLEU1* (deleted in leukaemia 1) as encoding a short protein as well as 33 non-coding transcripts (Supplementary Figure S12). RefSeq annotate the *DLEU1* as non-coding. *DLEU1* was added to UniProtKB in 1997, and in 2007 it was annotated as having 'protein evidence' because it appeared to interact with other proteins in large-scale protein-protein interaction experiments (50). There is little other evidence that *DLEU1* codes for a protein. There is no proper proteomics evidence, very poor cross species conservation, practically no conservation of the reading frame (12) and coding exons of *DLEU1* overlap with a SINE (MIRb) element. While UniProtKB annotators use evidence from large-scale protein-protein interaction experiments to label proteins

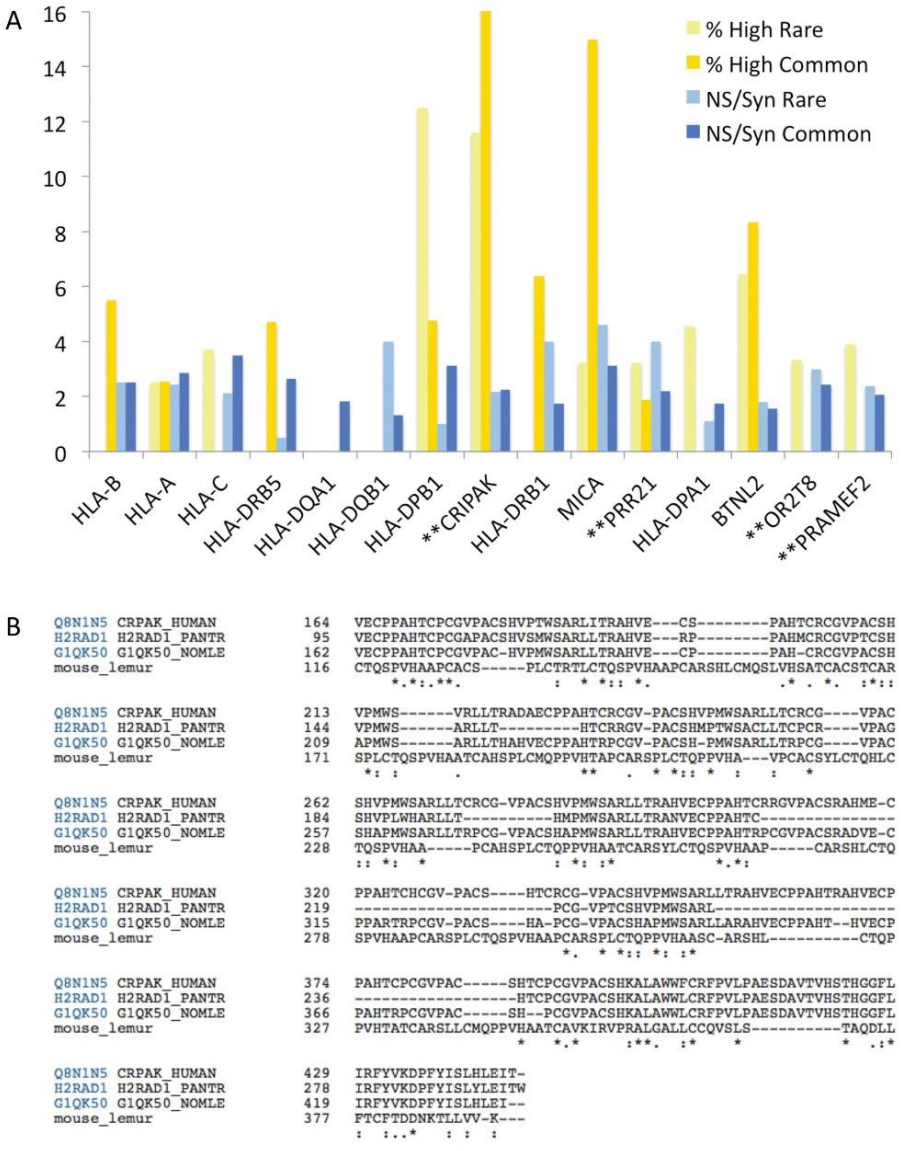


Figure 6. Genes with the highest proportion of high impact and non-synonymous variants. In (**A**), the percentage of high impact variants (yellow) and non-synonymous/synonymous ratios (blue) for the 15 genes with the highest rate of common non-synonymous variants. Minimum 30 common variants per gene. The darker colors show the values for common variants and the lighter shades show the values for rare variants. In (**B**), the alignment between human *CRIPAK* gene product and primate homologues annotated as *CRIPAK* in UniProtKB. There is very little evidence of conservation.

with the evidence code, 'protein evidence', evidence from large-scale protein-protein interaction experiments is not always sufficient to confirm protein-coding status. Large-scale interaction experiments construct proteins artificially and use these artificially generated proteins to see if they stick to other proteins. Proteins are generally sticky, even artificial ones, so binding between artificial constructs and real proteins is possible. While a great many of the detected *in vitro* interactions may also take place *in vivo*, a number will not. *DLEU1* is almost certainly a non-coding gene rather than a coding gene and will be reclassified as non-coding by Ensembl/GENCODE manual annotators.

There are many potential non-coding genes annotated with 'protein evidence' because of protein-protein interaction studies. These include *DRICH1* (which has a dN/dS above 1 between human and primates and a higher non-synonymous/synonymous ratio for common allele frequencies), *FAM218A* (which has very little evidence of homologues, even in primates), *PRR20C* (which has a dN/dS above 1 and homologues in primates only) and *RP11-*

511P7.5 (which appears to be a pseudogene at the 3′ end of *ZNF755*).

Some genes within the likely coding gene set also have conflicting evidence for their coding capability. Polo-like kinase 5 (*PLK5*) is detailed in Supplementary Figure S13. Glycine receptor subunit alpha-4 (*GLRA4*) is interesting because it is one of a number of coding genes that have human-specific stop codons. Glycine receptors are ligand-gated chloride channels and are highly conserved (chicken and mouse *GLRA4* are 94% identical over all but the first 40 residues). A mutation in the human version of *GLRA4* generates a protein that is truncated 39 amino acids from the C-terminus of the protein, removing the C-terminal transmembrane helix (Figure 7). This would almost certainly destabilize any pore, and would probably have considerable effect on the function.

The genetic variation for the four human glycine receptor genes is shown in Figure 7 along with data for GLRA4 from the Exac experiment (51). The family members with intact structure (GLRA1, GLRA2 and GLRA3) have no high impact mutations and the non-synonymous to synonymous ratio is higher for rare alleles than for common allelesIn contrast, 3% of common alleles variants in *GLRA4* are high impact and the non-synonymous to synonymous ratios are high for both rare and common alleles. The variation data suggest that *GLRA4* is not under selective pressure and is likely to be a unitary pseudogene.

The propagation of erroneous annotations

GVQW1 and *GVQW2* are short primate-specific genes with poor conservation (Supplementary Figure S14) that are classified as coding in all three reference databases, but that are tagged as potential non-coding in our study. *GVQW1* originated from an Alu SINE element, while *GVQW2* was annotated as coding recently. Pfam domains (13) are often used to help distinguish coding genes from non-coding genes and both genes seem to have been annotated as coding based on the presence of the domain GVQW.

Pfam annotators have recently removed the transposon-derived GVQW domain from the database as part of a revision of Pfam families because they no longer believe it is a true protein family. Unfortunately, when domains are removed from Pfam, there are no mechanisms to revise genes that were validated as coding based on these Pfam domains.

The now defunct domain seems to have been instrumental in the prediction of 1178 novel human coding genes by the CHESS database [BioRxiv: https://doi.org/10.1101/332825]. These novel predictions were based on RNAseq evidence and similarity to known proteins. More than half of these novel genes were similar to one of just nine UniProtKB proteins (eight human and one chimp). The alignment of the nine proteins (Supplementary Figure S15) shows that they are all closely related.

Two of these proteins came from *GVQW1* and *GVQW2*. Although *GVQW1* and *GVQW2* are in the process of being reclassified by GENCODE, they are still present in the UniProtKB and RefSeq reference sets. *GVQW1* and *GVQW2* are transposon-based, so it is reasonable to assume that all nine sequences are derived from Alu sequences (indeed isoforms from both *C16orf89* and *C9orf85* are among

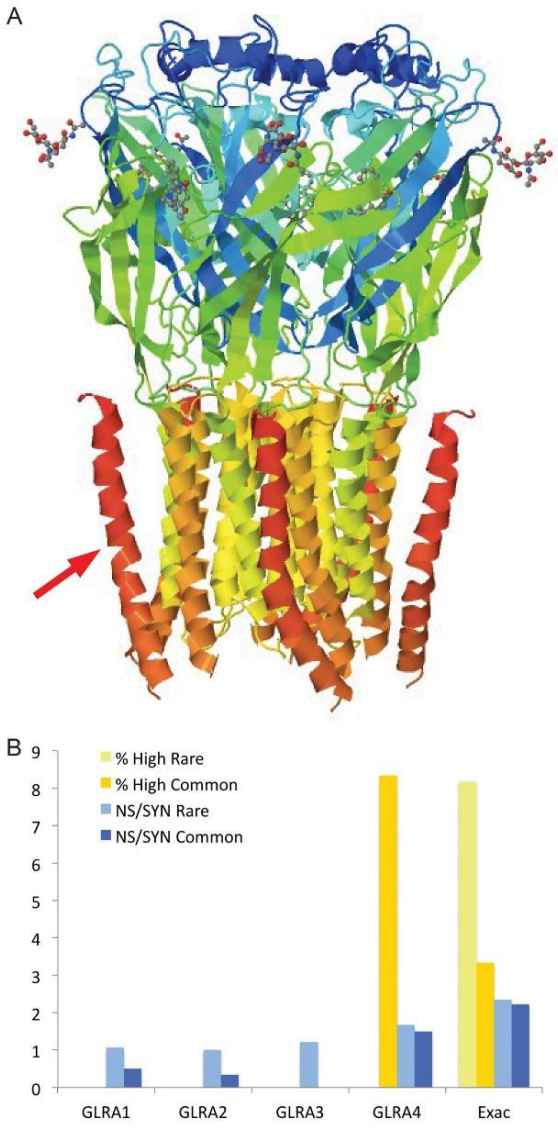


Figure 7. *GLRA4* loss of trans-membrane helix and genetic variation. (**A**) The cryo-EM structure of the *GLRA1* kinase domain from *Danio rerio* (PDB code: 3JAD), which is 80% sequence identical to human *GLRA4*. In *GLRA4*, the premature stop codon would lead to the loss of the dark orange trans-membrane helices in the figure (one of which is marked with a red arrow). From the point of view of the pore, this would mean the loss of five of the twenty helices, albeit the helices which are furthest away from the inside of the pore. This would almost certainly destabilize the pore, and would probably have considerable effect on the function. It would also leave the C-terminals of the protein on the cytoplasmic side instead of the extra-cellular side. (**B**) The percentage of high impact variants (yellow) and non-synonymous/synonymous ratios (blue) for the GLRA gene family. The percentage of high impact variants and non-synonymous/synonymous ratios for *GLRA4* from Exac are marked as "Exac". The darker colors show the values for common variants and the lighter shades show the values for rare variants. *GLRA4* does not have the same variation pattern as the other four genes.

the nine proteins) and are therefore erroneously annotated as coding. This in turn suggests that the novel sequences in CHESS predicted as coding because of their similarity to the nine proteins (more than half of the novel coding genes in CHESS) will also be transposon-related.

Clearly, misannotating genes as protein coding can have important downstream effects on a wide range of databases that depend on reliable predictions of coding genes. The CHESS database's prediction of hundreds of new coding genes based on a defunct, transposon-linked Pfam domain underscores how easily misclassifications can proliferate.

A number of other dead Pfam domains may have been used to help validate the potential non-coding genes, for example *C19orf48* and *C1orf145*. We also ran the Pfam-based tool Antifam (52) to check whether any genes had similarity to known non-coding domains and we found evidence for two more genes, *AC079355.1* and *AC118758.1*, which mapped to the same 'spurious ORF' domain. Both coding genes are automatic predictions.

DISCUSSION

There are >22 000 genes annotated as coding across the Ensembl/GENCODE, RefSeq and UniProtKB human proteomes. While manual annotators agree on >19 000 genes, one in eight of these genes are classified differently in at least one of the reference sets. Evidence from various sources suggests that many of the genes classified differently across the three reference sets are unlikely to code for essential proteins; these genes have poor UniProtKB evidence scores, a higher proportion of the most damaging germline variants and non-synonymous to synonymous substitution ratios that suggest many are under neutral selection.

To study differences between these genes and genes annotated as coding in all three reference sets we defined a set of 16 potential non-coding features from the Ensembl/GENCODE reference set. More than 11% of Ensembl/GENCODE coding genes had at least one potential non-coding feature and there were profound differences between these genes and the remaining 89% of genes. Only a handful of potential non-coding genes had reliable proteomics or antibody evidence, most had significantly lower transcript expression and their transcripts were detected in very few tissues. Non-coding genes are known to have much lower levels of expression than coding genes (53), so the fact that so many potential non-coding genes had low or negligible RNAseq expression levels supports the possibility that many will not code for proteins.

Data from genetic variation studies showed that potential non-coding genes had many more copy number variants, a much higher rate of potentially damaging variants, and larger non-synonymous to synonymous substitution ratios. The pattern of variants suggested that many of these genes are under neutral selection. Since neutral selection is not typical of coding genes, this reinforces the likelihood that many potential non-coding genes will not code for functional proteins.

There are 4234 coding genes that could be considered potentially non-coding across the three reference sets. These genes are either annotated differently across the three reference sets or were flagged as potential non-coding (Supplementary Figure S4). If the majority do not code for proteins, as the genetic variation patterns suggest, the number of coding genes will be much closer to the 19 446 genes common to the three reference sets than to the 22 210 genes in the union of those sets. However, it is still early to speculate on the precise number of coding genes because it is impossible to know how many potential non-coding genes will be reclassified by manual annotators, and because there is a steady trickle of new coding genes being annotated (54).

Human population variation data shows that two types of genes in particular appeared not under selection pressure and were therefore unlikely to code for functional proteins. The first are automatic gene predictions, genes in which all gene models are predicted, which make up approximately 1% of Ensembl/GENCODE coding genes and more than 4% of RefSeq coding genes. Our results suggest that these genes are adding little to the human reference annotation. The second group of genes are likely pseudogenes. Pseudogenes form the largest group of non-coding annotations and are especially difficult to distinguish from coding genes but have the clearest evidence for neutral selection of all the potential non-coding features. Likely pseudogenes are particularly prevalent in the UniProtKB unique subset.

Pseudogenes highlight the difficulties that manual annotators face when interpreting the available data (55). Most pseudogenes derive from protein coding genes, either by duplication or retrotransposition, and as a result often have large intact ORFs and protein-like features. In addition recent duplications usually have few obviously deleterious mutations, making the distinction between coding and pseudogene even more difficult. The Ubiquitin carboxyl-terminal hydrolase 17 family has 26 close to identical members, but while non-synonymous to synonymous ratios suggest that most or all are pseudogenes, they are all annotated as coding because there is no clear way of discriminating between them.

Experimental evidence is often ambiguous for many pseudogenes. Negative evidence (evidence to show that a gene does not code for proteins) does not exist, antibodies are rarely sufficiently specific to distinguish similar proteins and proteomics experiments can easily confuse similar peptides because of single-amino acid variations or post-translational modifications. Indeed, a number of the potential non-coding genes detected in the proteomics experiments may be false positive identifications. For example PeptideAtlas validates two peptides for potential non-coding gene *FO538757.2*. The two peptides identified for *FO538757.2* are just one amino acid different from the equivalent peptides from *WASH1*, a likely coding gene. Indeed UniProtKB annotates both these single amino acid differences as known *WASH1* sequence conflicts. It is more than probable that the peptides we detected for *FO538757.2* really came from *WASH1*. Here we should point out that although some identifications will be false positives, many potential non-coding genes, such as SEMG1 and SEMG2, sperm-specific potential non-coding genes with a primate origin, were identified with strong peptide evidence.

The increase in genetic variation data (42,51) should provide valuable support for manual annotators in this sense, though genetic diversity is not infinite and it will not be suitable for all genes. Most *bona fide* coding genes should

have very few high impact variants in common alleles and should have non-synonymous to synonymous ratios that are lower for common allele than they are for rare alleles. We have used genetic variation data to flag a number of possible pseudogenes that were not caught by our potential non-coding features (for example *PLK5*, *GLRA4*).

Over the years since the human genome sequence was released (4) rigorous manual annotation has brought us considerably closer to a final catalogue of human coding genes and annotators agree for more than 85% of coding genes. The final 12% of genes, those with the most conflicting evidence, will be more difficult to classify. One useful source of information to discriminate coding from non-coding genes makes use of the recent increase in the number of annotated mammalian genomes (27). With time and with more extensive data, large-scale genetic variation studies could also be a powerful tool to aid in the annotation of coding genes.

In order to flag potential non-coding genes we have built a pipeline that updates with the Ensembl/GENCODE reference set. This approach is a highly practical means of informing the curation of the human genome. The set of human coding genes needs to be as complete as possible for biomedical experiments, but inevitably some genes will be misannotated as coding. Once a gene has entered a reference set it may be propagated in large-scale databases and its coding potential may end up being validated via circular annotation. Detecting errors, retracing steps and rescinding the coding status of a gene once it is annotated as coding is a difficult process, so a system to catch and label genes that have conflicting or insufficient coding support is useful. The pipeline will be used to help pinpoint potential non-coding genes in the Ensembl/GENCODE human reference set. However, the approach could be made available for use by other annotation initiatives and could be extended to the annotation of other species. In fact, a pipeline has already been developed for the mouse reference set. Future releases of these analyses will be made publicly available.

Manual curators from the three main reference databases will investigate and debate the coding potential of these potential non-coding genes. It is important to note that while many potential non-coding genes will be reclassified, those that have evidence of coding capability will be maintained in the reference set. In addition a number of genes with conflicting evidence or insufficient evidence to determine coding status one way or another are also likely to be remain in the reference set. It may be possible to flag this second set of genes as potentially non-coding or pseudogene, while maintaining them as coding in the reference set.

Even if just half of these the potential non-coding genes we have highlighted turn out to be non-coding, this would clearly have a substantial impact on a range of fields. In particular, overestimating the number of coding genes inevitably complicates large-scale biomedical experiments, especially those that involve the mapping of disease-related variations to human genes. The more potential non-coding genes that are classified as coding as part of any analytical process, the noisier the results.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Iakes Ezkurdia and Jon Mudge for their input on this paper.

FUNDING

National Institutes of Health [2 U41 HG007234 to I.J., L.M., J.M.R. and M.L.T., R01 HG004037 to I.J.]. Funding for open access charge: NIH [2 U41 HG007234].
Conflict of interest statement. None declared.

REFERENCES

1. Harrison,P.M., Kumar,A., Lang,N., Snyder,M. and Gerstein,M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.*, **30**, 1083–1090.
2. Liang,F., Holt,I., Pertea,G., Karamycheva,S., Salzberg,S.L. and Quackenbush,J. (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.*, **24**, 239–240.
3. Wright,F.A., Lemon,W.J., Zhao,W.D., Sears,R., Zhuo,D., Wang,J.P., Yang,H.Y., Baer,T., Stredney,D., Spitzner,J. *et al.* (2001) A draft annotation and overview of the human genome. *Genome Biol.*, **2**, RESEARCH0025.
4. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
5. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
6. Southan,C. (2004) Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics*, **4**, 1712–1726.
7. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
8. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
9. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
10. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
11. Clamp,M., Fry,B., Kamal,M., Xie,X., Cuff,J., Lin,M.F., Kellis,M., Lindblad-Toh,K. and Lander,E.S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 19428–19433.
12. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**(Suppl 1), 1–9.
13. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
14. Rolland,T., Taşan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
15. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. and Aebersold,R. (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.
16. Goodstadt,L. and Ponting,C.P. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.*, **2**, e133.
17. Church,D.M., Goodstadt,L., Hillier,L.W., Zody,M.C., Goldstein,S., She,X., Bult,C.J., Agarwala,R., Cherry,J.L., DiCuccio,M. *et al.* (2009) Mouse Genome Sequencing Consortium. Lineage-specific

biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, **7**, e1000112.
18. Ezkurdia,I., Juan,D., Rodriguez,J.M., Frankish,A., Diekhans,M., Harrow,J., Vazquez,J., Valencia,A. and Tress,M.L. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.*, **23**, 5866–5878.
19. Yates,B., Braschi,B., Gray,K., Seal,R., Tweedie,S. and Bruford,E. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
20. Menashe,I., Aloni,R. and Lancet,D. (2006) A probabilistic classifier for olfactory receptor pseudogenes. *BMC Bioinformatics*, **7**, 393.
21. Buljan,M., Frankish,A. and Bateman,A. (2010) Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.*, **11**, R74.
22. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database*, **2016**, baw053.
23. Roux,J and Robinson-Rechavi,M. (2011) Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res.*, **21**, 357–363.
24. Huerta-Cepas,J., Dopazo,H., Dopazo,J. and Gabaldón,T. (2007) The human phylome. *Genome Biol.*, **8**, R109.
25. Cannarozzi,G., Schneider,A. and Gonnet,G. (2007) A phylogenomic study of human, dog, and mouse. *PLoS Comput. Biol.*, **3**, e2.
26. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
27. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.
28. Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A. and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.*, **46**, D213–D217.
29. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
30. Lopez,G., Maietta,P., Rodriguez,J.M., Valencia,A. and Tress,M.L. (2011) firestar–advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
31. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.
32. Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.
33. Käll,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
34. Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1197.
35. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
36. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419.
37. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Fritz,M.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
38. Sudmant,P.H., Mallick,S., Nelson,B.J., Hormozdiari,F., Krumm,N., Huddleston,J., Coe,B.P., Baker,C., Nordenfelt,S., Bamshad,M. *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**, aab3761.
39. Zarrei,M., MacDonald,J.R., Merico,D. and Scherer,S.W. (2015) A copy number variation map of the human genome. *Nat. Rev. Genet.*, **16**, 172–183.
40. Handsaker,R.E., Van Doren,V., Berman,J.R., Genovese,G., Kashin,S., Boettger,L.M. and McCarroll,S.A. (2015) Large multiallelic copy number variations in humans. *Nat. Genet.*, **47**, 296–303.
41. Abyzov,A., Li,S., Kim,D.R., Mohiyuddin,M., Stütz,A.M., Parrish,N.F., Mu,X.J., Clark,W., Chen,K., Hurles,M. *et al.* (2015) Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.*, **6**, 7256.
42. 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
43. NCBI Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
44. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S, Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
45. Deutsch,E.W., Overall,C.M., Van Eyk,J.E., Baker,M.S., Paik,Y.K., Weintraub,S.T., Lane,L., Martens,L., Vandenbrouck,Y., Kusebauch,U. *et al.* (2016) Human proteome project mass spectrometry data interpretation guidelines 2.1. *J. Proteome Res.*, **15**, 3961–3970.
46. Tress,M.L., Abascal,F. and Valencia,A. (2017) Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.*, **42**, 98–110.
47. Tress,M.L., Abascal,F. and Valencia,A. (2017) Most alternative isoforms are not functionally important. *Trends Biochem. Sci.*, **42**, 408–410.
48. Talukder,A.H., Meng,Q. and Kumar,R. (2006) CRIPak, a novel endogenous Pak1 inhibitor. *Oncogene*, **25**, 1311–1319.
49. Kortvely,E., Hauck,S.M., Behler,J., Ho,N. and Ueffing,M. (2016) The unconventional secretion of ARMS2. *Hum. Mol. Genet.*, **25**, 3143–3151.
50. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
51. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
52. Eberhardt,R.Y., Haft,D.H., Punta,M., Martin,M., O'Donovan,C. and Bateman,A. (2012) AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database*, **2012**, bas003.
53. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
54. Wright,J., Mudge,J.M., Weisser,H., Barzine,M.P., Gonzalez,J.M., Brazma,A., Choudhary,J.S. and Harrow,J. (2016) Improving GENCODE reference gene annotation using a high-stringencyproteogenomics workflow. *Nat. Commun.*, **7**, 11778.
55. Bruford,E.A., Lane,L. and Harrow,J. (2015) Devising a consensus framework for validation of novel human coding loci. *J. Proteome Res.*, **14**, 4945–4948.

(English) RESULTS AND MATERIAL & METHODS: Third article

An analysis of tissue-specific alternative splicing at the protein level

The alternative splicing of messenger RNA is a process that results in a single gene coding for multiple proteins (Pan *et al.*, 2008; Wang, E. T. *et al.*, 2008). In this process, exons can be included or excluded in combination to create a diverse range of mRNA transcripts (Nilsen & Graveley, 2010). Moreover, studies have described the role that alternative splicing plays in tissue differentiation at transcript level (Merkin *et al.*, 2012; Wang, E. T. *et al.*, 2008).

Here we carried out an analysis of a large-scale proteomics study (Kim, M.-S. S. *et al.*, 2014) that comprised 30 human tissues and hematopoietic cells, and compared that with the results of a large-scale RNA-seq analysis carried out by the Human Protein Atlas (Uhlén *et al.*, 2015). The RNA-seq analysis was performed on 36 different tissues and covers similar tissues to the Kim *et al.* analysis, though this analysis did not investigate fetal tissues or blood cells. We compared the two large-scale experiments by concentrating on the tissues present in both sets of experiments and by grouping the tissues by type.

As a first step we used the peptides identified in the Kim *et al.* analysis to define the set of alternative splice events that would be used in the analysis. We required that each side of a splice event was supported by a minimum of three peptides from different experiments (PEDs). Although we initially searched for tissue specificity using the tissues from the proteomics experiment, much of the analysis was carried out with the 30 tissues pooled into 10 groups of related tissues. This was done to amplify any signal. In order to compare both proteomics and transcriptomics analysis, we also grouped tissues from the large-scale RNA-seq analysis into 12 groups.

We were able to define a set of 255 alternative splicing events (ASE255) that came from 217 genes and in total 95 of these events had evidence for either tissue or group-specific splicing (37.3%) at protein level. From the transcriptomics analysis, we were able to map sufficient RNA-seq reads to both sides of the splicing event for 248 of the 255 events. There was evidence for tissue group-specific differences in transcript expression for a total of 159 of these 248 events (62.9%).

More than 50% of events that had significant differences in group-specific splicing at protein level were in nervous tissues (frontal cortex, fetal brain, spinal cord and retina) followed by muscle (almost 30%). Moreover, when we compared the groups of tissues between the proteomics analysis and transcriptomics analysis, the correlation between supporting PEDs and supporting reads was highest in nervous (0.799) and muscle (0.748) tissues and lowest in reproductive tissues (0.413).

We calculated disorder for all regions that differed between the main and alternative isoforms, and for both regions involved in the swap in the case of insertions/deletions (indels). However, there was no indication that disorder was related to tissue specificity either at the protein level, where 37.7% of tissue specific alternative regions were disordered against 48.1% of non-tissue specific regions, or at the transcript level.

We classified the ASE255 set according to two different criteria, the mechanism of the splicing process (transcript level) and the effect the splice event has on the resulting proteins (protein level). Then, we classified the effect at proteomics level, tissue-specific and non-specific events were more or less proportionally distributed within each even type. Homologous exon substitutions made up a third of all events with tissue-specific differences, with a Fisher exact

test of 0.0062 (against indels). By way of contrast, Micro-indel splice events had significantly fewer non-tissue specific events (Fisher exact test of 0.0018 against indels).

We also manually curated the relative age of the ASE255 set based on cross-species evidence. Manually curated event ages were defined as primate-derived, as from the eutheria/theria clades, as from the tetrapoda clade and as “ancient” (evolved before the sarcopterygii clade, more than 400 million years ago). More than half of the alternative events in the set evolved more than 400 million years ago and only 7.8% of the alternative events derived from the primate clade. Tissue-specific splice events were even more conserved. Almost three quarters (73.7%) of events with evidence of tissue specificity at the proteomics level evolved more than 400 million years ago and none of the tissue specific events were of primate-derived.

Tissue specificity at the transcript level also seemed to be associated with the conservation of splice events. Events that were tissue-specific in the transcriptomics analysis were older than events without significant tissue-specific. Remarkably, more than 95% of tissue specific events in which there is agreement between proteomics and RNA-seq analyses evolved prior to the ancestors of lobe-finned fish.

In order to compare the conservation of splice events against the whole genome we also estimated the relative age of alternative exons (defined by the APPRIS database). We found that 76% of alternative exons in the human genome appeared in the primate clade, within the last 90 million years, while just 5.7% were more than 400 million years old.

Finally, we calculated the significantly enriched GO terms for those genes with tissue-specific alternative splicing events. We found a strong relation between cytoskeleton-related genes, tissue specificity and conservation. Cytoskeleton genes with events that were tissue-specific at the proteomics level had a much higher proportion (82.1%) of events that evolved before or during the vertebrate clade.

Michael Tress conceived of the presented idea. Jose Manuel worked out almost all of the technical details, and performed the numerical calculations. Tomas di Domenico carried out the RNA-seq search and the search and post-processing of proteomics data. Fernando Pozo post-processed the RNA-seq data. Michael Tress encouraged Jose Manuel Rodríguez to investigate and supervised the findings of this work. Michael Tress and Jose Manuel Rodríguez wrote the manuscript. Jesús Vázquez and Fernando Pozo provided critical feedback and helped shape the manuscript.

(Español) RESULTADOS Y MATERIALES Y MÉTODOS: Tercer artículo

Un análisis de empalme alternativo específico de tejido a nivel de proteína

El empalme alternativo del ARN mensajero es un proceso que da como resultado un único gen que codifica a múltiples proteínas (Pan *et al.*, 2008; Wang, E. T. *et al.*, 2008). En este proceso, los exones se pueden combinar para crear una amplia gama de transcripciones de ARNm (Nilsen y Graveley, 2010). Además, estudios han descrito el papel que juega el empalme alternativo en la diferenciación de tejidos a nivel de transcripción (Merkin *et al.*, 2012; Wang, E. T. *et al.*, 2008).

En este trabajo, llevamos a cabo un análisis de un estudio de proteómica a gran escala (M.-SS Kim *et al.*, 2014) compuesto por 30 tejidos humanos y células hematopoyéticas, y lo comparamos con los resultados de un análisis de RNA-seq publicado por *Human Protein Atlas* (Uhlén *et al.*, 2015). El análisis de RNA-seq se realizó en 36 tejidos diferentes y cubre tejidos similares al análisis de Kim *et al.* Comparamos los dos estudios centrándonos en los tejidos presentes en ambos experimentos y agrupando los tejidos por tipo.

Como primer paso utilizamos los péptidos identificados en el análisis de Kim *et al.* para definir el conjunto de eventos de empalme alternativos que se utilizarían. Requerimos que cada lado del evento estuviera respaldado por un mínimo de tres péptidos de diferentes experimentos (PED). Aunque inicialmente buscamos la especificidad utilizando tejido, gran parte del análisis se llevó a cabo con los 30 tejidos agrupados en 10 grupos de tejidos relacionados. Esto se hizo para amplificar cualquier señal. Para comparar el análisis proteómico y transcriptómico, también agrupamos los tejidos del análisis de RNA-seq en 12 grupos.

Pudimos definir a nivel de proteína un conjunto de 255 eventos de empalme alternativo (ASE255) que provenían de 217 genes y en total 95 de estos eventos tenían evidencia de especificidad de tejido o de grupo (37,3%). A partir del análisis transcriptómico, 248 de los 255 eventos mapearon con suficientes lecturas RNA-seq a ambos lados del evento de empalme. Hubo evidencia de especificidad de grupos de tejidos en la expresión de la transcripción para un total de 159 de estos 248 eventos (62,9%).

Más del 50% de los eventos a nivel de proteína, tuvieron diferencias significativas en la especificidad del grupo de tejido que se produjeron en nervio; seguidos por músculo con casi el 30%. Además, cuando comparamos los grupos de tejidos entre el análisis proteómico y el análisis transcriptómico, la correlación entre los PED y las lecturas RNA-seq fue mayor en los tejidos nerviosos (0,799) y musculares (0,748) y más baja en los tejidos reproductivos (0,413).

Calculamos el desorden para todas las regiones que difirieron entre las isoformas principal y alternativa, y para ambas regiones involucradas en el intercambio de indels (inserciones/deleciones). Sin embargo, no hubo indicios de que el trastorno estuviera relacionado con la especificidad de tejido ni a nivel de proteína, donde el 37,7% de las regiones alternativas específicas de tejido estaban desordenadas, ni a nivel de transcripción, con un 48,1%.

Clasificamos el conjunto ASE255 de acuerdo con dos criterios diferentes, el mecanismo del *splicing* (nivel de transcripción) y el efecto que tiene el evento de *splicing* en las proteínas resultantes (nivel de proteína). Luego, clasificamos el efecto a nivel proteómico. Los eventos específicos de tejido, y los no específicos, se distribuyeron más o menos proporcionalmente

dentro de cada tipo. Las sustituciones de exones homólogos constituyeron un tercio de todos los eventos con diferencias específicas de tejido, haciendo una prueba Fisher de 0,0062 (contra indels). Por el contrario, los eventos de empalme de Micro-indel tuvieron significativamente menos eventos no específicos de tejido (prueba Fisher de 0,0018 contra indels).

También seleccionamos manualmente la edad relativa del conjunto ASE255 en función de la conservación en especies. Las clasificaciones de las edades se definieron como derivadas de primates, a partir de los *eutheria*/*theria*, a partir *tetrápoda* y "antiguas" (antes de *sarcopterygii*, más de 400 millones de años). Más de la mitad de los eventos alternativos en ASE255 evolucionaron hace más de 400 millones de años y solo el 7,8% de los eventos alternativos derivaron de los primates. Los eventos de empalme específicos de tejido estaban más conservados. El 73,7% de los eventos con evidencia de especificidad de tejido a nivel proteómico evolucionaron hace más de 400 millones de años y ninguno de este tipo de eventos fue derivado de primates. La especificidad del tejido a nivel de transcripción también pareció estar asociada con la conservación. Los eventos específicos de tejido en el análisis transcriptómico fueron más antiguos que los eventos sin tejido específicos. Sorprendentemente, más del 95% de los eventos específicos de tejido en los que existe un acuerdo entre la proteómica y los análisis de RNA-seq evolucionaron antes que los ancestros de los peces con aletas lobuladas.

Para comparar la conservación de los eventos de empalme con todo el genoma, también estimamos la edad relativa de los exones alternativos (definidos por APPRIS). Descubrimos que el 76% de los exones alternativos en el genoma humano aparecieron en primates, en los últimos 90 millones de años, mientras que solo el 5,7% tenían más de 400 millones de años.

Finalmente, calculamos los términos GO de los genes significativamente enriquecidos. Encontramos una fuerte relación entre los genes relacionados con el citoesqueleto, la especificidad del tejido y la conservación. Los genes del citoesqueleto con eventos que eran específicos de tejido a nivel proteómico, tenían una proporción mucho mayor (82,1%) de eventos que evolucionaron antes o durante los vertebrados.

Michael Tress concibió la idea presentada. José Manuel Rodríguez resolvió casi todos los detalles técnicos y realizó los cálculos numéricos. Tomas di Domenico llevó a cabo la búsqueda de RNA-seq y la búsqueda y posprocesamiento de datos proteómicos. Fernando Pozo procesó los datos de RNA-seq. MT animó a JMR a investigar y supervisó los hallazgos de este trabajo. MT y JMR redactaron el manuscrito. Jesús Vázquez y FP brindaron comentarios críticos y ayudaron a dar forma al manuscrito.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

An analysis of tissue-specific alternative splicing at the protein level

Jose Manuel Rodriguez[1], **Fernando Pozo**[2], **Tomas di Domenico**[2], **Jesus Vazquez**[1,3], **Michael L. Tress**[2]*

1 Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Calle Melchor Fernandez, Madrid, Spain, **2** Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Calle Melchor Fernandez, Madrid, Spain, **3** Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Madrid, Spain

* mtress@cnio.es

OPEN ACCESS

Citation: Rodriguez JM, Pozo F, di Domenico T, Vazquez J, Tress ML (2020) An analysis of tissue-specific alternative splicing at the protein level. PLoS Comput Biol 16(10): e1008287. https://doi.org/10.1371/journal.pcbi.1008287

Editor: Christine A. Orengo, University College London, UNITED KINGDOM

Received: February 10, 2020

Accepted: August 25, 2020

Published: October 5, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008287

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This study was supported by the National Institutes of Health (https://www.nih.gov) grant

Abstract

The role of alternative splicing is one of the great unanswered questions in cellular biology. There is strong evidence for alternative splicing at the transcript level, and transcriptomics experiments show that many splice events are tissue specific. It has been suggested that alternative splicing evolved in order to remodel tissue-specific protein-protein networks. Here we investigated the evidence for tissue-specific splicing among splice isoforms detected in a large-scale proteomics analysis. Although the data supporting alternative splicing is limited at the protein level, clear patterns emerged among the small numbers of alternative splice events that we could detect in the proteomics data. More than a third of these splice events were tissue-specific and most were ancient: over 95% of splice events that were tissue-specific in both proteomics and RNAseq analyses evolved prior to the ancestors of lobe-finned fish, at least 400 million years ago. By way of contrast, three in four alternative exons in the human gene set arose in the primate lineage, so our results cannot be extrapolated to the whole genome. Tissue-specific alternative protein forms in the proteomics analysis were particularly abundant in nervous and muscle tissues and their genes had roles related to the cytoskeleton and either the structure of muscle fibres or cell-cell connections. Our results suggest that this conserved tissue-specific alternative splicing may have played a role in the development of the vertebrate brain and heart.

Author summary

We manually curated a set of 255 splice events detected in a large-scale tissue-based proteomics experiment and found that more than a third had evidence of significant tissue-specific differences. Events that were significantly tissue-specific at the protein level were highly conserved; almost 75% evolved over 400 million years ago. The tissues in which we found most evidence for tissue-specific splicing were nervous tissues and cardiac tissues. Genes with tissue-specific events in these two tissues had functions related to important cellular structures in brain and heart tissues. These splice events may have been essential for the development of vertebrate heart and muscle. However, our data set may not be

number 2 U41 HG007234 (FP and MLT), the Spanish Ministry of Science, Innovation and Universities (https://www.ciencia.gob.es) grants BIO2015-67580-P and PGC2018-097019-B-I00 (JMR and JV), the Carlos III Institute of Health-Fondo de Investigación (https://www.isciii.es) Sanitaria grant PRB3, IPT17/0019 - ISCIII-SGEFI / ERDF, ProteoRed (JMR and JV), the Fundació MaratóTV3 (https://www.ccma.cat) grant 122/C/2015 (JMR and JV) and "la Caixa" Banking Foundation (https://obrasociallacaixa.org) project code HR17-00247 (JMR and JV). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

representative of alternative exons as a whole. We found that most tissue specific splicing was strongly conserved, but just 5% of annotated alternative exons in the human gene set are ancient. More than three quarters of alternative exons are primate-derived. Although the analysis does not provide a definitive answer to the question of the functional role of alternative splicing, our results do indicate that alternative splice variants may have played a significant part in the evolution of brain and heart tissues in vertebrates.

Introduction

Almost all multi-exon genes are able to undergo alternative splicing [1,2] via a range of mechanisms which include exon skipping, alternative splice site usage and alternative promoter and poly-A usage. This is reflected in the human reference set; at present human coding genes are annotated with an average of four distinct gene products [3]. Recent studies suggest that human coding genes generate on average more than ten alternative transcripts [4,5]. Assuming that all, or almost all of these transcripts are translated into functional alternative splice isoforms, we would expect the overall protein population to increase 10-fold from 20,000 (the number of human coding genes) to 200,000. This increase alone would have profound biological consequences. However, most proteins do not work in isolation but interact with other proteins, often as part of large complexes. If we take into account all the possible interactions of these distinct proteins [6], we would be likely to see an exponential increase in the number of cellular functions.

There has been much investigation at the transcript level to try to elucidate a role for alternative splicing and there is some indication that it may play a role in tissue differentiation. Approximately two thirds of alternative splice events have been shown to have tissue-specific differences. Wang *et al* [1] identified over 22,000 tissue-specific alternative transcript events and showed that between 47 and 65% of alternative events were tissue-specific depending on the type of splice event, while Gonzalez-Porta et al [7] found that the major transcript varied according to conditions across more than 60% of coding genes.

However, it seems that not all tissue-specific splice patterns are conserved across species. Merkin *et al* [2] found that despite the abundant evidence for tissue specificity of alternative transcripts the patterns of tissue-specific alternative splicing were only conserved in a few tissues between mammalian species and birds. Reyes et al. analysed tissue-specific splicing at the transcript level across six primate species [8]. They found that only a small number of exons had conserved splicing patterns. These exons with conserved patterns were enriched in untranslated regions and the protein coding regions were enriched in disordered regions. Meanwhile most tissue-specific alternative exons differed in their usage between species. They postulated that the different usage of exons was behind the tissue-specific "rewiring" of protein-protein interaction networks postulated by many groups [9,10] that would be essential for morphological differences between different species.

More recently, results from the large-scale GTEx consortium found that 84% of the variance between tissues was due to gene expression rather than alternative splicing [11] with the strong suggestion that at least a certain proportion of tissue-specific alternative splicing is stochastic [11]. A re-analysis of the GTEx data [12] found that 50% of genes had tissue-specific transcripts, but that most tissue-dependent splicing events would not affect proteome complexity of the cell since they involved untranslated exons.

Little research has been carried out into tissue-specific alternative splicing at the protein level. Examples of protein level tissue specificity have been highlighted in analyses of individual

research papers [13,14], but there are no large-scale analyses of tissue specificity at the protein level. One reason for this is that proteomics experiments detect many fewer alternative isoforms than would be expected [15,16], It is not clear why it is so hard to detect alternative protein isoforms. Although most alternative transcripts seem to be processed by the ribosome [17], it has been shown that transcript level differences between species decrease post-translation [18]. Alternative isoforms that are not detected in proteomics experiments could be expressed in quantities too low for mass spectrometry detection, or in fewer tissues, they could have a shorter cellular half-life, or ribosome control mechanisms [19] could reduce their translation.

In vitro experiments have suggested that most alternative isoforms would lose or change more than 50% of their binding partners [6] relative to the main protein isoform [15,20]. Such gross changes in interaction partners suggest that most alternative isoforms would be more than just minor variants of the main isoforms. However, it is difficult to know to what extent such *in vitro* experiments are representative of the cellular proteome.

The large-scale proteomics study of 30 human tissues and hematopoietic cells carried out by Kim et al [21] remains the best source of tissue level proteomics data, in part because it was carried out with replicates. The data from the Kim experiments has been analysed on a number of occasions [22–24], however no study has investigated tissue-specific splicing of alternative isoforms in detail. The original study highlighted distinct isoforms of *FYN* protein tyrosine kinase in brain and haematopoietic cells [21], while Wright *et al* suggested that most tissue-specific alternative splicing was in testis without revealing details [22]. The other two studies detailed evidence for tissue-specific alternative splicing in just a few genes mostly localised to brain and heart tissues [23] or to heart and testis [24].

Here we carried out an *in-depth* study of tissue-specific alternative splicing in tissues from the Kim *et al* proteomics experiments and contrasted it with data from a large-scale transcriptomics analysis. We find that there is strong evidence for tissue-specific splicing at the protein level in a minority of genes, and that these tissue-specific protein isoforms are generally found in muscle or nervous tissues. Almost three quarters of tissue-specific splice events detected at the protein level are conserved all the way back to jawed vertebrates.

Results

Proteomics evidence for alternative splicing

We first used the peptides identified in the proteomics experiment to define the set of alternative splice events that would be used in the analysis. For this set we required that each side of a splice event be supported by a minimum of three PEDs (peptides from different experiments, see Materials and Methods section). Since each side of a splicing event is different, we defined the two sides as either "main" and "alternative". The main side of each splicing event is the side supported by most PEDs (in the case of the proteomics experiments), or most RNAseq reads (in the case of the transcriptomics experiments). From the peptide data we were able to define a set of 255 alternative splicing events that came from 217 genes (see **S1 Table**). This dataset (ASE255) was used for all subsequent analyses.

Evidence for tissue-specific splicing in proteomics experiments

Initially we searched for evidence of both tissue-specific and group-specific differences at the protein level for the ASE255 set. For each tissue or tissue group we compared the distribution of PEDs for the main and alternative sides of each splice event, and also between the events and the rest of the protein. When we compared expression at the tissue level only, 51 events had significant differences in expression in at least one of the 30 tissues. When tissues were

grouped, we found significantly different levels of expression for 87 splice events. In total 95 of the 255 events had evidence for either tissue or group-specific splicing (37.3%), while there were 43 events with significant differences at both tissue and group level.

At the level of individual tissues, frontal cortex (17 tissue-specific events, S1 Fig) had the highest evidence of significant tissue-specific splicing. Fetal and adult heart tissues also had more than ten splice events with evidence for significant tissue-specific splicing at the protein level. When tissues were combined into groups, nervous tissues (frontal cortex, fetal brain, spinal cord and retina) had the highest level of tissue-specific alternative splicing (Fig 1). In fact, more than 50% of the 87 events that had significant differences in group-specific splicing were group-specific in nervous tissues (Fig 1).

Tissue specificity by event type

We classified the 255 splice events in the ASE255 set according to two different criteria, the mechanism of the splicing process and the effect the splice event has on the resulting proteins. We divided the splicing mechanisms into six types, skipped exons, mutually exclusively spliced exons, alternative 5´splice sites, alternative 3' splice sites, alternative promoters and alternative poly-A. Skipped exons were most numerous (93 events) with alternative 3' splice sites a distant second (37 events).

We divided the effect the splicing has on the protein into seven groups. Firstly, deletions and insertions for which we had peptides for each side of the event were pooled as "Indels"

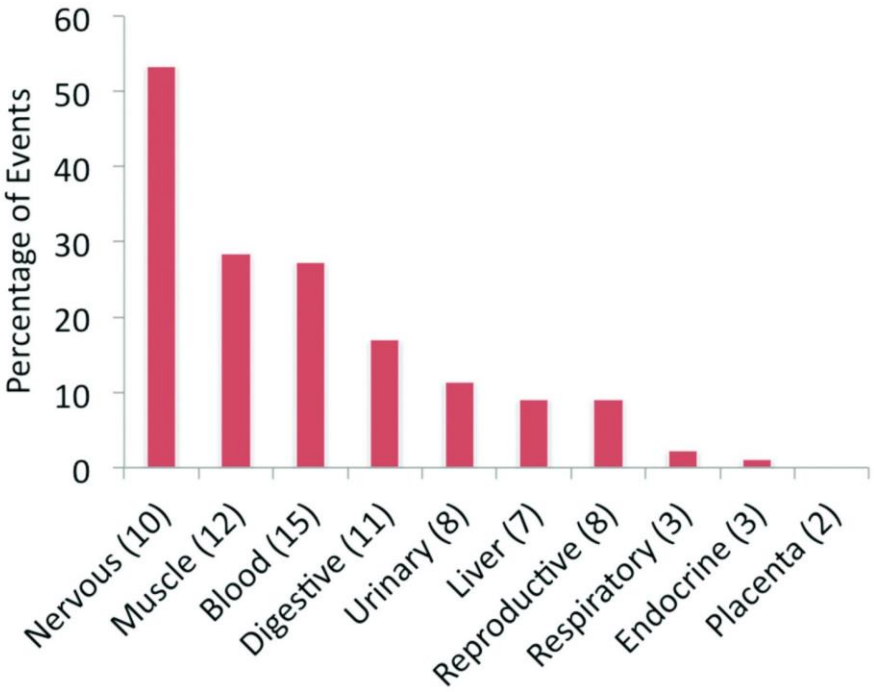


Fig 1. Tissue-specific alternative splicing events at the proteomics level. The percentage of significant tissue-specific alternative splicing events across the 10 proteomics tissue groups. The number of experiments for each tissue group is shown in the x-axis labels.

https://doi.org/10.1371/journal.pcbi.1008287.g001

since we did not know which was the alternative isoform. Meanwhile smaller indels, those that were smaller than four amino acid residues, were pooled as "Micro-Indels". Some of these mini-indels were produced by micro-exons, but most were generated by NAGNAG splicing [25]. If the substitutions were homologous [26], we pooled them in a different "Homologous Substitution" category because we have found that these exons often behave differently to most substitutions [16,23]. The remaining substitutions were tagged by their position in the protein sequence (C-terminal, N-terminal or internal), while protein isoforms that did not share any amino acid sequences were pooled into a seventh group ("Two Proteins"). A total of 104 events were classified as Indels. Almost three quarters of the indels were generated from skipped exons. There were just 5 events in the "Internal Substitution" category and 4 in the "Two Protein" category.

We calculated the proportion of each type of event among events with evidence of tissue specificity from the proteomics experiments and also among events that did not have evidence of protein-level tissue specificity. Two types of splicing mechanisms generated considerably more tissue specific events than non-tissue specific events (mutually exclusive exon and alternative poly A), while two mechanisms (alternative 5´splice sites and alternative 3' splice sites) produced a substantially lower proportion of tissue specific isoforms (Fig 2A).

When classified by the effect at the protein level, tissue-specific and non-specific events were more or less proportionally distributed within each class (Fig 2B), but two types of events had distributions that were significantly different from the others. Homologous substitution events were enriched among tissue-specific events. Homologous exon substitutions made up a third of all events with tissue-specific differences, with a Fisher exact test of 0.0062 (against indels). By way of contrast Micro-indel category splice events had significantly fewer non-tissue specific events (Fisher exact test of 0.0018 against indels). More than half of the events in the Micro-indel category were formed via alternative 3' splice sites.

Half of the homologous swaps were produced from mutually exclusive exons, while the other half were produced from alternative Poly A and alternative promoters. The proportion of tissue specific events among non-homologous alternative Poly A events (all of which are C-terminal substitutions) and non-homologous alternative promoter events (all N-terminal substitutions) decreases considerably once the homologous substitution events have been removed. This suggests that sequence homology was more important than the mechanism of action of the splicing process in the gain of tissue specific splicing.

Disorder and tissue specificity

Protein disorder has been strongly linked to alternative splicing [27] and to tissue-specific splicing in general [9]. We analysed the proportion of events with disorder in the ASE255 set and found that alternative exons in the set of splice events were enriched in disorder. A total of 43.6% of alternative exons were predicted to be disordered against 32.8% of the genes that made up the ASE255 set. However, there was no indication that disorder was related to tissue specificity either at the protein level, where 37.7% of tissue specific alternative regions were disordered against 48.1% of non-tissue specific regions, or at the transcript level (S2 Fig). Tissue specific skipped (cassette) exons are also depleted in predicted disordered regions in this set (S2 Fig).

Most tissue-specific splicing occurs in ancient splice events

We manually curated the relative age of each of the 255 splicing events based on cross species evidence. Manually curated event ages were defined as primate-derived (up to approximately 75 million years old), as evolved during the eutheria/theria clades (between 75 and 160 million years ago), as evolved during the tetrapoda clade (evolved after sarcopterygii, between 160 and

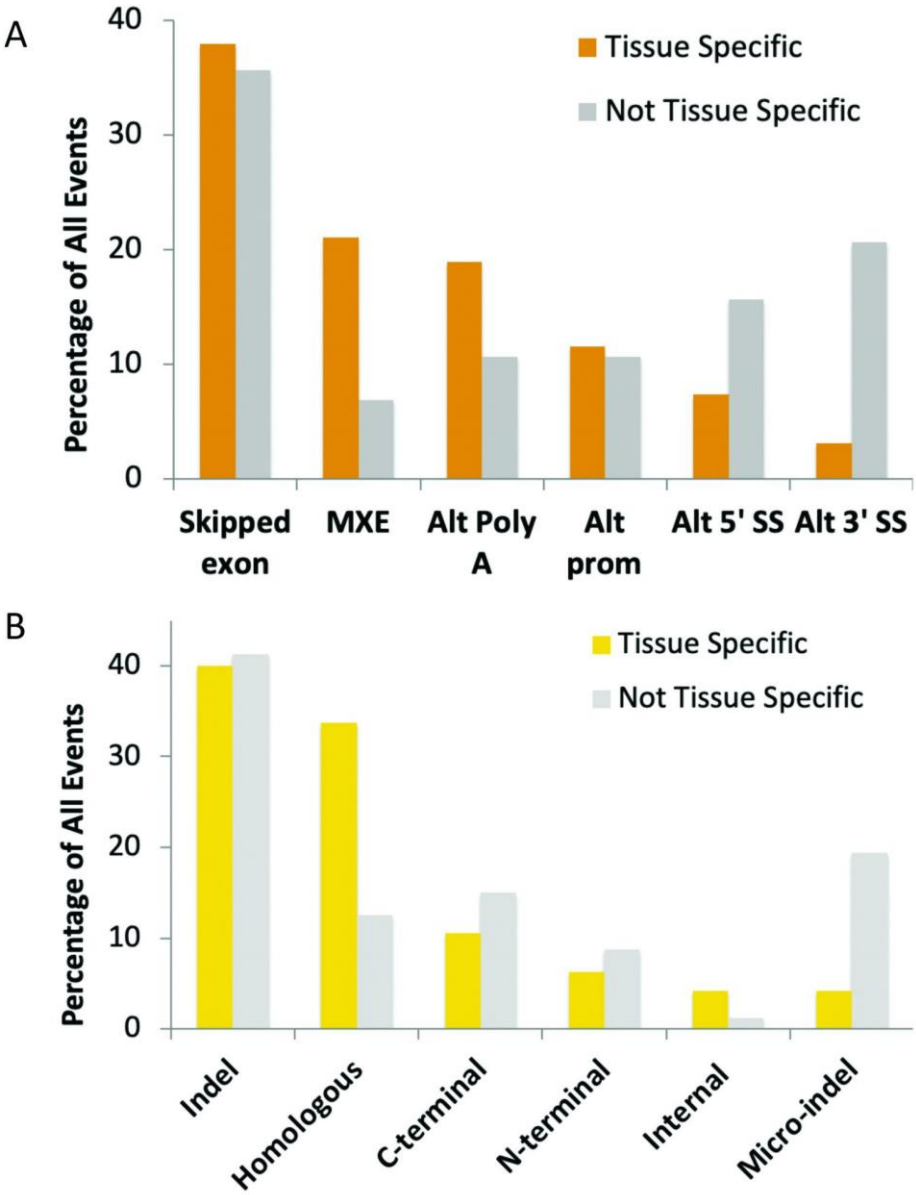


Fig 2. Tissue-specific expression at the proteomics level by type of splice event. (A) The breakdown of events by alternative splicing mechanism and tissue specificity, or lack of. The proportion of the 95 events that are tissue or group specific events are in orange. The proportions of event types among the 160 non-tissue specific events are shown in grey. (B) The relative proportions of tissue specificity of different protein level events. The proportion of the 95 tissue or group specific events that make up each event type is shown in yellow. The proportions of event types among the 160 non-tissue specific events are shown in light blue. All event types are defined in the main text.

https://doi.org/10.1371/journal.pcbi.1008287.g002

400 million years ago) and as ancient (evolved before the sarcopterygii clade, more than 400 million years ago).

In order to compare the results against the whole genome we also estimated the relative age of alternative exons. Alternative exons were defined from their annotations in the APPRIS database and we analysed just those exons that had a minimum of 42 bases (see Materials and Methods section). This automatic estimation of alternative exon age is an approximation, but it does provide an idea of the relative proportions of the four age groups among alternative exons in the genome. We found that 76% of alternative exons in the human genome appeared in the primate clade, within the last 90 million years, while just 5.7% were more than 400 million years old (Fig 3).

By way of contrast to annotated alternative exons, alternative splice events detected in proteomics experiments were considerably more conserved: more than half of the alternative events in the ASE255 set evolved more than 400 million years ago and only 7.8% of the alternative events in the ASE255 set derived from the primate clade. We have previously shown that proteins from ancient gene families are more likely to be detected in proteomics experiments [28] and that there is little reliable proteomics evidence for primate-derived coding genes [28,29]. Hence, it is not surprising that we also found most evidence for ancient splice events and little evidence of alternative splicing events derived from the primate clade.

Not only was the set of alternative events detected at the protein level enriched in ancient events, but tissue-specific splice events were even more conserved. Almost three quarters

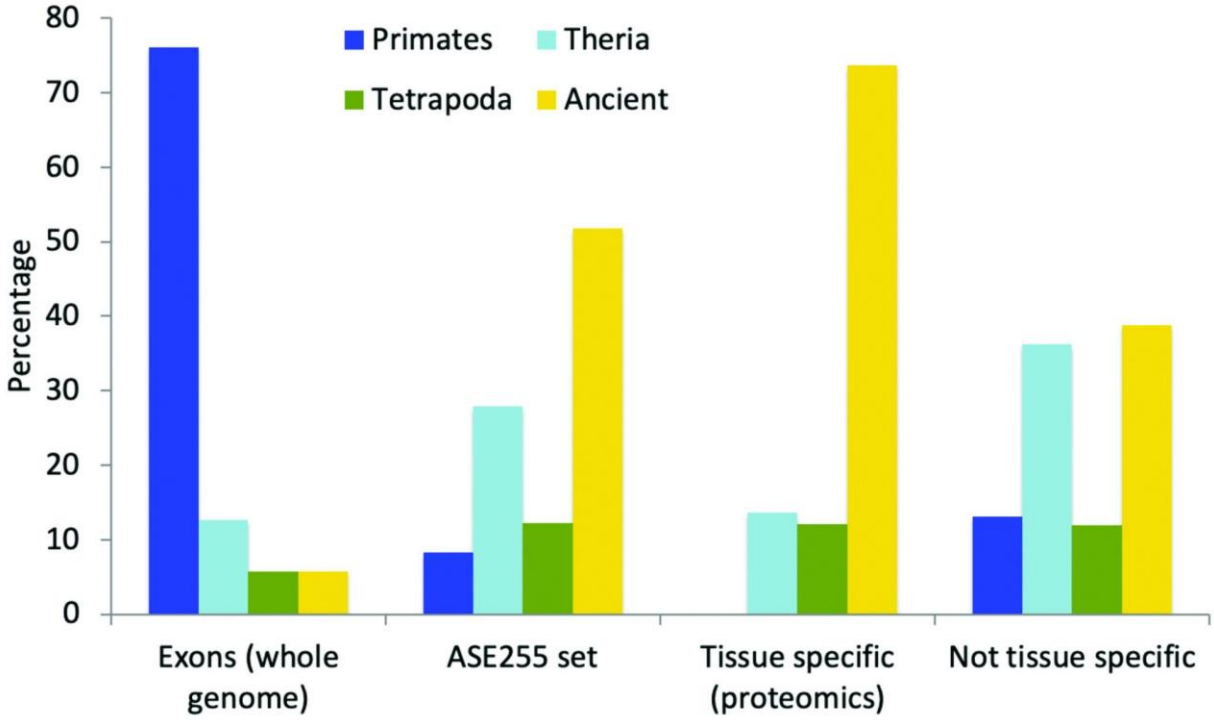


Fig 3. The age of alternative exons versus subsets of splicing events detected in proteomics experiments. "Exons (whole genome)" are all alternative exons in the human genome, "ASE255 set"is the set of 255 alternative splicing events detected in the proteomics analysis, "Tissue-specific (proteomics)" are the 95 events that have significant tissue or group-specific differences at the protein level and "Not tissue specific" are the 161 events that do not have tissue-specific enrichment in proteomics experiments.

https://doi.org/10.1371/journal.pcbi.1008287.g003

(73.7%) of events with evidence of tissue specificity at the proteomics level evolved more than 400 million years ago (Fig 4). At the same time there is no evidence at all for tissue-specific splicing of primate-derived splice events at the protein level (Fig 4). Tissue-specific alternative splicing events detectable at the proteomics level are highly conserved.

Functional relevance of tissue-specific splicing

We have previously found that alternative splice variants detected in proteomics experiments are enriched in functions related to the cytoskeleton [16]. We find similar results with the alternative splice events used in this analysis. This is not entirely surprising since proteomics analyses tend to be enriched in the most abundant proteins, including ribosome proteins and actin cytoskeleton-related proteins, and depleted in integral membrane proteins [30].

The top ten highest scoring GO terms for the 217 genes in the ASE255 set as a whole were all cytoskeleton-related and included *cell-cell adherens junction* (Benjamini-Hochberg adjusted q-value of 2.7E-08), *Z-disc* (6.8E-10), *structural constituent of muscle* (5.4E-10), *focal adhesion* (1.7E-07), and actin filament binding (1.5E-07). See S2 Table for more details. A total of 111 of the 217 ASE255 genes were labelled with terms related to the cytoskeleton. There was a strong relation between cytoskeleton-related genes, tissue specificity and conservation. Cytoskeleton genes with events that were tissue-specific at the proteomics level had a much higher proportion (82.1%) of events that evolved before or during the vertebrate clade (see S3 Fig). In fact, events in cytoskeleton genes were significantly more likely to be tissue specific (Fisher exact test, 0.0004) than non-cytoskeleton genes.

We also analysed the two subsets of events that had the most evidence of group-specific splicing at the protein level (Fig 1): those events with group-specific splicing at the protein level in nervous tissues (43 events from 37 genes) and those events with group-specific splicing at the protein level in muscle tissues (24 events, 18 genes). While the results were similar to the results for the ASE255 set, there were specific differences.

Terms for the nervous tissues specific events were related to adhesion, morphology and cellular communications and included *cadherin binding involved in cell-cell adhesion*" (0.002), *cell-cell adherens junction* (0.004), *stress fiber* (1.1E-04) and *plasma membrane* (4E-04), whereas terms for the events specific to muscle tissues were enriched in those terms more related to muscle, such as *Z-disc* (3.2E-05), *structural constituent of muscle* (2.3E-09), *actin filament organization* (3.2E-04) and *muscle thin filament tropomyosin* (5.7E-04). There was one term in common in the top 10 significantly enriched GO terms, *actin filament binding*. Tissue-specific splicing in the two sets of genes seemed to be related to cytoskeleton organization of the specific tissues. See S2 Table for more details.

Comparison to tissue-specific splicing at the transcript level

We were able to map sufficient RNAseq reads to both sides of the splicing event for 248 of the 255 events. According to the criteria we used in our analysis (a difference of at least 1 standard deviation in expression levels between alternative events) there was evidence for group-specific differences in expression for a total of 159 of the 248 events (62.9%). This concurs with what has already been found by numerous groups; approximately two thirds of alternative splicing events are strongly tissue-specific at the transcript level [1].

The tissue group with the most evidence for group-specific expression for the events in the ASE255 set was digestive tissues (see S4 Fig), but nervous tissues, reproductive tissues, fat and muscle tissues also had high levels of group-specific expression at the transcript level. There was little evidence of tissue specificity at the transcript level among the 248 transcript level events we analysed in either liver or endocrine tissues.

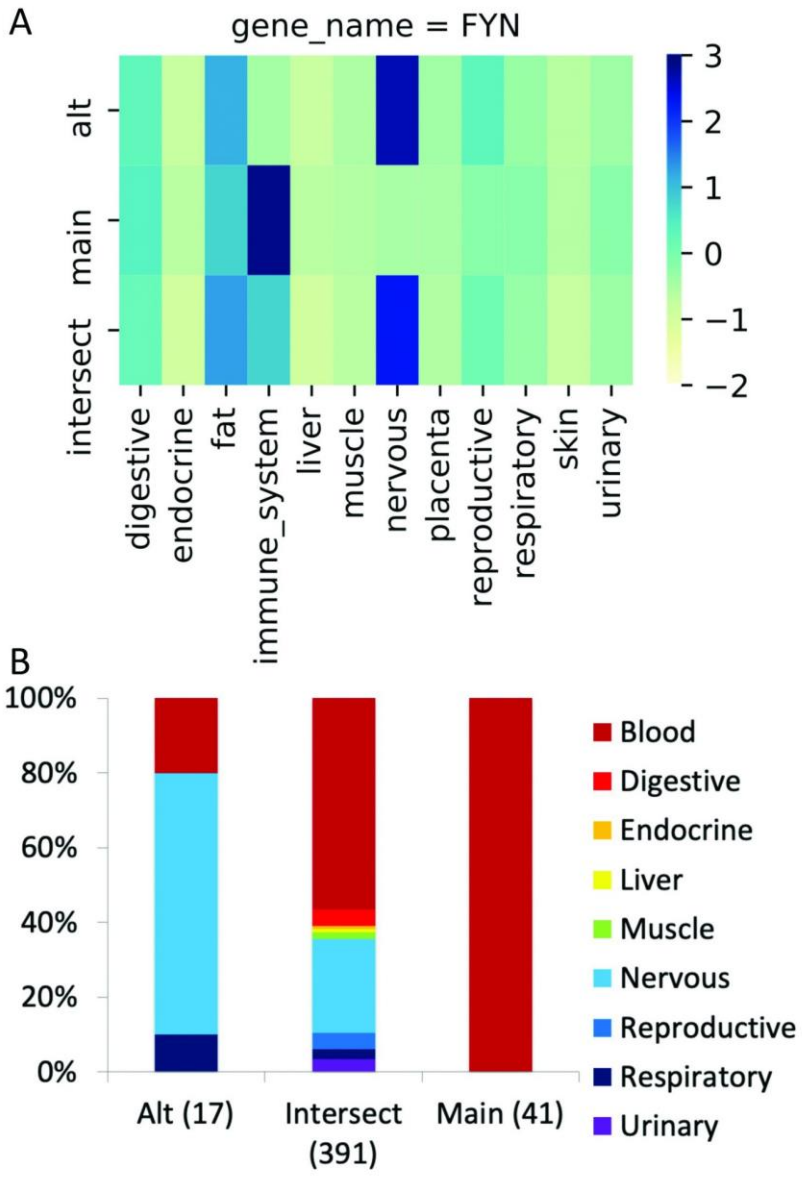


Fig 4. The group-specific splicing event in *FYN*. (A) Group-specific distribution of reads that support each side of the *FYN* splice junction (main, alternative) and those that support the common protein sequence (intersect) coloured by standard deviation from the mean; the darker the colour, the greater the positive standard deviation. The main transcript has more than one standard deviation of reads than the alternative for immune system tissues and the alternative transcript has more than one standard deviation of reads than the main transcript in nervous tissues. (B) The distribution of the PEDs from the proteomics experiments for the alternative and main isoforms ("Alt" and "Main") and those that map to the remainder of the common amino acid sequence ("Intersect"). The numbers of PEDs that belong to each group are shown in brackets. Fisher tests show that the alternative side of the event is significantly enriched in peptides from nervous tissues, just as in the RNAseq experiment. The main side of the event is significantly enriched in peptides from blood cells. Although both sides of the event are enriched in different tissue groups, *FYN* does not count twice towards the total of 99 cases in the PGE99 set because the main side of the event is enriched in blood cells.

https://doi.org/10.1371/journal.pcbi.1008287.g004

Coincidence of tissue-specific splicing at protein and transcript level

In order to make a direct comparison between the proteomics and transcriptomics data set, we had to generate a set of paired events from the splice events in the ASE255 set. For the comparison we required that the event was enriched in any of the tissue groups apart from blood. The tissue groups that we included had to be present in both proteomics and RNAseq analysis and the hematopoietic cells analysed in the Kim *et al* experiments [21] did not have a comparable tissue in the Uhlen *et al* analysis [31].

For the comparison we only considered those events in which one side of the event (main or alternative) was significantly enriched at the protein level in at least one of the grouped tissues. We left out events that were only significant at the tissue level and events in which the peptide evidence for the event was depleted rather than enriched. If an event was significantly enriched in more than one tissue group, we counted each tissue in which it was enriched as a distinct case. In total there were 99 cases of proteomics group-specific enrichment in which either the main or alternative side of a splice event was significantly enriched in a tissue group. An example is shown in Fig 4. The 99 cases came from 76 distinct events and are referred to here as the PGE99 set.

Reassuringly, we found that two thirds (66) of the protein level enrichments were also enriched in the same tissue group at the transcript level. The proportion of events significantly enriched at both the protein and transcript level differed substantially between tissue groups (Fig 5A). Many of the events enriched in muscle and nervous tissues at the transcript level were also enriched at the protein level; 32 of 78 splice events enriched at the transcript level in nervous tissues and 21 of 48 events enriched in muscle tissues were significantly enriched at the protein level. However, the same was not true of the other tissue groups. Only 7 of the 85 events enriched in digestive tissues and just 3 of 71 events enriched in reproductive tissues in the transcriptomics experiments were significantly enriched in the same tissues in proteomics experiments. There was significant enrichment for one event in both protein and transcript analyses in liver (*ACOX1*), one significantly enriched in respiratory tissues (*NEBL*) and one significantly enriched in urinary tissues (*TPM4*).

The higher proportions of events enriched at both transcript and protein-level in muscle and nervous tissues were statistically significant. Fisher's exact tests showed significant differences between nervous and placenta (p-value = 0.0005), nervous and digestive (p-value < 0.00001), nervous and reproductive (p-value < 0.00001), and even nervous and respiratory (p-value = 0.0495) and nervous and urinary tissues (p-value = 0.0272). The comparisons between muscle tissues and digestive, placenta, reproductive, respiratory and urinary tissues were similarly statistically significant.

In all but three of the 66 cases of significant enrichment in the same tissue at both the protein and transcript level the event could be traced back to a common ancestor with fish. Tissue specificity at the transcript level also seemed to be associated with conservation of splice events (Fig 5B). Events that were tissue-specific in the transcriptomics analysis were older than events without significant tissue-specific splicing (Fig 5B).The proportion of ancient (> 400 million years old) splice events that were tissue enriched at the transcript level was also significantly greater than the proportion of ancient splice events that were not tissue specific (Fisher exact test < 0.00001).

Event age and tissue groups

To analyse why there was considerably more coincidence between protein level and transcript level tissue specific splicing in nervous and muscle tissues than in reproductive and digestive tissues we defined the sides of each significantly enriched transcript level event as either enriched (the side of the event with significantly more transcript evidence) or depleted.

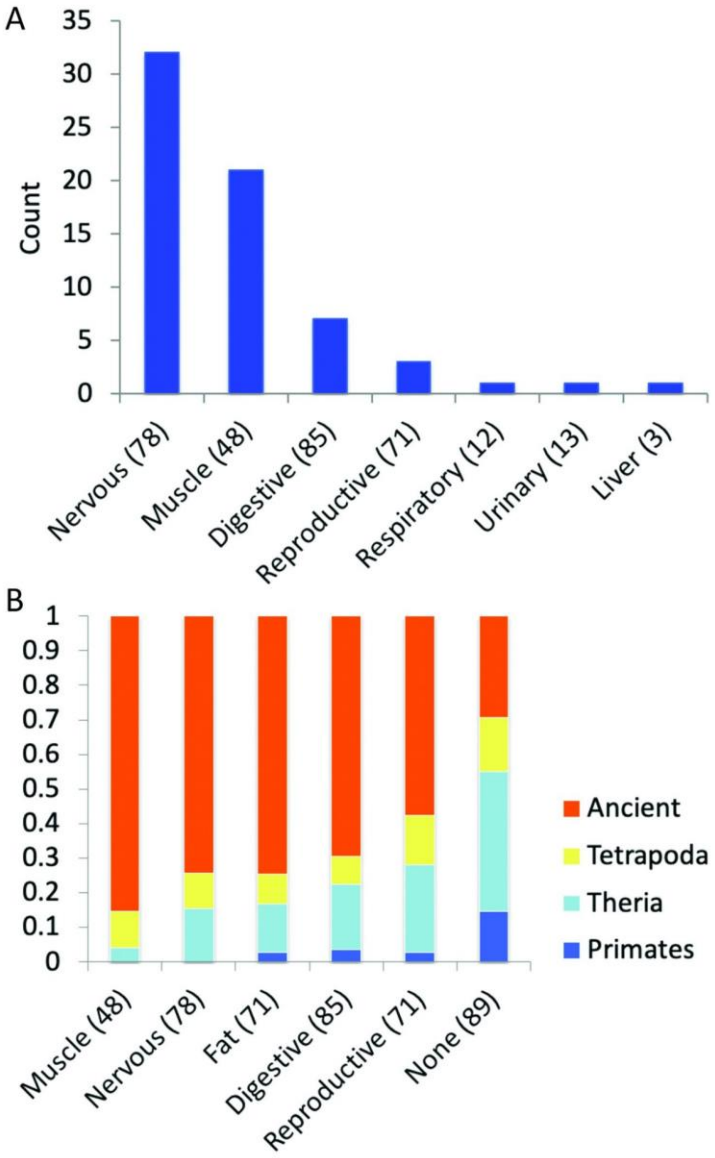


Fig 5. Comparison of RNA-level tissue-specific events. (A) The number of transcript-level enriched events also enriched at the protein level. Each bar shows the number of transcript-level tissue group enriched events that are enriched in the same tissue group in proteomics experiments. Enriched events were compared over the 9 tissue groups that coincided in both transcriptomics and proteomics experiments. The number of events that were tissue-specific in the transcriptomics experiments for each group is shown in brackets. (B) The age of the events enriched in RNAseq studies in the five most populated tissue groups and those not enriched at all (None). At the RNAseq level more of the muscle and nervous tissue enriched events are ancient than those in any other tissue. Results shown for tissues with a minimum of 48 tissue-specific enriched events.

https://doi.org/10.1371/journal.pcbi.1008287.g005

With the two sides of each event defined we were able to sum the PEDs that supported each side of digestive, muscle, nervous and reproductive tissue specific events (S5 Fig). We found that there were significantly more PEDs for the transcript-enriched side of events than the depleted side in all four tissues (Fisher's exact tests: digestive 0.00001, muscle 0.0, nervous 0.0, reproductive 0.0007),

We calculated the percentages of supporting PEDs for the enriched and depleted sides of each individual event and generated scatter plots for each of the tissues (Fig 6). Most events had proportionally more supporting PEDs for the enriched side of the splice event than the depleted side in all four tissue groups. Where muscle and nervous tissues differed was that the PEDs that support the enriched side of the event were often highly enriched. Many of the points in muscle and nervous tissues fall a long way from the diagonal representing equal proportions of supporting PEDs (Fig 6) and the enriched side is supported by 100% of the PEDs in many events that are enriched in these tissues. By contrast none of the enriched sides of

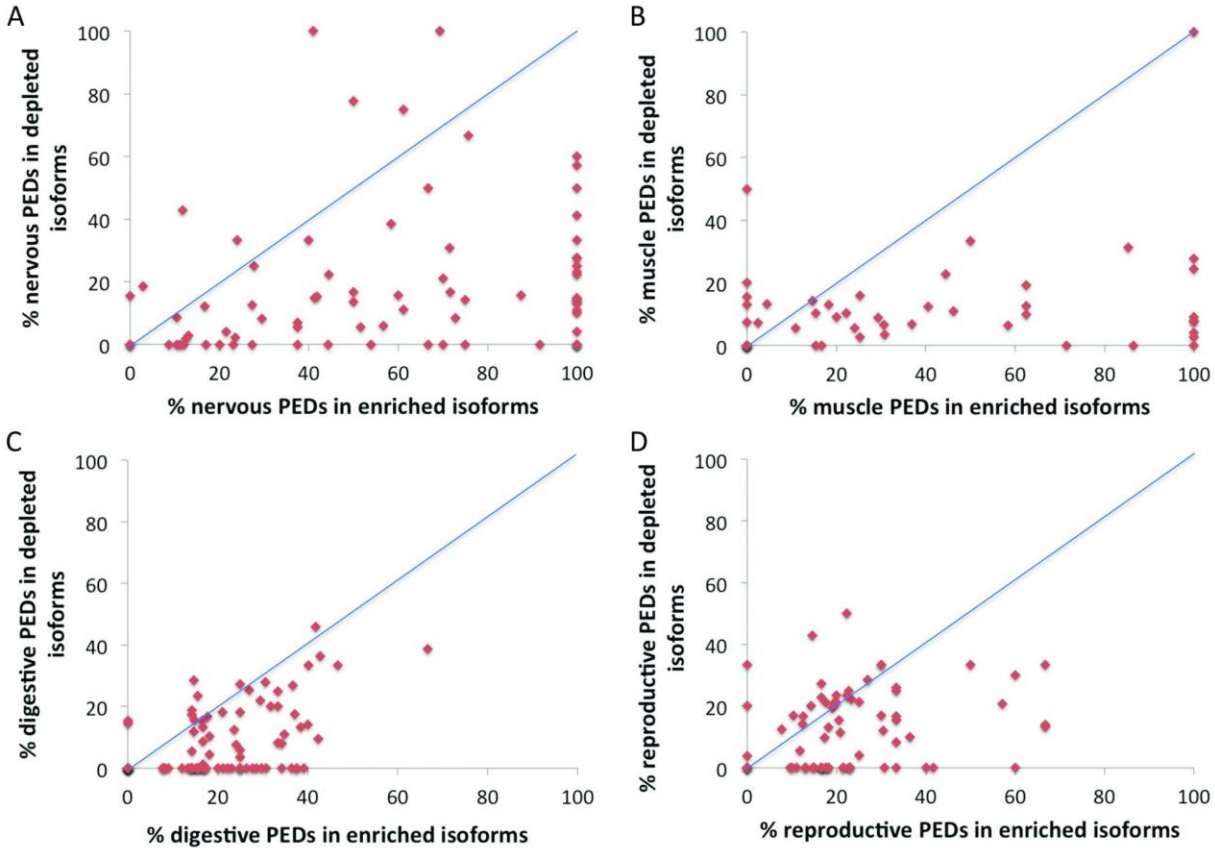


Fig 6. Scatter plot of the percentages of PEDs supporting transcript level enrichment. The figure shows scatter plots of the percentage of PEDs that support the selected tissue for the side of the event that is enriched in that tissue group in the transcriptomics experiments (X-axis) versus the percentage of PEDs that support the selected tissue for the side of the event depleted in the transcriptomics experiments (Y-axis). The diagonal line shows where the percentage of PEDs that support the side of the event that is enriched in the transcriptomics experiments is identical to the percentage of PEDs that support the side of the event that is depleted in that tissue group in the transcriptomics experiments. Events that have proportionally more PEDs on the transcript-enriched side of the event (those that agree with the transcriptomics evidence) ought to be below the line. Tissues shown are A. Nervous B. Muscle C. Digestive and D. Reproductive.

https://doi.org/10.1371/journal.pcbi.1008287.g006

events in reproductive or digestive tissues is supported by more than 70% of total PEDs. The proteomics evidence suggests that many transcript-enriched events in digestive and reproductive tissues may also be enriched at the proteomics level, but the enrichment is often minimal, as seen by the clustering around the diagonal in these two tissues (Fig 6).

The data also suggests there may be a higher proportion of noisy splicing at the transcript level in reproductive tissues than in the other tested tissues, although it is difficult to draw firm conclusions from tissues which include testis. In more than a third (35.6%) of events that are reproductive-enriched at the transcript level there is as much or more evidence for the depleted side of the event at the protein level as there is for the enriched side. The 31 events that were tissue-specific at the transcript level in reproductive tissues that evolved most recently (since the split with fish) are not as a whole significantly enriched at the protein level (S6 Fig).

Is there correlation between proteomics and transcriptomic data at the event level?

Since we had already calculated the percentage of PEDs that supported both sides of tissue specific splice events, we also calculated the percentage of RNAseq reads that supported splice events that were tissue specific in digestive, muscle, nervous and reproductive tissues. We determined the correlation between the percentage of PEDs and the percentage of RNAseq reads that supported each side of a splice event. Here there were also substantial differences between digestive, muscle, nervous and reproductive tissues here (Fig 7). The correlation between supporting PEDs and supporting reads was highest in nervous (0.799) and muscle (0.748) tissues and lowest in reproductive tissues (0.413). Plots of supporting reads against supporting PEDs are available for the four tissues (S5 Fig).

We used the whole of the ASE255 set to determine the correlation for each tissue by age of splice event (Fig 7). When comparing supporting PEDs and reads over all splice events the correlation will in part be due to gene expression rather than alternative splicing because we are including events that are not significantly tissue specific. This explains much of the high correlation among theria-derived events in reproductive tissues, for example. Despite this, it is clear that the correlation between proteomics and transcriptomics support is considerably worse for those splice events that arose in the primate clade. Correlation coefficients for primate-derived events (which make up more than three quarters of annotated alternative exons in the human gene set) ranged from 0.003 (in muscle) to 0.319 (digestive tissues).

Caveats

There are a number of caveats to the comparison of proteomics and tissue level alternative splicing. Firstly, the analysis was carried out on a small number of alternative splicing events. This was inevitable because even large-scale mass spectrometry-based proteomics experiments detect few alternative isoforms reliably [15]. Secondly, even though we analysed those splice events with the most proteomics support, the relatively low numbers of discriminating peptides for each event limits statistical power and makes it harder to identify tissue specificity. A deeper exploration at the protein level is likely to show that there are tissue-specific differences for some events that we are not detecting.

Thirdly, the comparison between the proteomics and transcriptomics experiments was handicapped by the fact that the experiments were not paired (experiments did not come from the same individuals). Finally, we were only able to interrogate 9 tissue groups, and it is likely that other groups may also display further tissue-specific alternative splicing. For example, there was also substantial evidence for tissue-specific alternative splicing at the protein level in

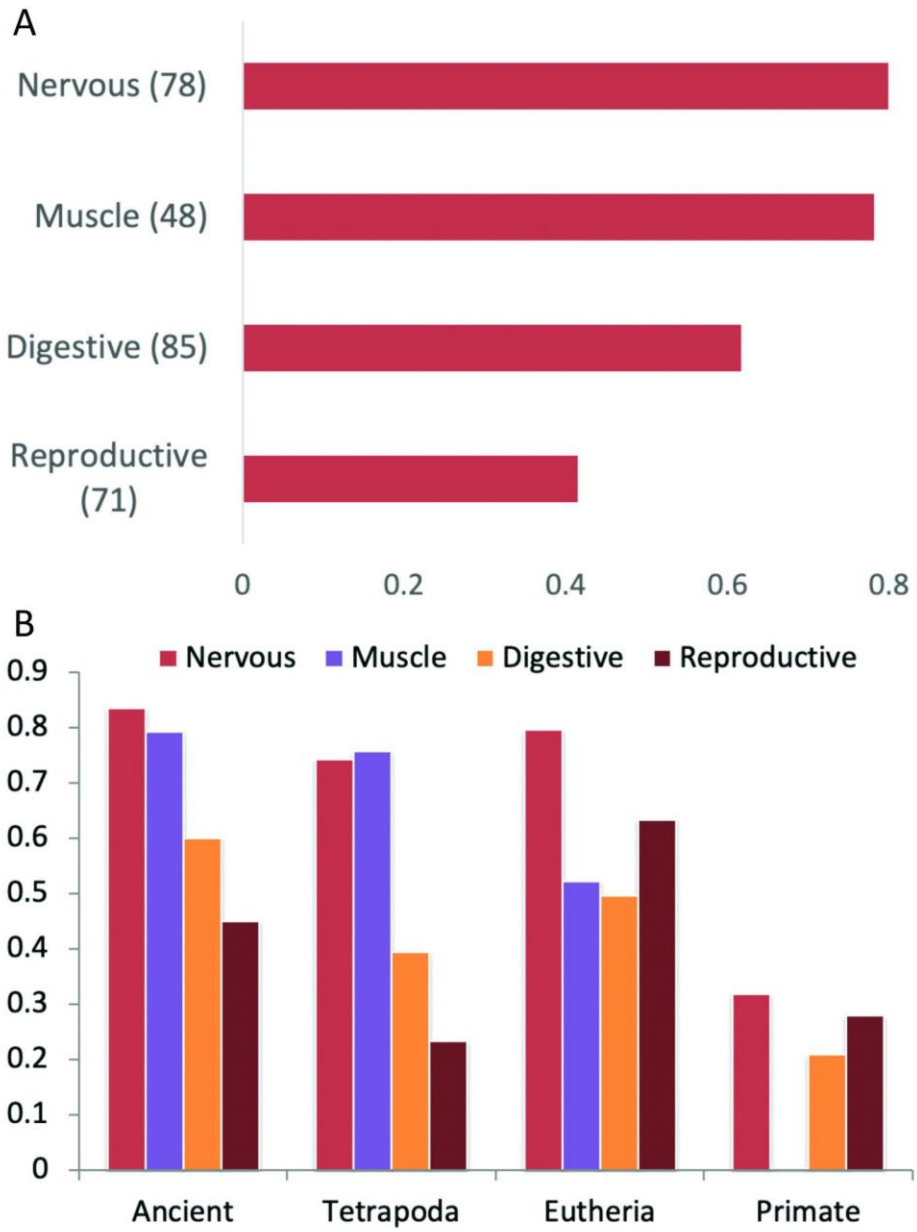


Fig 7. Correlation between PED and RNAseq read support. (A) Correlation between percentage PED and read support for those splice events enriched in grouped digestive, muscle, nervous and reproductive tissues at the transcript level. (B) Correlation between percentage PED and read support for splice events grouped by event age.

https://doi.org/10.1371/journal.pcbi.1008287.g007

blood and at the transcript level in fat. Within these 9 tissue groups it was also harder to detect tissue-specific splicing at the protein level in tissue groups with fewer replicate experiments (placenta, endocrine and lung tissues).

Given the nature of proteomics experiments and the minimum requirement for three PEDs, we cannot demonstrate that any splice event is *not* tissue or tissue group specific, or even cell type specific. The lack of coverage and/or the different make up of tissues does play a role in the differences between the results at the transcript and protein level. For example, events in *ABI2*, and *ATP1B4* were enriched in brain tissue in the transcriptomics analysis and we identified multiple discriminating PSM for isoforms of in frontal cortex, but in both cases the differences were not significant at the protein level due to the lack of peptide coverage. Events in six genes, *FMN1*, *RAP1GAP*, *NECTIN1*, *IDE3B*, *FRMD5* and *ATP2B3*, that had apparent tissue specific differences at the protein level (two had PEDs in frontal cortex, one was apparently enriched in retina, one in fetal heart, one in adult heart, and one in pancreas) were left out of the analysis because one side of each event only had 2 PEDs. All eight of these events could be traced back to a common ancestor with fish.

Conclusions

Transcript level studies consistently show that the majority of alternatively spliced exons are tissue specific. In this analysis we also find substantial tissue-specific alternative splicing also exists at the protein level. Just over a third of the 255 splice events validated in our proteomics analysis are significantly tissue specific.

Manual curation of the protein level tissue-specific splice events detected in our analysis found that almost three quarters had homologues in fish. No tissue specific splice event was primate-derived. This is in sharp contrast to the alternative exons in the human gene set, more than three quarters of which arose in the primate clade. Reyes *et al* found similar differences in tissue specific splice patterns: while a minority of conserved exons had large amplitude tissue-specific differences, exons with little variations in tissue specific usage were not conserved between species [8].

The stark differences in conservation between tissue specific splice events with evidence at the protein level and alternative exons in the human gene set mean that our results cannot be extrapolated to the whole genome. The lack of detectable tissue-specific splicing among recently evolved splice events suggests that primate-derived splice events are likely to have different tissue-specific behaviour and many may have low amplitude tissue differences, if they have any at all. The weak link between protein level tissue-specificity and recent splice events suggests that tissue-specific alternative splicing is unlikely to generate important species-specific differences.

The theory that alternative splicing might be responsible for large-scale tissue-specific protein-protein interaction networks [9,32] is based in part on evidence for tissue specific splicing, and in part on evidence that alternative exons are enriched in predicted disorder. While we find that alternative exons with evidence of translation are more disordered than would be expected, we find contrasting results for tissue specific splicing events. The set of protein level tissue specific splice events actually have proportionally fewer disordered regions than non-tissue specific splice events.

There is some overlap between our data set and the exons used in these two analyses. For example, *BIN1*, illustrated in the Ellis et al study [32], is part of our ASE255 set. However, our set is highly enriched in exons that evolved during or prior to the vertebrate clade and more recently evolved splice events are significantly enriched in predicted disordered regions (S2 Fig). Recently evolved splice events have significantly more disordered regions than those that evolved more than 400 million years ago (Fisher exact test value < 0.00001). Although tissue specific alternative splicing is likely to affect protein-protein interactions, our study suggests that the role of disorder may not be as important as has been suggested.

Most protein level tissue enrichment at the protein level occurred in either muscle or nervous tissues. By way of contrast to other analyses [22,24], which found considerable evidence of tissue-specific splicing in testis at the protein level, we detected little evidence for tissue-specific splicing at the protein level in testis or in grouped reproductive tissues as a whole. Very few events were significantly enriched at the protein level in reproductive tissues and more than a third of the 71 events enriched at the transcript level were actually depleted at the protein level.

Nervous and cardiac tissues have been shown to have an important number of conserved tissue-specific splice events [33,34]. Our protein-level results are in agreement with an analysis of transcript level splicing signatures across multiple vertebrate species [2], which found that brain and heart/muscle tissues had strong conserved splicing signatures, while remaining tissues clustered by species rather than by tissue.

Functional analysis showed that protein level tissue-specific events were significantly enriched in genes annotated with functional terms related to the cytoskeleton. Genes with significant tissue-specific alternative splicing in muscle tissues (principally heart) were related to the composition and function of muscle and the Z-discs in the sarcomere, while genes with significant tissue-specific alternative splicing in nervous tissues were related to cytoskeletal connections and cell-cell contacts.

The importance of tissue-specific alternative splicing in two specialised tissues like brain and heart, the clear evidence of deep conservation, and the functional terms that are associated with the cytoskeleton and cellular differentiation paints a picture in which tissue-specific alternative splicing has been decisive in the development of nervous and muscle tissues. Our results are supported by previous data that document that tissue-specific splicing plays an important role in the development of brain and heart tissues [35–37].

In this study we have identified many functional alternative isoforms along with the tissues in which they are most expressed. The challenge is to determine exact functional roles for those isoforms where none is known. The gene *NEBL*, for example, has two main isoforms that differ in their N-terminals, the longer is called nebulette and the shorter LASP2. We find that nebulette is expressed exclusively in cardiac tissues, while LASP2 is found most often in nervous and urinary tissues and not in muscle tissues. Although the role of nebulette in binding Z-disc associated desmin filaments in cardiac tissues has been known for several years [38], LASP2 has only recently been shown to play a crucial role in post-synaptic development in the brain [39]. In order to further the investigation into the roles of these undoubtedly important alternative isoforms, we have listed many of the tissue specific alternative isoforms analysed in this study on the APPRIS web site [20].

Material and methods

Human reference genome

This study was based on the annotations in v27 of the GENCODE human reference gene set. The manual annotations in GENCODE v27 [3] are equivalent to Ensembl 90 and were produced in June 2017. The GENCODE v27 gene set had 19,881 protein coding genes.

Proteomics analysis

We reanalysed the data from the Kim *et al* [21] proteomics experiments. The data comprised spectra from high-resolution Fourier-transform mass spectrometry experiments of 30 histologically normal human samples, including 17 adult tissues, 7 foetal tissues and 6 purified primary haematopoietic cells. In total there were 79 usable experiments, 18 covering fetal tissues and 61 covering adult tissues and haematopoietic cells. All tissues had at least two replicate

experiments, though the number of replicates varied. Adult heart had five replicates, for example.

Spectra from each experiment were downloaded from ProteomeXchange [40] and were searched against the GENCODE v27 human reference proteome, a decoy database [41] and a list of common contaminants, using the COMET search engine [42]. COMET allowed fixed post-translational modifications of methionine. The peptide spectrum matches (PSMs) from COMET were post-processed with Percolator [43]. We were more interested in reducing false positives than in increasing coverage, so we selected those PSMs that had a posterior error probability (PEP) lower than 0.001. PEP values of less than 0.001 in our analysis equated to PSM q-values of less than 0.0001. In addition, peptides were also limited to those that were fully tryptic, had no more than a single missed cleavage and had a length between 7 and 40 residues. Peptides that mapped to more than one gene were also discarded. With these rules in place we identified at least 2 PSM for 11,065 coding genes in the GENCODE v27 reference set.

Although we searched for tissue specificity using the 30 distinct tissues in the Kim *et al* analysis, much of the analysis was based on pooling the 30 tissues and hematopoietic cells into 10 groups of related tissues. This was done to amplify any signal. The proteomics tissue groups are detailed in **S3 Table**.

Alternative splicing analysis

We analysed the tissue specificity of splice events rather than the tissue specificity of entire transcripts and splice isoforms because RNAseq reads and peptides are too short to cover more than short regions of sequence. Transcript reconstruction methods can be used to predict alternative transcript levels, but these methods are inaccurate [44] and there is no equivalent method for proteomics data.

In order to analyse splice events it is necessary to introduce the idea that splice events have two sides. Discriminating peptides and RNAseq reads will map to one side or the other of a splice event. For example, in the case of an indel one side of the event will be an insertion and the other a deletion, while there will be two different amino acid sequences at the protein level in the case of substitutions. We distinguished each side of the event as the main (the side of the event with most protein evidence) or alternative.

Analysis of the proteomics data allowed us to detect the presence or absence of peptides in distinct tissue-based experiments. Given the format of the experiments we were analysing (label-free experiments, replicates for all tissues) we chose to count the number of experiments in which splice event distinguishing peptides were detected. Each peptide was associated with a peptide-experiment detections (PEDs) count, which represented the number of experiments in which a peptide was detected. A peptide that was identified in every single experiment would therefore have 79 PEDs; peptides identified in a single experiment would be associated with just one PED. For the analysis we required that each side of a splicing event (main and alternative) was supported by a minimum of three PEDs (**Fig 8A**). This threshold was applied because it is not possible to detect the significance of tissue specificity for events supported by fewer than three PEDs.

Protein level tissue specificity calculations

To carry out tissue-specific analysis at the protein level we annotated one side of each splice event as belonging to the main isoform, the side of the event with the most supporting PEDs, while the other side of the event was determined to belong to an alternative isoform. For each event, PEDs that supported each gene were separated into three types (**Fig 8B**), those that supported the side of the splice event with most evidence (that would give rise to the main

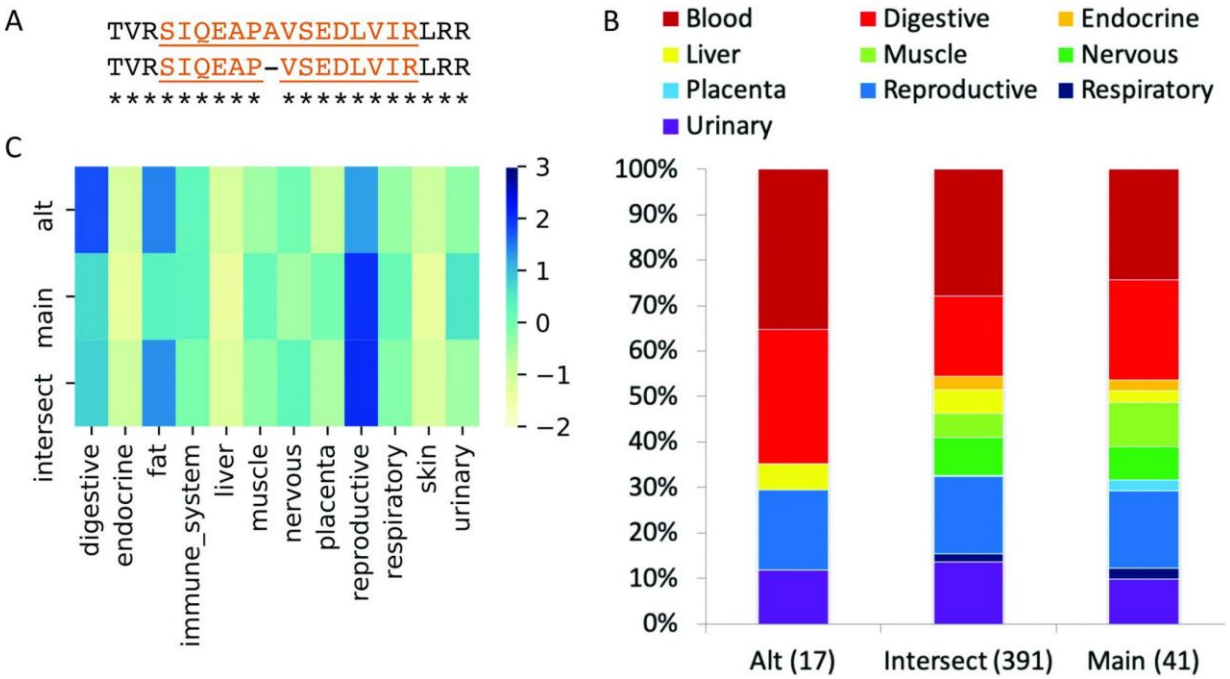


Fig 8. Group-specific splicing in gene *TOR1AIP1*. (A) We found peptide evidence for a single primate-derived splice event for *TOR1AIP1*. This NAGNAG splicing event resulted in the loss/gain of a single amino acid. The discriminating peptides we detected are highlighted. (B) The distribution of the PEDs for the discriminating peptides ("Alt" and "Main") and the remaining peptides that mapped to TOR1AIP1, but that did not distinguish one isoform from the other ("Intersect"). The number of PEDs in each set is shown in brackets. Fisher tests show that the distributions of PEDs between the Main and Alt peptides are not significantly different over any of the ten tissue groups. (C) The group-specific distribution of reads that support each side of the splice junction (main, alt) and those that support remaining common protein sequence (intersect) coloured by standard deviation from the mean; the darker the colour, the greater the positive standard deviation. There is more than one standard deviation between the reads for the digestive and reproductive groups, so the *TOR1AIP1* event is determined to be group specific at the transcript level for these tissue groups.

https://doi.org/10.1371/journal.pcbi.1008287.g008

isoform), those that supported the other side of the splice event (the alternative isoform) and those that did not discriminate between the main isoform and the alternative isoform (the intersect).

We used the PEDs to calculate three sets of contingency tables for Fisher's exact tests, always comparing one tissue or group against the rest of tissues or groups. Fisher's exact tests were carried out for all tissues between main isoform and alternative isoform, main isoform and intersect, and alternative isoform and intersect.

Transcript expression data

We downloaded data from the large-scale RNAseq analysis carried out by the Human Protein Atlas [31]. The RNAseq analysis was performed on 36 different tissues. It covers similar tissues to the Kim *et al.* analysis, though this analysis did not investigate fetal tissues or blood cells (S3 Table). We aligned the RNAseq data to GENCODE v27 using STAR 2.6 [45], forcing end-to-end read alignments to avoid unwanted alignments to repetitive regions. The maximum number of multiple alignments allowed was 50 and the rest of parameters were set by default.

We grouped tissues from the Human Protein Atlas analysis into 12 groups, where possible using just those tissues analysed in the proteomics experiments. Transcriptomics analysis

tissue groupings are shown in S3 Table. Tissues not interrogated in the proteomics experiments (such as skeletal muscle and duodenum) were left out of the groupings. There were three groups that did not appear in the proteomics analysis (skin, fat and immune system) and one proteomics analysis group that did not have an equivalent in the transcriptomics analysis (blood).

For the 255 splice events in the protein level alternative splicing set, we summed the reads into three groups in the same way that we did for the peptides, those reads that distinguished either the alternative or the main side of the splicing event and those that did not distinguish either side of the event. We calculated the mean number of reads across all the tissue groups for each gene and used the mean to calculate standard deviations for each set of reads that mapped to each peptide (Fig 1C). Events were counted as tissue group specific when the reads that mapped to one side of the splice event (equivalent to the main protein isoform or the alternative splice isoform) were at least one standard deviation higher than the other side of the splice event. Heat maps for the splice events in the ASE255 set are shown in S7 Fig.

Alternative exon age

We calculated the age of the splice events in the ASE255 set manually by searching for supporting evidence in the UniProtKB database. We carried out BLAST [46] searches against vertebrate sequences with the residues that made up each side of the event and manually noted the presence or absence of the required sequence. If the event was shorter than 20 amino acid residues, we added flanking amino acids so that the search sequence was at least 20 amino acids long. We complemented BLAST searches with multiple alignments of vertebrate sequences.

To analyse the age of events in the genome as a whole, we calculated cross-species conservation scores for the alternative exons in the human reference set. Alternative exons were defined at the genome level using the APPRIS database [20]. APPRIS selects a representative protein isoform as the principal isoform for every coding gene. APPRIS determines principal isoforms based on protein structural and functional information and a score representing cross-species conservation and we have demonstrated that a single main isoform is the reality for the majority of coding genes and that APPRIS is the best predictor of this main isoform [7]. Alternative isoforms were all isoforms that were not tagged as principal. Alternative exons were those that did not overlap at all with exons that produced principal isoforms.

Ideally we would also calculate exon age manually, but this is not feasible at the genome level. Instead we calculated exon age from the cross-species conservation of the amino acid sequence corresponding to each exon. Cross-species conservation was calculated from BLAST searches against a protein database. We limited our analysis to alternative exons with a minimum of 42 bases to reduce the error rate.

Searches were carried out in two ways. Firstly, we searched for similarity to the translated exon itself, and secondly, we searched for similarity to the translation of the exon joined to the neighbouring exon (in the case 3' and 5' exon substitutions), or the exon plus both flanking exons (in the case of inserted or substituted exons). For searches with both sets of exons we recorded the species of those homologous sequences that had fewer than four residue insertions. The most distant homologue in each search was taken to represent the predicted age of the exon. The final exon age was the minimum of the predicted ages in the two analyses (the single exon and multiple exon calculations).

Disorder predictions

We downloaded the IUPred2 disorder predictor [47] to make predictions for disorder for the splice isoforms in the ASE255 set. We calculated long disorder for all regions that differed

between the main and alternative isoforms. For indels we calculated disorder for the insertion, for substitutions we calculated disorder for both regions involved in the swap and took the region with the highest proportion of disorder as the representative score for that event. Events that were four or fewer amino acids in length were left out of the analysis. IUPred defines a disordered residue as having a score of 0.5. We defined a region as disordered if more than half of the amino acid residues scored more than 0.5.

GO term calculations

We used DAVID [48] to calculate the significantly enriched GO terms within the genes we detect alternative splicing for, within those genes that had tissue-specific alternative splicing in nervous tissue in both proteomics and transcriptomics experiments, and within those genes that had tissue-specific alternative splicing in muscle tissue in both proteomics and transcriptomics experiments. As a background we used the 10,485 genes that we detected in the Kim *et al* experiments that had at least two distinct non-overlapping peptides. This was to remove in-built biases of the proteomics experiments and to limit to those genes for which it was minimally possible to detect two distinct splice isoforms.

Supporting information

S1 Fig. Significant tissue specific alternative splicing cases in proteomics tissues. The count of the number of times we recorded tissue specific differences at the protein level in each of the 30 tissues.
(PDF)

S2 Fig. Predicted order and disorder for alternative exons. Mean order and disorder predicted by IUPred for various subsets. *Protein TS* are those events that are tissue specific at the protein level. *Transcript TS* are those events that are tissue specific at the transcript level. *Cassette TS* are skipped exon events that are tissue-specific at the protein level. *Protein Not* are those events that are not tissue specific at the protein level. *Transcript Not* are those events that are not tissue specific at the transcript level. *Cassette TS* are skipped exon events that are not tissue specific at the protein level. *Ancient* are those events that manual curation has shown to evolve more than 400 million years ago. *Recent* are all other events.
(PDF)

S3 Fig. The relative ages of splice events in cytoskeleton-related genes. The number of events with evidence in four different clades (vertebra to primates) separated into four groups by whether or not they were present in cytoskeleton-related genes ("Cytoskeleton" and "Other genes"), and whether or not the event was found to be significantly tissue specific at the protein level ("TS" or "Not"). There was a significantly higher proportion of vertebrate-derived events among the tissue specific events in cytoskeleton-related genes (Fisher's exact tests: 0.0093 vs Other genes TS, less than 0.00001 for the other two non-tissue specific groups).
(PDF)

S4 Fig. The number of events that were tissue-specific in each of the 12 transcriptomics tissue groups.
(PDF)

S5 Fig. Correlation between supporting PEDs and supporting reads. For each enriched/depleted event in the corresponding tissue the chart shows the percentage of reads support one side of the event that are detected in the corresponding tissue, plotted against the percentage of all PEDs for the same side of the event detected in proteomics experiments for that tissue.

Results are shown just for those events that are enriched/depleted in transcriptomics experiments in (A) digestive, (B) muscle, (C) nervous and (D) reproductive tissues.
(PDF)

S6 Fig. Percentage of PEDs supporting the transcript level enrichment. The figure shows the percentage of supporting PEDs for the four tested tissue groups (digestive, muscle, nervous and reproductive) from events are enriched (or depleted) in these groups in transcriptomics experiments. The percentage of supporting PEDs among all PEDs detected are shown for the sides of the events that are enriched in transcriptomics experiments (dark red) and for the sides of the events depleted in transcriptomics experiments (light blue). The percentage of PEDs are shown over all events enriched in transcriptomics experiments (*All*), over the subsets of events enriched in transcriptomics experiments that evolved after the split from fish (*Tetrapoda*) and over those that evolved after the split from monotremes (*Theria*). The number of events enriched in transcriptomics experiments and in each subset is shown in the x-axis. Asterisks above the bars show where the number of PEDs supporting the enriched side of the events were significantly different from the number of PEDs on the depleted sides of the events as would be expected if the events were group specific as a whole.
(PDF)

S7 Fig. Heatmaps with the standard deviation for the AS events. The darker the colour, the greater the standard deviation. We calculated the mean number of reads across all the tissue groups for each gene (*intersect*) and used the mean to calculate standard deviations for each set of reads that mapped to each event. Events were counted as tissue group specific when the reads that mapped to the *main* or alternative (*alt*) side of the splice event were at least one standard deviation higher than the other side of the splice event.
(PDF)

S1 Table. The ASE255 set.
(XLSX)

S2 Table. GO terms for tissue specific events.
(XLSX)

S3 Table. List of groups of tissues in proteomics and RNA-seq experiments. Human Body Map tissue proteomics experiments collected in tissue groups (tab 1) and Human Protein Atlas tissue transcriptomics experiments collected in tissue groups (tab 2).
(XLSX)

Acknowledgments

The authors would like to thank Federico Abascal for his invaluable input on this paper.

Author Contributions

Conceptualization: Michael L. Tress.

Data curation: Jose Manuel Rodriguez, Fernando Pozo, Tomas di Domenico.

Formal analysis: Jose Manuel Rodriguez, Michael L. Tress.

Funding acquisition: Jesus Vazquez, Michael L. Tress.

Investigation: Jose Manuel Rodriguez, Michael L. Tress.

Methodology: Jose Manuel Rodriguez, Jesus Vazquez, Michael L. Tress.

Project administration: Michael L. Tress.

Resources: Fernando Pozo, Tomas di Domenico.

Software: Jose Manuel Rodriguez.

Supervision: Jesus Vazquez, Michael L. Tress.

Validation: Jose Manuel Rodriguez, Michael L. Tress.

Visualization: Jose Manuel Rodriguez, Michael L. Tress.

Writing – original draft: Jose Manuel Rodriguez, Michael L. Tress.

Writing – review & editing: Jose Manuel Rodriguez, Fernando Pozo, Jesus Vazquez, Michael L. Tress.

References

1. Wang E, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. https://doi.org/10.1038/nature07509 PMID: 18978772
2. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012; 338:1593–1599. https://doi.org/10.1126/science.1228186 PMID: 23258891
3. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019; 47:D766–D773. https://doi.org/10.1093/nar/gky955 PMID: 30357393
4. Hu Z, Scott HS, Qin G, Zheng G, Chu X, Xie L, et al. Revealing missing human protein isoforms based on ab initio prediction, RNA-seq and proteomics. Sci Rep. 2015; 5 10940. https://doi.org/10.1038/srep10940 PMID: 26156868
5. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018; 19:208. https://doi.org/10.1186/s13059-018-1590-2 PMID: 30486838
6. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell. 2016; 164:805–817. https://doi.org/10.1016/j.cell.2016.01.029 PMID: 26871637
7. 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. https://doi.org/10.1186/gb-2013-14-7-r70 PMID: 23815980
8. Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. Drift and conservation of differential exon usage across tissues in primate species. Proc Natl Acad Sci U S A. 2013; 110:15377–15382. https://doi.org/10.1073/pnas.1307202110 PMID: 24003148
9. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell. 2012; 46:871–883. https://doi.org/10.1016/j.molcel.2012.05.039 PMID: 22749400
10. Ghadie MA, Lambourne L, Vidal M, Xia Y. Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. PLoS Comput Biol. 2017; 13:e1005717. https://doi.org/10.1371/journal.pcbi.1005717 PMID: 28846689
11. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015; 348:660–665. https://doi.org/10.1126/science.aaa0355 PMID: 25954002
12. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res. 2018; 46:582–592. https://doi.org/10.1093/nar/gkx1165 PMID: 29202200
13. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. Function of alternative splicing. Gene. 2013; 514:1–30. https://doi.org/10.1016/j.gene.2012.07.083 PMID: 22909801
14. Bhuiyan SA, Ly S, Phan M, Huntington B, Hogan E, Liu CC, Liu J, Pavlidis P. Systematic evaluation of isoform function in literature reports of alternative splicing. BMC Genomics. 2018; 19:637. https://doi.org/10.1186/s12864-018-5013-2 PMID: 30153812

15. Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res. 2015; 14:1880–1887. https://doi.org/10.1021/pr501286b PMID: 25732134

16. Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, et al. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. Mol Biol Evol. 2012; 29:2265–2283. https://doi.org/10.1093/molbev/mss100 PMID: 22446687

17. Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. The ribosome-engaged landscape of alternative splicing. Nat Struct Mol Biol. 2016; 23:1117–1123. https://doi.org/10.1038/nsmb.3317 PMID: 27820807

18. Wang SH, Hsiao CJ, Khan Z, Pritchard JK. Post-translational buffering leads to convergent protein expression levels between primates. Genome Biol. 2018; 19:83. https://doi.org/10.1186/s13059-018-1451-z PMID: 29950183

19. Inada T. The ribosome as a platform for mRNA and nascent polypeptide quality control. Trends Biochem Sci., 2017; 42:5–15. https://doi.org/10.1016/j.tibs.2016.09.005 PMID: 27746049

20. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res. 2018; 46:D213–D217. https://doi.org/10.1093/nar/gkx997 PMID: 29069475

21. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014; 509:575–581. https://doi.org/10.1038/nature13302 PMID: 24870542

22. Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, et al. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat Commun. 2016; 7:11778. https://doi.org/10.1038/ncomms11778 PMID: 27250503

23. Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, Vázquez J, et al. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS Comput Biol. 2015; 11:e1004325. https://doi.org/10.1371/journal.pcbi.1004325 PMID: 26061177

24. Lau E, Han Y, Williams DR, Thomas CT, Shrestha R, Wu JC, Lam MPY. Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. Cell Rep. 2019; 29:3751–3765. https://doi.org/10.1016/j.celrep.2019.11.026 PMID: 31825849

25. Bradley RK, Merkin J, Lambert NJ, Burge CB. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol. 2012; 10:e1001229. https://doi.org/10.1371/journal.pbio.1001229 PMID: 22235189

26. Kondrashov FA, Koonin EV. Origin of alternative splicing by tandem exon duplication. Hum Mol Genet. 2001; 10:2661–2669. https://doi.org/10.1093/hmg/10.23.2661 PMID: 11726553

27. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, et al. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A. 2006; 103:8390–8395. https://doi.org/10.1073/pnas.0507916103 PMID: 16717195

28. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet. 2014; 23:5866–5878. https://doi.org/10.1093/hmg/ddu309 PMID: 24939910

29. Abascal F, Juan D, Jungreis I, Kellis M, Martinez L, Rigau M, et al. Loose ends: almost one in five human genes still have unresolved coding status. Nucleic Acids Res. 2018; 46:7070–7084. https://doi.org/10.1093/nar/gky587 PMID: 29982784

30. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011; 7:548. https://doi.org/10.1038/msb.2011.81 PMID: 22068331

31. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015; 347:1260419. https://doi.org/10.1126/science.1260419 PMID: 25613900

32. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell. 2012; 46:884–892. https://doi.org/10.1016/j.molcel.2012.05.037 PMID: 22749401

33. Kalsotra A, Xiao X, Ward AJ, Castle JC, Johnson JM, Burge CB, Cooper TA. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. Proc Natl Acad Sci U S A. 2008; 105:20333–20338. https://doi.org/10.1073/pnas.0809045105 PMID: 19075228

34. Vuong CK, Black DL, Zheng S. The neurogenetics of alternative splicing. Nat Rev Neurosci. 2016; 17:265–281. https://doi.org/10.1038/nrn.2016.27 PMID: 27094079

35. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet. 2011; 12:715–729. https://doi.org/10.1038/nrg3052 PMID: 21921927

36. Lara-Pezzi E, Gómez-Salinero J, Gatto A, García-Pavía P. The alternative heart: impact of alternative splicing in heart disease. J Cardiovasc Transl Res. 2013; 6:945–995. https://doi.org/10.1007/s12265-013-9482-z PMID: 23775418

37. Jacko M, Weyn-Vanhentenryck SM, Smerdon JW, Yan R, Feng H, Williams DJ, et al. Rbfox Splicing Factors Promote Neuronal Maturation and Axon Initial Segment Assembly. Neuron. 2018; 97:853–868. https://doi.org/10.1016/j.neuron.2018.01.020 PMID: 29398366

38. Hernandez DA, Bennett CM, Dunina-Barkovskaya L, Wedig T, Capetanaki Y, Herrmann H, Conover,G. M. Nebulette is a powerful cytolinker organizing desmin and actin in mouse hearts. Mol Biol Cell. 2016; 27:3869–3882. https://doi.org/10.1091/mbc.E16-04-0237 PMID: 27733623

39. Myers KR, Yu K, Kremerskothen J, Butt E, Zheng JQ. The Nebulin Family LIM and SH3 Proteins Regulate Postsynaptic Development and Function. J Neurosci. 2020; 40:526–541. https://doi.org/10.1523/JNEUROSCI.0334-19.2019 PMID: 31754010

40. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 2017; 45:D1100–D1106. https://doi.org/10.1093/nar/gkw936 PMID: 27924013

41. Wright JC, Choudhary JS. DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. J Proteomics Bioinform. 2016; 9:176–180. https://doi.org/10.4172/jpb.1000404 PMID: 27418748

42. Eng JK, Jahan TA, Hoopmann, MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013; 13:22–24. https://doi.org/10.1002/pmic.201200439 PMID: 23148064

43. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom. 2016; 27:1719–1727. https://doi.org/10.1007/s13361-016-1460-7 PMID: 27572102

44. Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015; 31:3938–3945. https://doi.org/10.1093/bioinformatics/btv488 PMID: 26338770

45. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

47. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018; 46:W329–W337. https://doi.org/10.1093/nar/gky384 PMID: 29860432

48. Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Liu D, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics. 2007; 8:426. https://doi.org/10.1186/1471-2105-8-426 PMID: 17980028

DISCUSSION

APPRIS: principal isoforms for multiple gene sets

Many experiments and large-scale analyses require a single representative for each gene. The standard method for selecting a representative is to choose the longest isoform, but the longest isoform is not always the main isoform. APPRIS automatically selects a principal isoform for coding genes based on the available biological information. APPRIS deploys a range of computational methods to annotate alternative isoforms with protein 3D structure information, functionally important residues, Pfam domains, signal peptides and transmembrane helices, and a score for the cross-species conservation of each transcript model. These high-quality annotations are used to select the principal isoform.

The motivation behind APPRIS is the observation that most alternative isoforms either lack regions of conserved structure or function, or have exons that are evolving at measurably different rates compared with their principal counterparts (Tress *et al.*, 2008). APPRIS selects as a principal isoform the isoform with the most conserved protein features and most evidence of cross-species conservation, while those isoforms with unusual, missing or non-conserved protein features are flagged as alternative.

Results from our group and others (Ezkurdia *et al.*, 2012, 2015; Sheynkman *et al.*, 2013; Tress *et al.*, 2017) suggest that many genes have a single, clearly definable dominant protein isoform and that the alternative isoforms are either expressed less frequently, in limited tissues or in unique developmental stages, or have a much shorter half-life. The dominant protein isoform is almost always the APPRIS principal isoform: APPRIS principal isoforms overwhelmingly coincide with the manually annotated unique CCDS variants and with the main isoforms detected in large-scale proteomics experiments (Ezkurdia *et al.*, 2015). Further corroboration of the importance of APPRIS principal isoforms comes from large-scale genetic variation studies, which show that exons from principal isoforms are under purifying selection, while alternative exons appear to be evolving neutrally (Liu & Lin, 2015; Tress *et al.*, 2017).

The principal isoform is the most representative isoform for each coding gene. However, not all APPRIS principal isoforms are alike. Principal isoforms are tagged with a score from 1 to 5 depending on the reliability of the selection, with 1 being the most reliable and 5 being the method of last resort, selecting the longest remaining candidate isoform. APPRIS determines a most reliable isoform for 75%-95% of protein-coding genes annotated depending on the gene set and the species. In the case of human, the current version of the APPRIS database determines a principal isoform without resorting to sequence length in 99% of protein-coding genes, compared to the previous version that identified a principal isoform for 85% of the human protein-coding genes.

The reliability of each APPRIS module is continually revised using the Ensembl/GENCODE human reference gene set. We determined that the gold standard set for principal isoforms are those genes with just one CCDS variant because the agreement between the main experimental isoforms and unique CCDS variants was 98.6% across those genes that had a single CCDS isoform (Ezkurdia *et al.*, 2015; Tress *et al.*, 2017). Since the first published version of APPRIS, there has been a steady increase in the agreement between the unique CCDS variants and the APPRIS principal isoforms (and of course the results from the individual methods).

One example that displays the recent improvements in the APPRIS methods and the principal isoform decisions is the *ASIC4* gene. Acid-sensing ion channel 4 is a cation channel with an

affinity for sodium. This gene has two variants (ASIC4-201 and ASIC4-202). Both isoforms map to many 3D structures (e.g. 3IJ4), but ASIC4-201 has a deletion that by homology to *ASIC1* would remove part of the thumb region, a region crucial to the regulation of the ion channel (see Figure 9). Although the Pfam domain (ASC) is broken in both isoforms, the ASIC4-201 isoform with the deletion would have an extra gap. In previous versions of APPRIS, Matador3D and SPADE disagreed over which isoform was most likely to be the principal isoform because the Pfam domain was broken in both isoforms. The new version of SPADE recognizes the extra break in the ASC domain caused by the deletion in ASIC4-201 and selects ASIC4-202 as the main isoform. Since both Matador3D and SPADE now agree on the isoform that most represents the conserved protein features, ASIC4-202 is now selected as the principal isoform.

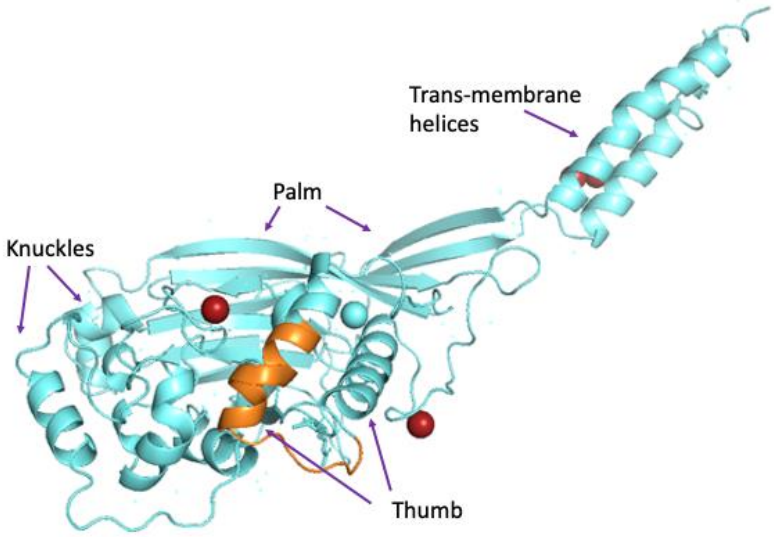


Figure 9. ***The deletion in ASIC4-201 mapped onto the structure of chicken ASIC1.*** *The region deleted is in orange. Caesium cations shown in red. The deletion would lead to the complete refolding of the "thumb" region of the protein, a region that is important for the regulation of the ASIC1 ion channel (Hanukoglu, 2017). The image was generated using Pymol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC) based on PDB structure 3IJ4, chicken ASIC1.*

APPRIS principal isoforms have a wide range of uses, from the determination of principal and alternative isoforms for genes in individual research projects, to the determination of principal and alternative exons for use in genome-wide analysis of variants. Clarifying which splice isoform (or isoforms) is functionally relevant is important for understanding biological systems and the effect of mutations (Abascal *et al.*, 2016). Indeed, we have found that just 0.6% of ClinVar (Landrum *et al.*, 2018) pathogenic mutations supported by publications map to exons defined as alternative by APPRIS. Even then the phenotypic effect of half of these mutations is likely to be a result of interference with principal transcript splice sites rather than an effect on the predicted alternative isoform (unpublished results).

APPRIS is also providing a wealth of data that are being used in ongoing projects to further investigate the role and importance of AS, such as the analysis of tissue-specific AS (Rodriguez *et al.*, 2020), the labelling of potential non-coding genes (Abascal *et al.*, 2018), the prediction of functional alternative isoforms (unpublished results).

Before this publication, the APPRIS database covered five Ensembl species (human, mouse, rat, pig and zebra fish). With the publication, we extended the database to three more species: one vertebrate (chimpanzee) and two invertebrate genomes (*Drosophila* and *Caenorhabditis elegans*). However, APPRIS is continually expanding based on the needs of the scientific community, and now APPRIS has two more vertebrate species: chicken and

cow. The chicken genome was a request from the large-scale Bird 10,000 Genomes Project (Zhang, 2015).

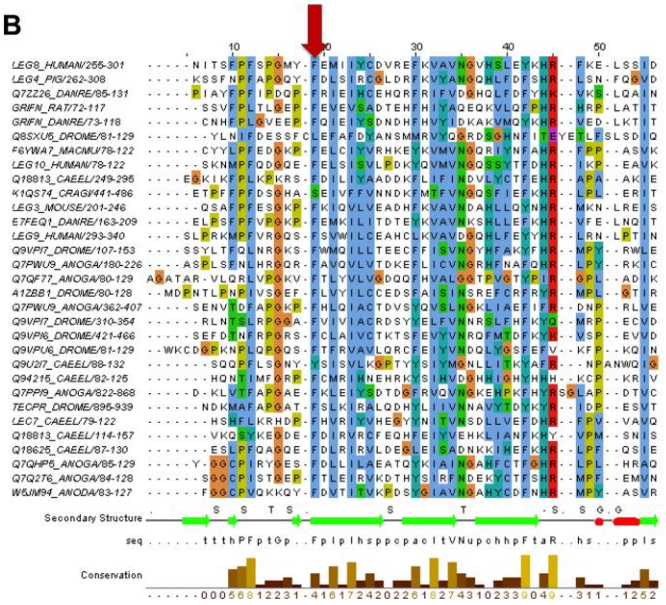
The extension of APPRIS annotations to the RefSeq gene sets and UniProtKB proteomes, a part of Ensembl/GENCODE, is very useful for investigators and genome research. We have also created merged gene sets for vertebrate species by cross-referencing the Ensembl/GENCODE, RefSeq and UniProtKB reference sets. We established a common gene set (Intersection).

The merged gene set, Intersection, allows us to identify isoforms missing in the individual gene sets. This information is fed back to manual annotators to inform gene models. For example, the principal isoform from the Intersection set for the gene *GRIFIN* is annotated in Ensembl/GENCODE (ENST00000614228) and UniProtKB (A4D1Z8), but not in RefSeq (Figure 10). The principal isoform has annotation evidence from cross-species alignments and the C-terminal extension in the Ensembl/GENCODE and UniProtKB (but not in the RefSeq variants) is also established in mammals. In addition, the principal isoform maps better to known 3D structures, and has an unbroken Pfam domain.

A

Ensembl Id.	RefSeq Id.	UniProt Id.	Length (aa)	No. Functional Residues	3D Structure Score	Domain Score	Conservation Score
-	NM_001291784	-	148	8	93.6	72.6	0
-	XM_011515405	-	208	8	93.4	72.6	0
ENST00000614228	-	A4D1Z8	144	8	109.1	86.8	5.8

B



C

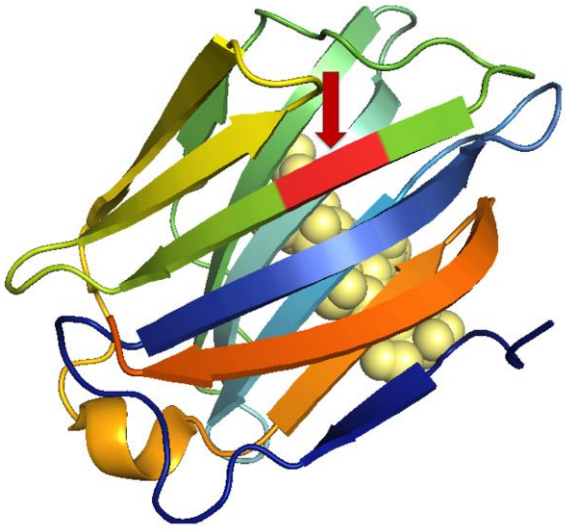


Figure 10. ***APPRIS annotations for gene GRIFIN - figure from (Rodriguez et al., 2018).*** *(A) APPRIS results for the three protein-coding variants composed of Ensembl/GENCODE, RefSeq and UniProtKB. APPRIS identifies the same isoform ENST00000614228 (Ensembl) and A4D1Z8 (UniProtKB) as the principal isoform (highlighted in green). A selection based on the 3D structure, the functional domains and the conservation in related species. (B) Alignment for a section of the Pfam galectin family of proteins. The red arrow shows where 8-extra residues in the RefSeq variants would disrupt a region of the functional domain of GRIFIN. (C) The 3D structure of 4LBJ that has identity with variants ENST00000614228+A4D1Z8. The red arrow shows where the 8-extra residues would break the structure.*

Loose ends: almost one in five human genes still have unresolved coding status

The initial step to merge the three human reference sets Ensembl/GENCODE, RefSeq and UniProtKB was based on a process developed for APPRIS. Afterwards, we carried out an extensive manual selection to integrate the three data sets. The manual curation produced a combined set of 22,210 protein-encoding genes. At the same time, 19,446 genes were annotated in the intersection of the three sets, which meant that one in eight protein-coding genes were classified differently in at least one of the reference sets (Figure 11).

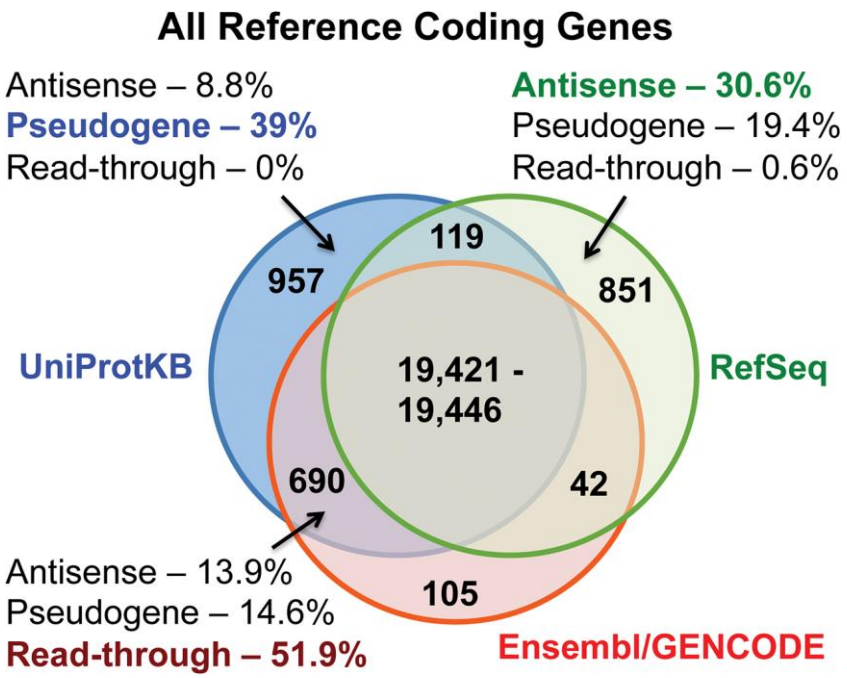


Figure 11. ***The overlap between Ensembl/GENCODE, RefSeq and UniProtKB genes - figure from (Abascal et al., 2018).*** *The diagram shows the number of coding genes for each reference database and the intersection between them. The number of genes in the intersection of the three sets is variable because RefSeq and Ensembl/GENCODE disagree on gene boundaries for a number of genes. The figure also shows the percentage of annotated coding genes classified as antisense and pseudogene in other databases, or known to be based on read-through transcripts for coding genes unique to one reference set.*

Those genes that are not classified as coding in all three reference sets have a range of alternative classifications (Table 1). For example, 51.9% of the genes annotated as coding in Ensembl/GENCODE and UniProtKB but not in RefSeq are read-through genes (Figure 11). Read-through genes, genes that are composed of transcripts that skip the last coding exon to read through to exons from neighboring genes or pseudogenes, are currently classified as coding by the RefSeq and Ensembl/GENCODE annotations even though there is little evidence they encode proteins. Potential "antisense" genes, non-coding genes on the opposite strand of protein-coding loci, account for 30.6% of genes classified as coding in the RefSeq unique subset.

Genes that are classified as pseudogenes in other reference sets make up 39% of the genes that are coding in UniProtKB alone (Figure 11). These genes have homology to known protein-coding genes but contain a frameshift and/or stop codon(s), which disrupt the ORF and most arise through duplication followed by loss of function. These genes are especially difficult to distinguish from coding genes. Distinguishing whether a locus should be a

pseudogene or protein coding gene is often complicated and changing predictions to coding genes involves investigating variation of haplotypes, underlying genome assembly errors and using extremely stringent mapping options to confidentially (Bruford *et al.*, 2015).

Type	Ensembl Single	RefSeq Single	UniProtKB Single	Ensembl - RefSeq	Ensembl - UniProtKB	RefSeq - UniProtKB
Antisense	7	260	84		96	22
Duplicate (technical)	41			2	1	
IG/TR genes	6		120	33	16	
LncRNA		141	126		1	47
Other ncRNA		39	40		23	7
Sense overlapping		44				
Pseudogene	2	165	373		101	39
Read-through	38	5		7	358	1
Retroviral gene			26			
Sense intronic	2	31			16	
Alt genome sequence		6	84			
Not in reference	9	160	104		78	3
Total	**105**	**851**	**957**	**42**	**690**	**119**

Table 1. The annotations of genes not classified as coding in all three sets. *The table shows the classification for those genes classified as coding by just one or two reference sets. Genes annotated, but not in the reference set are tagged as "Alt genome sequence". Genes that are not present in other reference sets are labelled as "Not in reference".*

In the paper we defined a set of 16 potential non-coding features. A total of 2,278 (11%) Ensembl/GENCODE coding genes were tagged with at least one of the potential non-coding features. This included almost all the genes outside of the intersection of the three reference sets. This suggests that many or even most of the "coding" genes outside of the intersection may not code for cellular proteins.

In order to compare the potential non-coding (PNC) genes with genes that are likely to be coding, we analyzed experimental transcriptomics, proteomic and antibody binding data. Few potential non-coding genes had reliable proteomics or antibody evidence and they also had less transcript support. In fact, PNC genes had significantly lower transcript expression and were detected in very few tissues. Since non-coding genes are known to have much lower levels of expression (Derrien *et al.*, 2012), the low or negligible RNA-seq expression levels is further evidence for the suggestion that many PNC genes will not code for proteins.

Genetic variation data is a good indicator of selective pressure. Most coding genes should have very few high impact variants in common alleles and should have non-synonymous to synonymous ratios that are lower for common alleles than they are for rare alleles. We found that PNC genes had a much higher proportion of high impact variants and greater non-synonymous to synonymous ratios than likely coding genes (Figure 12).

The higher proportions of high impact variants among PNC genes and the similarity in non-synonymous to synonymous ratios in both common alleles and rare alleles suggests that many of these genes are not under purifying selection. Since neutral selection is not characteristic of coding genes, this implies the suggestion that many PNC genes are unlikely to code for functional proteins. Those PNC genes annotated by two or fewer reference sets (Subsets) had worse ratios than the PNC from the intersection of the three sets (Intersect).

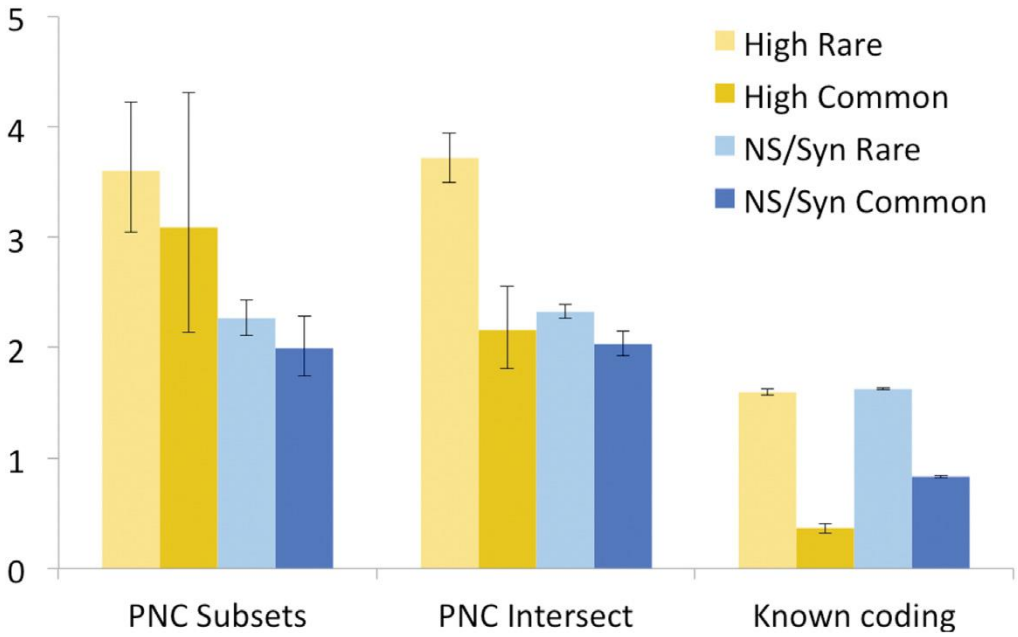


Figure 12. ***Genomic variation in likely coding genes and possible noncoding genes - figure from (Abascal et al., 2018).*** *Percentage high impact variants (yellow) and nonsynonymous/ synonymous ratios (blue) for known coding genes (likely coding genes), for possible non-coding genes in the intersection of the three sets (PNC Intersect) and for those PNC genes classified as coding by just one or two reference sets. Read-through genes were removed when calculating variants. The darker colours show the values for common variants and the lighter shades show the values for rare variants.*

Although we predicted that many of the PNC genes will not code for proteins, the determination of whether a gene is coding or not is complicated and ambiguous. Even what should be unequivocal coding evidence itself may not always be what it seems. Antibody experiments are not specific enough to distinguish similar proteins, and proteomics experiments can easily confuse similar peptides due to single amino acid variations or post-translational modifications. In fact, after the publication of the paper we sent a dozen antibody identifications of PNC to the Human Protein Atlas to discover which were most likely to be real. The Human Protein Atlas told us that none of the identifications should be regarded as high confidence.

To complicate matters further, while positive evidence for coding potential is often hard to validate, support for non-genes does not exist: it is impossible to prove that a gene can never code for a protein. In the end classification as non-coding, pseudogene, artifact or coding is usually decided by manual curators on the balance of all the available evidence.

The distinct methods of curation in RefSeq, UniProt and Ensembl/GENCODE means that there are likely to be many disagreements between the annotators over the genes that are annotated differently in the three reference sets. However, as a result of our paper the three annotation groups are now working more closely together.

If most of the genes not classified as coding across the three reference sets do not code for proteins, the number of coding genes will be much closer to the 19,446 genes common to the sets. However, it is still early to speculate on the exact number of coding genes because it is impossible to know how many new coding genes may appear (Wright, J. C. *et al.*, 2016).

With the publication of the "finished" version of the Human Genome Project (International Human Genome Sequencing Consortium, 2004), the number of coding genes decreased between 20,000 and 25,000. The most recent version of GENCODE (GENCODE v35 08/2020) contains 19,954 genes. Rigorous manual annotation has brought us considerably closer to a final catalog of human coding genes, where the annotators coincide in more than 85% of the coding genes.

An analysis of tissue-specific alternative splicing at the protein level

Studies have shown consistently that most alternatively spliced exons are tissue specific at the transcript level. In this analysis, we also found substantial tissue-specific alternative splicing at protein level. Given the relatively low coverage of proteomics experiments, it should be more difficult to detect tissue specific isoforms, yet we found that just over a third of the 255 splice events validated in our proteomics analysis are significantly tissue specific. Tissue specific alternative protein forms were particularly abundant in nervous and muscle tissues (see Figure 13).

Both nervous and muscle tissues have previously been shown to have an important number of conserved tissue specific splice events (Kalsotra *et al.*, 2008; Vuong *et al.*, 2016). Our protein-level results are in agreement with an analysis of transcript level splicing signatures across multiple vertebrate species (Merkin *et al.*, 2012), which found that brain and heart/muscle tissues had strong conserved splicing signatures, while remaining tissues clustered by species rather than by tissue.

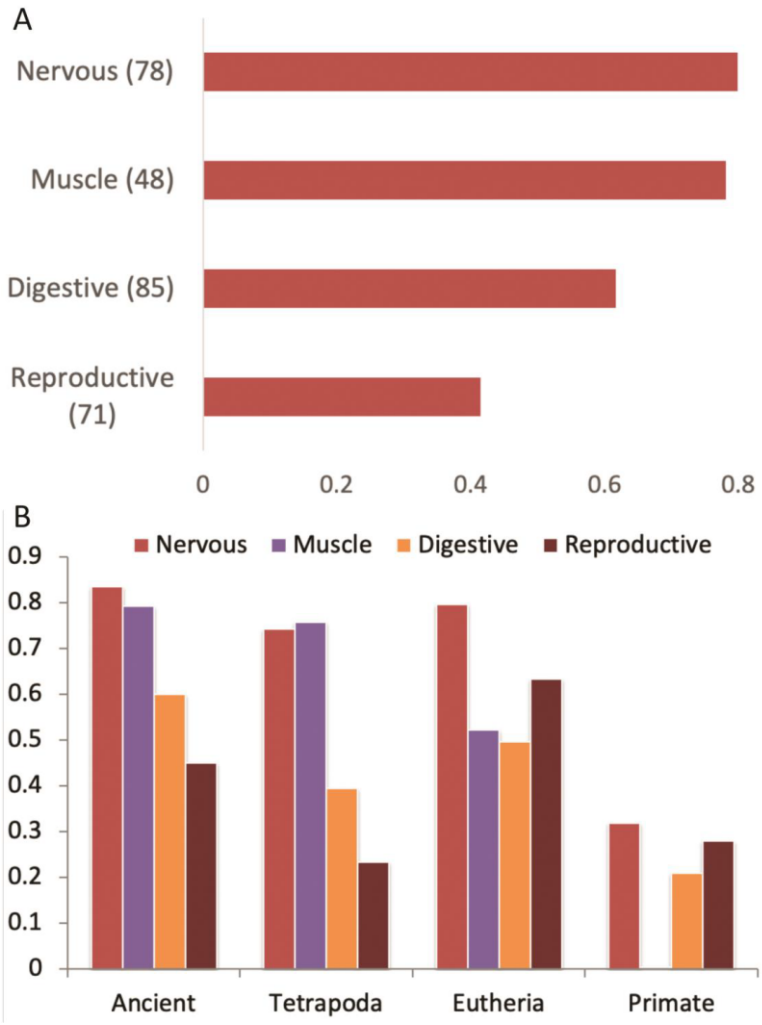


Figure 13. ***Correlation between PED and RNA-seq read support - figure from (Rodriguez et al., 2020).*** *(A) Correlation between percentage PED and read support for those splice events enriched in grouped digestive, muscle, nervous and reproductive tissues at the transcript level. (B) Correlation between percentage PED and read support for splice events grouped by event age.*

We found that although many events enriched in reproductive and digestive tissues at the transcript level were also enriched at the protein level, these differences were almost never statistically significant. In contrast to other analyses (Lau *et al.*, 2019; Wright, J. C. *et al.*, 2016), which found considerable evidence of tissue-specific splicing in testis at the protein level, we found that very few events were significantly enriched at protein level and more than a third of the events enriched at the transcript level were actually depleted at the protein level.

Although we detected substantial evidence of tissue specific alternative splicing at the protein level, there was evidence to suggest that the high number of tissue specific isoforms might be specific to the set of highly expressed splice isoforms in this paper. Alternative splice events detected in proteomics experiments were considerably more conserved than those in the genome as a whole: more than half of the alternative events in the 255 alternative splicing events evolved more than 400 million years ago and only 7.8% of the alternative events in our set derived from the primate clade (Figure 14). Ezkurdia *et al.* previously showed that proteins from ancient gene families are more likely to be detected in proteomics experiments (Ezkurdia *et al.*, 2014) and that there is little reliable proteomics evidence for primate-derived coding genes (Abascal *et al.*, 2018; Ezkurdia *et al.*, 2014). Hence, it is not surprising that we also found most evidence for ancient splice events and little evidence of alternative splicing events derived from the primate clade.

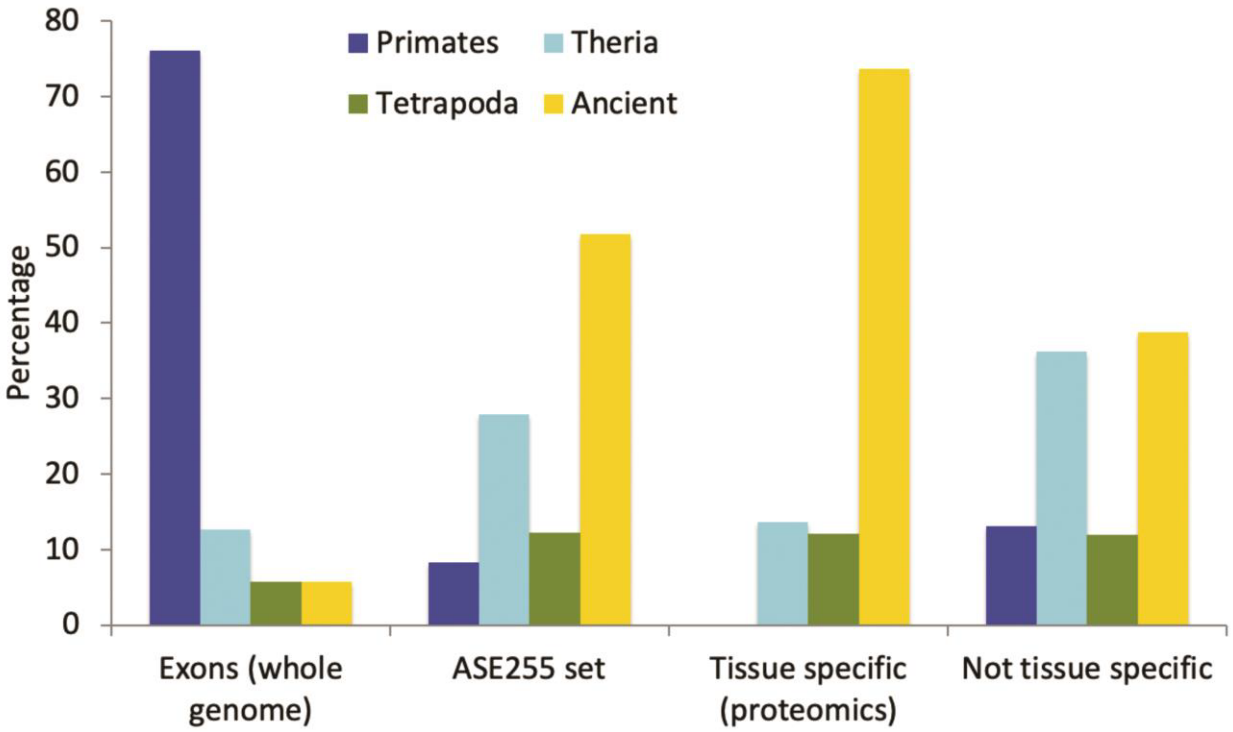


Figure 14. ***The age of alternative exons versus subsets of splicing events detected in proteomics experiments - figure from (Rodriguez et al., 2020).*** *"Exons (whole genome)" – all alternative exons in the human genome, "ASE255 set" – initial data set with the 255 alternative splicing events detected in the proteomics analysis, "Tissue-specific (proteomics)" - events that have significant tissue or group-specific differences at the protein level and "Not tissue specific" – events without tissue-specific enrichment in proteomics experiments.*

However, not only was the set of alternative events detected at the protein level enriched in ancient events, but tissue-specific splice events were even more conserved. Almost three quarters of events with evidence of tissue specificity at the proteomics level evolved more than 400 million years ago. No tissue specific splice event was primate-derived. This is in sharp contrast to the alternative exons in the human gene set, more than three quarters of which arose in the primate clade. Reyes *et al* found similar differences in tissue specific splice patterns: while a minority of conserved exons had large amplitude tissue-specific differences, exons with little variations in tissue specific usage were not conserved between species (A. Reyes *et al.*, 2013).

The stark differences in conservation between tissue specific splice events with evidence at the protein level and alternative exons in the human gene set mean that our results cannot be extrapolated to the whole genome. The lack of detectable tissue-specific splicing among recently evolved splice events suggests that primate-derived splice events are likely to have different tissue-specific behaviour and many may have low amplitude tissue differences, if they have any at all. This weak link between protein level tissue-specificity and recently evolved splice events makes it unlikely that important species-specific differences arose from tissue-specific alternative splicing.

The theory that alternative splicing might be responsible for large-scale tissue-specific protein-protein interaction networks (Buljan *et al.*, 2012; Ellis *et al.*, 2012) is based in part on evidence for tissue specific splicing, and in part on evidence that alternative exons are enriched in predicted disorder (Romero *et al.*, 2006). We analysed the proportion of events with disorder in the ASE255 set and found that alternative exons in the set of splice events

were enriched in disorder. However, there was no indication that disorder was related to tissue specificity either at the protein level or transcript level.

The functional analysis showed again that alternative splicing at the protein level is highly enriched in terms related to the cytoskeleton (Ezkurdia *et al.*, 2012). Among genes with nervous and muscle tissue specific alternative splicing, there were clear differences between the functional terms enriched in these two subsets of genes. Genes with significant tissue specific alternative splicing in muscle tissues (principally heart) were related to the composition and function of muscle and the Z-discs in the sarcomere, while genes with significant tissue specific alternative splicing in nervous tissues were related to cytoskeletal connections and cell-cell contacts.

The importance of tissue specific alternative splicing in two specialised tissues like brain and heart, the clear evidence of deep conservation, and the functional terms that are associated with the cytoskeleton and cellular differentiation paints a picture in which tissue specific alternative splicing has been important in the development of nervous and muscle tissues. Our results are supported by previous data that document that tissue-specific splicing plays an important role in the development of brain and heart tissues (Jacko *et al.*, 2018; Kalsotra & Cooper, 2011; Lara-Pezzi *et al.*, 2013).

The challenge now is to determine exact functional roles for those isoforms where none is known. The gene *NEBL*, for example, has two main isoforms that differ in the N-terminal; the longer is called nebulette and the shorter *LASP2*. We find that nebulette is expressed exclusively in cardiac tissues, while *LASP2* is found most often in nervous and urinary tissues and not in muscle tissues. Although the role of nebulette in binding Z-disc associated desmin filaments in cardiac tissues has been known for several years (Hernandez *et al.*, 2016), *LASP2* has only recently been shown to play a crucial role in post-synaptic development in the brain (Myers *et al.*, 2020). In order to further the investigation into the roles of these undoubtedly important alternative isoforms, we have listed many of the tissue specific alternative isoforms analysed in this study on the APPRIS web site (Rodriguez *et al.*, 2018).

CONCLUSIONS

1. The three human reference gene sets currently overestimate the number of human coding genes, complicating and adding noise to genome research and large-scale biomedical experiments. We find that one in eight of these genes are classified differently in at least one of the reference sets. The set of human coding genes needs to be as complete and consistent as possible for basic research and large-scale projects.

2. The designation of a single representative protein reflects in most cases the biological reality of the cell and this also seems to be true regardless of cell type. This dominant protein isoform is almost always the APPRIS principal isoform, which highlights the importance of extending APPRIS principal isoforms to all model species and to RefSeq and UniProtKB gene sets. The selection of a principal isoform is a critical first step for any genome-wide analysis.

3. Although the data supporting alternative splicing is limited at the protein level, we found that over 95% of splice events that were tissue-specific in both proteomics and RNA-seq analyses evolved at least 400 million years ago.

4. Tissue-specific alternative protein isoforms in the proteomics analysis were abundant in nervous and muscle tissues and their genes had functions related to either the structure of muscle fibres or cell-cell connections. Our results suggest that tissue specific alternative splicing may have played a crucial role in the development of the brain and the heart in vertebrates.

CONCLUSIONES

1. Actualmente, las tres bases de datos de referencia de genes humanos sobreestiman el número de genes codificantes, lo que complica y agrega ruido a la investigación del genoma y experimentos biomédicos a gran escala. Encontramos que uno de cada ocho de estos genes se clasifica de manera diferente en al menos una de las bases de datos de referencia. El conjunto de genes codificantes debe ser lo más completo y congruente posible para la investigación básica y para los proyectos a gran escala.

2. La designación de una única proteína representativa refleja, en la mayoría de los casos, la realidad biológica de la célula; y esto también parece ser cierto independientemente del tipo de célula. Esta isoforma de proteína dominante es casi siempre la isoforma principal de APPRIS, lo que resalta la importancia de extender las isoformas principales de APPRIS a todas las especies modelo y a los conjuntos de genes de RefSeq y UniProtKB. La selección de una isoforma principal es un primer paso fundamental para cualquier análisis genómico.

3. Aunque los datos que respaldan el empalme alternativo son limitados a nivel de proteína, encontramos que más del 95% de los eventos de empalme que eran específicos de tejido tanto en proteómica como en análisis de RNA-seq, evolucionaron hace al menos 400 millones de años.

4. Las isoformas alternativas de proteínas específicas de tejido en el análisis proteómico eran abundantes en los tejidos nerviosos y musculares, y sus genes tenían funciones relacionadas con la estructura de las fibras musculares o con las conexiones célula-célula. Nuestros resultados sugieren que el empalme alternativo específico de tejido puede haber jugado un papel crucial en el desarrollo del cerebro y el corazón en los vertebrados.

REFERENCES

Abascal, F., Corvelo, A., Cruz, F., Villanueva-Cañas, J. L., Vlasova, A., Marcet-Houben, M., Martínez-Cruz, B., Cheng, J. Y., Prieto, P., Quesada, V., Quilez, J., Li, G., García, F., Rubio-Camarillo, M., Frias, L., Ribeca, P., Capella-Gutiérrez, S., Rodríguez, J. M., Câmara, F., … Godoy, J. a. (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1090-1

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M., del Pozo, A., Vázquez, J., Valencia, A., & Tress, M. L. (2015). Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLoS Computational Biology*, *11*(6), 1–29. https://doi.org/10.1371/journal.pcbi.1004325

Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J. M., Vazquez, J., & Tress, M. L. (2018). Loose ends: Almost one in five human genes still have unresolved coding status. *Nucleic Acids Research*, *46*(14), 7070–7084. https://doi.org/10.1093/nar/gky587

Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., Hurles, M., Korbel, J. O., Lam, H. Y. K., Lee, C., & Gerstein, M. B. (2015). Erratum: Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. In *Nature Communications* (Vol. 6). Nature Publishing Group. https://doi.org/10.1038/ncomms9389

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. In *Nature*. https://doi.org/10.1038/nature01511

Alekseyenko, A. v., Kim, N., & Lee, C. J. (2007). Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*. https://doi.org/10.1261/rna.325107

Alt, F. W., Bothwell, A. L. M., Knapp, M., Siden, E., Mather, E., Koshland, M., & Baltimore, D. (1980). Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3′ ends. *Cell*. https://doi.org/10.1016/0092-8674(80)90615-7

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. In *Nucleic Acids Research*. https://doi.org/10.1093/nar/25.17.3389

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Antequera, F., & Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.90.24.11995

Ast, G. (2004). How did alternative splicing evolve? *Nature Reviews Genetics*, *5*(10), 773–782. https://doi.org/10.1038/nrg1451

Auton, A., & Salcedo, T. (2015). The 1000 genomes project. In *Assessing Rare Variation in Complex Traits: Design and Analysis of Genetic Studies*. https://doi.org/10.1007/978-1-4939-2824-8_6

Bahar, T., Ben, S., & Terry, G. (2011). Distribution of alternatively spliced transcript isoforms within human and mouse transcriptomes. *J Omics Res*.

Benjamin Lewin. (1990). Genes IV. *Oxford University Press*. https://doi.org/10.1016/0968-0004(90)90239-8

Birzele, F., Küffner, R., Meier, F., Oefinger, F., Potthast, C., & Zimmer, R. (2008). ProSAS: A database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Research*, *36*(SUPPL. 1), 63–68. https://doi.org/10.1093/nar/gkm793

Black, D. L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry*, *72*(1), 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720

Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., & Hubbard, T. (2011). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and “resurrected” pseudogenes in the mouse genome. *Genome Research*. https://doi.org/10.1101/gr.114272.110

Brosius, J. (2009). The fragmented gene. *Annals of the New York Academy of Sciences*, *1178*, 186–193. https://doi.org/10.1111/j.1749-6632.2009.05004.x

Bruford, E. A., Lane, L., & Harrow, J. (2015). Devising a Consensus Framework for Validation of Novel Human Coding Loci. *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.5b00688

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, *46*(6), 871–883. https://doi.org/10.1016/j.molcel.2012.05.039

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg775

Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., & Briggs, S. P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0811066106

Chang, K. Y., Ryan Georgianna, D., Heber, S., Payne, G. A., & Muddiman, D. C. (2010). Detection of alternative splice variants at the proteome level in aspergillus flavus. *Journal of Proteome Research*. https://doi.org/10.1021/pr900602d

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2777

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., & Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human

genome. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49). https://doi.org/10.1073/pnas.0709013104

Colak, R., Kim, T. H., Michaut, M., Sun, M., Irimia, M., Bellay, J., Myers, C. L., Blencowe, B. J., & Kim, P. M. (2013). Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003030

Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S., Bye, J. M., Beare, D. M., & Dunham, I. (2003). Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Research*. https://doi.org/10.1101/gr.695703

Crick, F. (1970). Central dogma of molecular biology. *Nature*. https://doi.org/10.1038/227561a0

Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quétier, F., Saurin, W., & Weissenbach, J. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics*. https://doi.org/10.1038/76118

Das, M., Burge, C. B., Park, E., Colinas, J., & Pelletier, J. (2001). Assessment of the total number of human transcription units. *Genomics*. https://doi.org/10.1006/geno.2001.6620

de La Fuente, L., Arzalluz-Luque, Á., Tardáguila, M., del Risco, H., Martí, C., Tarazona, S., Salguero, P., Scott, R., Lerma, A., Alastrue-Agudo, A., Bonilla, P., Newman, J. R. B., Kosugi, S., McIntyre, L. M., Moreno-Manzano, V., & Conesa, A. (2020). TappAS: A comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biology*, *21*(1), 119. https://doi.org/10.1186/s13059-020-02028-w

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*. https://doi.org/10.1101/gr.132159.111

Desiere, F. (2006). The PeptideAtlas project. *Nucleic Acids Research*, *34*(90001), D655–D658. https://doi.org/10.1093/nar/gkj040

Deutsch, E. W., Sun, Z., Campbell, D., Kusebauch, U., Chu, C. S., Mendoza, L., Shteynberg, D., Omenn, G. S., & Moritz, R. L. (2015). State of the human proteome in 2014/2015 As viewed through peptideatlas: Enhancing accuracy and coverage through the atlas prophet. *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.5b00500

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*. https://doi.org/10.1038/nature11233

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., & Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell*. https://doi.org/10.1016/0092-8674(80)90617-0

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky995

Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T. H., Calarco, J. A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P. M., Wrana, J. L., Blencowe, B. J., Çolak, R., Irimia, M., Kim, T. H., Calarco, J. A., Wang, X., Pan, Q., O'Hanlon, D., … Blencowe, B. J. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell*, *46*(6), 884–892. https://doi.org/10.1016/j.molcel.2012.05.037

Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.2000.3903

The ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K. B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

The ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*.

Ewing, B., & Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*. https://doi.org/10.1038/76115

Ezkurdia, I., del Pozo, A., Frankish, A., Rodriguez, J. M., Harrow, J., Ashman, K., Valencia, A., & Tress, M. L. (2012). Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Molecular Biology and Evolution*, *29*(9), 2265–2283. https://doi.org/10.1093/molbev/mss100

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., & Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddu309

Ezkurdia, I., Rodriguez, J. M., Carrillo-de Santa Pau, E., Vázquez, J., Valencia, A., & Tress, M. L. (2015). Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform. *Journal of Proteome Research*. https://doi.org/10.1021/pr501286b

Ezkurdia, I., Valencia, A., & Tress, M. (2014). *Analyzing the First Drafts of the Human Proteome*.

Farrell, C. M., O'Leary, N. A., Harte, R. A., Loveland, J. E., Wilming, L. G., Wallin, C., Diekhans, M., Barrell, D., Searle, S. M. J., Aken, B., Hiatt, S. M., Frankish, A., Suner, M. M., Rajput, B., Steward, C. A., Brown, G. R., Bennett, R., Murphy, M., Wu, W., … Pruitt, K. D. (2014). Current status and new features of the Consensus Coding Sequence database. In *Nucleic Acids Research* (Vol. 42, Issue D1). https://doi.org/10.1093/nar/gkt1059

Fields, C., Adams, M. D., White, O., & Venter, J. C. (1994). How many genes in the human genome? *Nature Genetics*. https://doi.org/10.1038/ng0794-345

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky955

Geiger, T., Wehner, A., Schaab, C., Cox, J., & Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular and Cellular Proteomics*. https://doi.org/10.1074/mcp.M111.014050

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., & Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. In *Genome Research*. https://doi.org/10.1101/gr.6339607

Ghadie, M. A., Lambourne, L., Vidal, M., & Xia, Y. (2017). Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS Computational Biology*, *13*(8), e1005717. https://doi.org/10.1371/journal.pcbi.1005717

Gilbert, W. (1978). Why genes in pieces? In *Nature*. https://doi.org/10.1038/271501a0

Gingeras, T. R. (2007). Origin of phenotypes: Genes and transcripts. In *Genome Research*. https://doi.org/10.1101/gr.6525007

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., & Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, *14*(7), R70. https://doi.org/10.1186/gb-2013-14-7-r70

Görg, A., Weiss, W., & Dunn, M. J. (2004). Current two-dimensional electrophoresis technology for proteomics. In *Proteomics*. https://doi.org/10.1002/pmic.200401031

Guigó, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V. B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T. R., Harrow, J., Hubbard, T., Lewis, S. E., & Reese, M. G. (2006). EGASP: the human ENCODE Genome Annotation Assessment Project. In *Genome biology*. https://doi.org/10.1186/gb-2006-7-s1-s2

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*. https://doi.org/10.1038/13690

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M., & Kent, W. J. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1095

Handsaker, R. E., van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & Mccarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, *47*(3), 296–303. https://doi.org/10.1038/ng.3200

Hanukoglu, I. (2017). ASIC and ENaC type sodium channels: conformational states and the structures of the ion selectivity filters. In *FEBS Journal* (Vol. 284, Issue 4, pp. 525–545). Blackwell Publishing Ltd. https://doi.org/10.1111/febs.13840

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., Lagarde, J., Gilbert, J. G. R., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S. E., & Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biology*. https://doi.org/10.1186/gb-2006-7-s1-s4

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Harte, R. A., Farrell, C. M., Loveland, J. E., Suner, M. M., Wilming, L., Aken, B., Barrell, D., Frankish, A., Wallin, C., Searle, S., Diekhans, M., Harrow, J., & Pruitt, K. D. (2012). Tracking and coordinating an international curation effort for the CCDS Project. *Database*, *2012*. https://doi.org/10.1093/database/bas008

Hatje, K., Mühlhausen, S., Simm, D., & Kollmar, M. (2019). The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *BioEssays*, *41*(11), 1–14. https://doi.org/10.1002/bies.201900066

Hernandez, D. A., Bennett, C. M., Dunina-Barkovskaya, L., Wedig, T., Capetanaki, Y., Herrmann, H., & Conover, G. M. (2016). Nebulette is a powerful cytolinker organizing desmin and actin in mouse hearts. *Molecular Biology of the Cell*. https://doi.org/10.1091/mbc.E16-04-0237

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. *Database*. https://doi.org/10.1093/database/bav096

Hnilicová, J., & Staněk, D. (2011). Where splicing joins chromatin. *Nucleus*. https://doi.org/10.4161/nucl.2.3.15876

Hoskins, A. A., Friedman, L. J., Gallagher, S. S., Crawford, D. J., Anderson, E. G., Wombacher, R., Ramirez, N., Cornish, V. W., Gelles, J., & Moore, M. J. (2011). Ordered and dynamic assembly of single spliceosomes. *Science*. https://doi.org/10.1126/science.1198830

Hoskins, A. A., & Moore, M. J. (2012). The spliceosome: A flexible, reversible macromolecular machine. In *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2012.02.009

Hui, J. Y. (2009). Regulation of mammalian pre-mRNA splicing. In *Science in China, Series C: Life Sciences*. https://doi.org/10.1007/s11427-009-0037-0

Hu, Z., Scott, H. S., Qin, G., Zheng, G., Chu, X., Xie, L., Adelson, D. L., Oftedal, B. E., Venugopal, P., Babic, M., Hahn, C. N., Zhang, B., Wang, X., Li, N., & Wei, C. (2015). Revealing Missing Human Protein Isoforms Based on Ab Initio Prediction, RNA-seq and Proteomics. *Scientific Reports*. https://doi.org/10.1038/srep10940

International Human Genome Sequencing Consortium. (2004). International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. https://doi.org/nature03001 [pii]\r10.1038/nature03001

Jacko, M., Weyn-Vanhentenryck, S. M., Smerdon, J. W., Yan, R., Feng, H., Williams, D. J., Pai, J., Xu, K., Wichterle, H., & Zhang, C. (2018). Rbfox Splicing Factors Promote Neuronal Maturation and Axon Initial Segment Assembly. *Neuron*. https://doi.org/10.1016/j.neuron.2018.01.020

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl677

Juntawong, P., Girke, T., Bazin, J., & Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1317811111

Jurica, M. S., & Moore, M. J. (2003). Pre-mRNA splicing: Awash in a sea of proteins. In *Molecular Cell*. https://doi.org/10.1016/S1097-2765(03)00270-3

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2004.03.016

Kalsotra, A., & Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3052

Kalsotra, A., Xiao, X., Ward, A. J., Castle, J. C., Johnson, J. M., Burge, C. B., & Cooper, T. A. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0809045105

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene*, *514*(1), 1–30. https://doi.org/10.1016/j.gene.2012.07.083

Kim, E., Goren, A., & Ast, G. (2008). Alternative splicing: Current perspectives. In *BioEssays*. https://doi.org/10.1002/bies.20692

Kim, E., Magen, A., & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl924

Kim, M.-S. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., … Pandey, A. (2014). A draft map of the human proteome. *Nature*, *509*(7502), 575–581. https://doi.org/10.1038/nature13302

Koren, E., Lev-Maor, G., & Ast, G. (2007). The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.0030095

Lahens, N. F., Kavakli, I. H., Zhang, R., Hayer, K., Black, M. B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R. S., Grant, G. R., & Hogenesch, J. B. (2014). IVT-seq reveals

extreme bias in RNA-sequencing. *Genome Biology*, *15*(6), R86. https://doi.org/10.1186/gb-2014-15-6-r86

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*. https://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lara-Pezzi, E., Gómez-Salinero, J., Gatto, A., & García-Pavía, P. (2013). The alternative heart: Impact of alternative splicing in heart disease. *Journal of Cardiovascular Translational Research*, *6*(6), 945–955. https://doi.org/10.1007/s12265-013-9482-z

Lau, E., Han, Y., Williams, D. R., Thomas, C. T., Shrestha, R., Wu, J. C., & Lam, M. P. Y. (2019). Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. *Cell Reports*, *29*(11), 3751-3765.e5. https://doi.org/10.1016/j.celrep.2019.11.026

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., & Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics*. https://doi.org/10.1038/76126

Licatalosi, D. D., & Darnell, R. B. (2010). RNA processing and its regulation: Global insights into biological networks. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2673

Li, H. D., Menon, R., Govindarajoo, B., Panwar, B., Zhang, Y., Omenn, G. S., & Guan, Y. (2015). Functional networks of highest-connected splice isoforms: From the chromosome 17 human proteome project. *Journal of Proteome Research*, *14*(9), 3484–3491. https://doi.org/10.1021/acs.jproteome.5b00494

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., & Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*. https://doi.org/10.1038/10890

Liu, T., & Lin, K. (2015). The distribution pattern of genetic variation in the transcript isoforms of the alternatively spliced protein-coding genes in the human genome. *Molecular BioSystems*, *11*(5), 1378–1388. https://doi.org/10.1039/c5mb00132c

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics*. https://doi.org/10.1038/ng.2653

Lopez, A. J. (1998). ALTERNATIVE SPLICING OF PRE-mRNA: Developmental Consequences and Mechanisms of Regulation. *Annual Review of Genetics*, *32*(1), 279–305. https://doi.org/10.1146/annurev.genet.32.1.279

Lopez, G., Maietta, P., Rodriguez, J. M., Valencia, A., & Tress, M. L. (2011). Firestar--Advances in the Prediction of Functionally Important Residues. *Nucleic Acids Research*, *39*(Web Server issue), W235-41. https://doi.org/10.1093/nar/gkr437

Low, T. Y., vanHeesch, S., vandenToorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., vanBreukelen, B., Mohammed, S., Cuppen, E., Heck, A. J. R., & Guryev, V. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Reports*. https://doi.org/10.1016/j.celrep.2013.10.041

Ly, T., Ahmad, Y., Shlien, A., Soroka, D., Mills, A., Emanuele, M. J., Stratton, M. R., & Lamond, A. I. (2014). A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *ELife*. https://doi.org/10.7554/eLife.01630

Mallick, P., & Kuster, B. (2010). Proteomics: A pragmatic perspective. *Nature Biotechnology*, *28*(7), 695–709. https://doi.org/10.1038/nbt.1658

Martelli, P. L., D'Antonio, M., Bonizzoni, P., Castrignanò, T., D'Erchia, A. M., de Meo, P. D. O., Fariselli, P., Finelli, M., Licciulli, F., Mangiulli, M., Mignone, F., Pavesi, G., Picardi, E., Rizzi, R., Rossi, I., Valletti, A., Zauli, A., Zambelli, F., Casadio, R., & Pesole, G. (2011). ASPicDB: A database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Research*, *39*(SUPPL. 1), 80–85. https://doi.org/10.1093/nar/gkq1073

Mattick, J. S. (2003). Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. In *BioEssays*. https://doi.org/10.1002/bies.10332

Melamud, E., & Moult, J. (2009). Stochastic noise in splicing machinery. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkp471

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. v., Djebali, S., Niarchou, A., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., … Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science*. https://doi.org/10.1126/science.aaa0355

Menon, R., Zhang, Q., Zhang, Y., Fermin, D., Bardeesy, N., DePinho, R. A., Lu, C., Hanash, S. M., Omenn, G. S., & States, D. J. (2009). Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-08-2145

Mercer, T. R., & Mattick, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Research*, *23*(7), 1081–1088. https://doi.org/10.1101/gr.156612.113

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science (New York, N.Y.)*, *338*(6114), 1593–1599. https://doi.org/10.1126/science.1228186

Muller, H. J. (1950). Our load of mutations. *American Journal of Human Genetics*.

Munoz, J., Low, T. Y., Kok, Y. J., Chin, A., Frese, C. K., Ding, V., Choo, A., & Heck, A. J. R. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Molecular Systems Biology*. https://doi.org/10.1038/msb.2011.84

Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., & Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology*. https://doi.org/10.1038/msb.2011.81

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, *463*(7280), 457–463. https://doi.org/10.1038/nature08909

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*. https://doi.org/10.1038/ng.259

Pearson, H. (2006). What is a gene? In *Nature*. https://doi.org/10.1038/441398a

Peled-Zehavi, H., Berglund, J. A., Rosbash, M., & Frankel, A. D. (2001). Recognition of RNA Branch Point Sequences by the KH Domain of Splicing Factor 1 (Mammalian Branch Point Binding Protein) in a Splicing Factor Complex. *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.21.15.5232-5241.2001

Pertea, M., & Salzberg, S. L. (2010). Between a chicken and a grape: Estimating the number of human genes. In *Genome Biology* (Vol. 11, Issue 5, p. 206). BioMed Central. https://doi.org/10.1186/gb-2010-11-5-206

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y. C., Madugundu, A. K., Pandey, A., & Salzberg, S. L. (2018). CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1590-2

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M. M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, *19*(7), 1316–1323. https://doi.org/10.1101/gr.080531.108

Ramensky, V. E., Nurtdinov, R. N., Neverov, A. D., Mironov, A. A., & Gelfand, M. S. (2008). Positive Selection in Alternatively Spliced Exons of Human Genes. *American Journal of Human Genetics*, *83*(1), 94–98. https://doi.org/10.1016/j.ajhg.2008.05.017

Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., & Huber, W. (2013). Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences*, *110*(38), 15377–15382. https://doi.org/10.1073/pnas.1307202110

Reyes, Alejandro, & Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1165

Rodriguez, J. M., Carro, A., Valencia, A., & Tress, M. L. (2015). APPRIS WebServer and WebServices. *Nucleic Acids Research*, *43*(W1), W455–W459. https://doi.org/10.1093/nar/gkv512

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J. J., Lopez, G., Valencia, A., & Tress, M. L. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, *41*(D1), 110–117. https://doi.org/10.1093/nar/gks1058

Rodriguez, J. M., Rodriguez-Rivas, J., di Domenico, T., Vázquez, J., Valencia, A., & Tress, M. L. (2018). APPRIS 2017: Principal isoforms for multiple gene sets. *Nucleic Acids Research*, *46*(D1), D213–D217. https://doi.org/10.1093/nar/gkx997

Rodriguez, J. M., Pozo F., di Domenico, T., Vázquez, J., & Tress, M. L. (2020). An analysis of tissue-specific alternative splicing at the protein level. *PLoS Computational Biology* (Accepted)

Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., & Dunker, A. K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0507916103

Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Green, R. K., Goodsell, D. S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A. S., Shao, C., … Burley, S. K. (2017). The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1000

Sakabe, N. J., & de Souza, S. J. (2007). Sequence features responsible for intron retention in human. *BMC Genomics*. https://doi.org/10.1186/1471-2164-8-59

Sánchez-Pla, A., Reverter, F., Ruíz de Villa, M. C., & Comabella, M. (2012). Transcriptomics: mRNA and alternative splicing. *Journal of Neuroimmunology*. https://doi.org/10.1016/j.jneuroim.2012.04.008

Sheynkman, G. M., Shortreed, M. R., Frey, B. L., & Smith, L. M. (2013). Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular and Cellular Proteomics*. https://doi.org/10.1074/mcp.O113.028142

Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K. I., & Go, M. (2009). AS-ALPS: A database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Research*, *37*(SUPPL. 1), 305–309. https://doi.org/10.1093/nar/gkn869

Shirota, M., & Kinoshita, K. (2016). Discrepancies between human DNA, mRNA and protein reference sequences and their relation to single nucleotide variants in the human population. *Database : The Journal of Biological Databases and Curation*, *2016*. https://doi.org/10.1093/database/baw124

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., … Kasprzyk, A. (2015). The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, *43*(W1), W589–W598. https://doi.org/10.1093/nar/gkv350

Smith, C. W. J., & Valcárcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in Biochemical Sciences*, *25*(8), 381–388. https://doi.org/10.1016/S0968-0004(00)01604-2

Southan, C. (2004). Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics*, *4*(6), 1712–1726. https://doi.org/10.1002/pmic.200300700

Southan, C. (2017). Last rolls of the yoyo: Assessing the human canonical protein count. *F1000Research*, *6*(0), 1–23. https://doi.org/10.12688/f1000research.11119.1

Spuhler, J. N. (1948). On the number of genes in man. *Science*. https://doi.org/10.1126/science.108.2802.279-a

Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome: Motors, clocks, springs, and things. In *Cell*. https://doi.org/10.1016/S0092-8674(00)80925-3

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S. E., Behr, J., Bertone, P., Bohnert, R., Bucher, P., Cloonan, N., Derrien, T., Djebali, S., Du, J., Dudoit, S., Gerstein, M., Gingeras, T. R., … Zhang, M. Q. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, *10*(november), 7–9. https://doi.org/10.1038/nmeth.2714

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L. B., Posukh, O. L., Sahakyan, H., Watkins, W. S., Yepiskoposyan, L., Abdullah, M. S., Bravi, C. M., Capelli, C., Hervig, T., … Eichler, E. E. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science*, *349*(6253). https://doi.org/10.1126/science.aab3761

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H. Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Sugnet, C. W., Kent, W. J., Ares, M., & Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. https://doi.org/10.1142/9789812704856_0007

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P., & Bafna, V. (2007). Improving gene annotation using peptide mass spectrometry. *Genome Research*. https://doi.org/10.1101/gr.5646507

Tazi, J., Bakkour, N., & Stamm, S. (2009). Alternative splicing and disease. In *Biochimica et Biophysica Acta - Molecular Basis of Disease*. https://doi.org/10.1016/j.bbadis.2008.09.017

Tress, M. L., Abascal, F., & Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, *42*(2), 98–110. https://doi.org/10.1016/j.tibs.2016.08.008

Tress, M. L., Bodenmiller, B., Aebersold, R., & Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology*, *9*(11), R162. https://doi.org/10.1186/gb-2008-9-11-r162

Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Ólason, P. Í., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R. A., López, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., … Valencia, A. (2007). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(13), 5495–5500. https://doi.org/10.1073/pnas.0700800104

Tress, M. L., Wesselink, J. J., Frankish, A., López, G., Goldman, N., Löytynoja, A., Massingham, T., Pardi, F., Whelan, S., Harrow, J., & Valencia, A. (2008). Determination and validation of principal gene products. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btm547

Trowitzsch, S., Weber, G., Lührmann, R., & Wahl, M. C. (2009). Crystal Structure of the Pml1p Subunit of the Yeast Precursor mRNA Retention and Splicing Complex. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2008.10.087

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

The UniProt Consortium. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky092

U.S. Department of Health and Human Services & Department of Energy. (1990). *Understanding our genetic inheritance. The U.S. Human Genome project: The first five years: fiscal years 1991-1995*.

Venables, J. P. (2004). Aberrant and alternative splicing in cancer. In *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-04-1910

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science*. https://doi.org/10.1126/science.1058040

Viklund, H., & Elofsson, A. (2004). Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Science*. https://doi.org/10.1110/ps.04625404

Vogel, F. (1964). A preliminary estimate of the number of human genes. *Nature*. https://doi.org/10.1038/201847a0

Vuong, C. K., Black, D. L., & Zheng, S. (2016). The neurogenetics of alternative splicing HHS Public Access. *Nat Rev Neurosci*. https://doi.org/10.1038/nrn.2016.27

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, G. S., & Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2164

Washburn, M. P., Wolters, D., & Yates, J. R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*. https://doi.org/10.1038/85686

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., … Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*. https://doi.org/10.1038/nature13319

Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D., Wang, J. P., Yang, H. Y., Baer, T., Stredney, D., Spitzner, J., Stutz, A., Krahe, R., & Yuan, B. (2001). A draft annotation and overview of the human genome. *Genome Biology*. https://doi.org/10.1186/gb-2001-2-7-research0025

Wright, J. C., Mudge, J., Weisser, H., Barzine, M. P., Gonzalez, J. M., Brazma, A., Choudhary, J. S., & Harrow, J. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature Communications*. https://doi.org/10.1038/ncomms11778

Xing, Y., & Lee, C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(38), 13526–13531. https://doi.org/10.1073/pnas.0501213102

Xuan, Z., Wang, J., & Zhang, M. Q. (2003). Computational comparison of two mouse draft genomes and the human golden path. *Genome Biology*. https://doi.org/10.1186/gb-2002-4-1-r1

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. In *Nature Reviews Genetics* (Vol. 16, Issue 3, pp. 172–183). Nature Publishing Group. https://doi.org/10.1038/nrg3871

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1098

Zhang, G. (2015). Genomics: Bird sequencing project takes off. In *Nature* (Vol. 522, Issue 7554, p. 34). Nature Publishing Group. https://doi.org/10.1038/522034d