Steward et al. Genome Medicine (2017) 9:49
DOI 10.1186/s13073-017-0441-1

Genome Medicine

REVIEW Open Access

Genome annotation for clinical genomic diagnostics: strengths and weaknesses

Charles A. Steward[1,2*], Alasdair P. J. Parker[3], Berge A. Minassian[4,5], Sanjay M. Sisodiya[6,7], Adam Frankish[2,8] and Jennifer Harrow[2,9]

Abstract

The Human Genome Project and advances in DNA sequencing technologies have revolutionized the identification of genetic disorders through the use of clinical exome sequencing. However, in a considerable number of patients, the genetic basis remains unclear. As clinicians begin to consider whole-genome sequencing, an understanding of the processes and tools involved and the factors to consider in the annotation of the structure and function of genomic elements that might influence variant identification is crucial. Here, we discuss and illustrate the strengths and weaknesses of approaches for the annotation and classification of important elements of protein-coding genes, other genomic elements such as pseudogenes and the non-coding genome, comparative-genomic approaches for inferring gene function, and new technologies for aiding genome annotation, as a practical guide for clinicians when considering pathogenic sequence variation. Complete and accurate annotation of structure and function of genome features has the potential to reduce both false-negative (from missing annotation) and false-positive (from incorrect annotation) errors in causal variant identification in exome and genome sequences. Re-analysis of unsolved cases will be necessary as newer technology improves genome annotation, potentially improving the rate of diagnosis.

Background

Advances in genomic technologies over the past 20 years have provided researchers with unprecedented data relating to genome variation in different diseases [1]. However, even after whole-exome sequencing (WES), the genetic basis for a particular phenotype remains unclear in a considerable proportion of patients. Here, we examine how genomic annotation might influence variant identification, using examples mostly from both common and rarer neurological disorders. We highlight why the present technology can fail to identify the pathogenic basis of a patient's disorder, or produce an incorrect result where the wrong variant is labelled as causative. For these reasons, we believe it is important to re-analyse unresolved cases as newer technology and software improve gene and genome annotation. The aim of this paper is to make common genomic techniques accessible to clinicians through the use of figures and examples that help to explain genome sequencing, gene classification and genome annotation in the context of pathogenic sequence variation. Finally, we discuss how new genomic techniques will improve our ability to identify pathogenic sequence variation.

Genome sequencing

The Human Genome Project (HGP) was launched officially in 1987 by the US Department of Energy to sequence the approximately 3 billion base-pairs (bp) that constitute the human genome [2]. The first draft sequence was published in 2001 and computational annotation, a process that attributes a biological function to the genomic elements, described 30,000 to 40,000 protein-coding genes across 22 pairs of autosomes and the X and Y sex chromosomes in a genome of 2.9 billion bases (gigabases, Gb) [2]. The precise size and gene count of the reference human genome remains uncertain to this day because sequence gaps remain, while the

* Correspondence: charles.steward@congenica.com
[1]Congenica Ltd, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK
[2]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
Full list of author information is available at the end of the article

classification of genes becomes more refined [3]. Consequently, additions are continually made to the genome to fill sequence gaps [4]. The most recent published estimates suggest that just under 20,000 protein-coding genes [5] are present in a genome of approximately 3.1 Gb [6]. The HGP enabled initial research examining sequence variation on chromosome 22 [7], to more recent medical advances that now see DNA sequencing used routinely in large-scale research programs, such as the Deciphering Developmental Disorders (DDD) study [8, 9]. Sequencing for the HGP used the chain terminator method [10], more commonly known as 'Sanger sequencing', and owing to the better-quality sequence data and read-length associated with Sanger sequencing compared with current sequencing technologies, Sanger sequencing is still used to confirm sequence variants [11].

Current methods for producing the raw sequence data for whole-genome sequencing (WGS) are placed into two categories based upon the length of the nucleotide sequence produced, or sequence 'read'. Short-read technology comes from Illumina Inc. [12] and uses well-established chemistry to identify the sequence of nucleotides in a given short segment of DNA. Illumina sequencing platforms such as the HiSeq X produce base-pair reads of lengths from 150 to 250 bp in a given DNA segment and are used to read sequences from both ends of a DNA fragment. This 'next-generation' technology is a dramatic improvement over older Sanger sequencing methods that produced longer reads but at much higher cost [13]. More recently, 'third-generation' technologies from Pacific Biosciences (PacBio) and Oxford Nanopore are gaining users and making an impact. These third-generation methods generate longer reads, up to tens of thousands of base-pairs per read, but with higher error rates.

The speed of DNA sequencing, the amount of sequence that can be produced and the number of genomes that can be sequenced have increased massively with next-generation sequencing (NGS) techniques [14]. Such advances have enabled large collaborative projects that look at variation in a population, such as the 1000 Genomes Project [15], as well as those investigating the medical value of WGS, such as the UK 100,000 Genomes Project [16]. It is hoped that WGS will facilitate the research, diagnosis and treatment of many diseases.

Once a patient genome has been sequenced, it needs to be aligned to the reference genome and analysed for variants. Typically, software algorithms such as the Burrows-Wheeler Aligner (BWA) are used for short- [17] and long-read [18] alignment and the Genome Analysis Toolkit (GATK) is used to identify or 'call' sequence variants [19]. Figure 1 illustrates a typical genome analysis pipeline, describing the different file formats commonly used—FASTQ [20], BAM [21] and VCF [22].

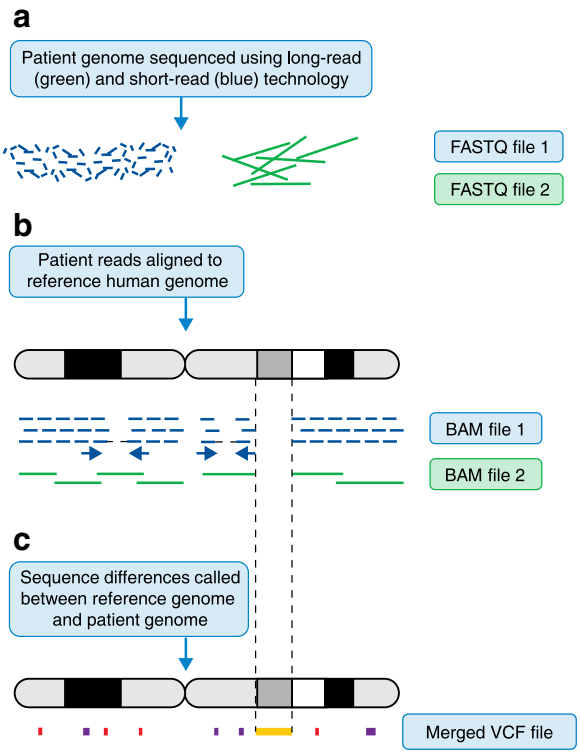

Fig. 1 The genome analysis pipeline. Note that, for clarity, some steps have been omitted. Figure illustrations are not to scale and are only meant to be illustrative of the differences between short- and long-read sequencing. **a** Unaligned reads from sequencing machines are stored as FASTQ file formats. This is a text-based format for storing both a DNA sequence and its corresponding quality scores. **b** Reads are aligned to the genome. Short reads provide deep coverage, whereas reads that have been sequenced from both ends (*blue arrows*) help to orientate unaligned contigs. It is difficult to align short reads confidently across repetitive sequences when the repeating genome sequence is longer than the sequence read. Long-read sequences help to order contigs across larger regions, particularly with repetitive sequences, but do not provide the necessary depth needed to be confident of calling a base at a certain position. Note that there is a large region where there is no read coverage at all. This is indicative of structural variation. Here, the patient has a large deletion with respect to the reference genome. Once the reads have been aligned to the reference genome they are stored in a BAM file. A BAM file (.bam) is the binary version of a sequence alignment map (SAM file format). The latter is a tab-delimited text-based format for storing DNA sequences aligned to a reference sequence. **c** The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing genetic sequence variations. VCF files are much smaller than FASTQ and BAM files. Note that single-nucleotide variants (SNVs) and small insertions and deletions ('indels') are illustrated as *red* and *purple blocks*, whereas a much larger structural variant is indicated by an *orange block*

Pathogenic sequence variation can range in size from single-nucleotide variants (SNVs), small insertions and deletions ('indels') of fewer than 50 base-pairs in length, to larger structural variants (SVs) [23], which are generally classified as regions of genomic variation greater

than 1 kb, such as copy-number variants (CNVs), insertions, retrotransposon elements, inversions, segmental duplications, and other such genomic rearrangements [24, 25]. Currently, the consequence of non-synonymous variants of the protein-coding elements only can be routinely automatically predicted by algorithms such as SIFT and PolyPhen [26], yet many different types of variants are implicated in disease. As sequencing techniques begin to move away from 'gene panel' testing to WGS, it is crucial to understand the structure of genes and any regulatory features that might lie within intra/intergenic regions as changes in any of these regions might have a crucial impact on the function of a gene.

Recently, the American College of Medical Genetics and Genomics (ACMG) recommended a set of standards and guidelines to help medical geneticists assign pathogenicity using standardized nomenclature and evidence used to support the assignment for Mendelian disorders [27]. For example, the terms 'mutation' and 'polymorphism' have often been used misleadingly, with assumptions made that 'mutation' is pathogenic, whereas 'polymorphism' is benign. As such, one recommendation that ACMG makes is that both these terms are replaced by 'variant', with the following modifiers (1) pathogenic, (2) likely pathogenic, (3) uncertain significance, (4) likely benign, or (5) benign [27]. As such, here, we use the term variant. A standard gene-variant nomenclature is maintained and versioned by the Human Genome Variation Society (HGVS) [28]. Both ACMG and HGVS examples are illustrated in Table 1.

Classifying genes and other genomic elements

Current gene sets identify under 20,000 protein-coding genes and over 15,000 long non-coding RNAs (lncRNAs) [29, 30]. In this section, for clinicians who might not be familiar with gene structure and function, we present the important elements of different parts of protein-coding genes, and other categories of genomic elements, such as pseudogenes and elements of the non-coding genome such as lncRNAs, and we highlight their potential functionality, illustrated with examples of their roles in disease. We demonstrate the importance of classifying such regions correctly and why incorrect classification could impact the interpretation of sequence variation.

Important elements of protein coding genes

A eukaryotic gene is typically organized into exons and introns (Fig. 2), although some genes, for example *SOX3*, which is associated with X-linked mental retardation [31], can have a single exon structure. The functional regions of protein-coding genes are typically designated as the coding sequence (CDS) and the 5′ and 3′ untranslated regions (UTRs) (Fig. 2).

The 5′ UTR of a transcript contains regulatory regions. For example, some upstream open reading frames (uORFs; which are sequences that begin with an ATG codon and end in a stop codon, meaning that they have the potential to be translated) in the 5′ UTR are translated to produce proteins that could enhance or suppress the function of the main CDS [32]. Experimental techniques such as cap-analysis gene expression (CAGE) [33] are used to identify transcription start sites (TSSs) (Fig. 2a).

Table 1 Examples of disease-causing variation with associated HGVS nomenclature

Location	Gene	Variation	HGVS nomenclature	ACMG clinical significance	Associated disorder	Reference
5′ UTR	*FMR1*	Expansion	NM_002024.5(FMR1):c.-128_-126(200)	Pathogenic	Fragile X syndrome	[186]
CDS	*GRIN2A*	Nonsense	NM_000833.4(GRIN2A):c.2041C > T (p.Arg681Ter)	Pathogenic	Idiopathic focal epilepsy (IFE) with rolandic spikes is the most common childhood epilepsy	[187]
CDS	*GABRB3*	Missense	NM_021912.4(GABRB3):c.745C > A (p.Gln249Lys)	Pathogenic	Early infantile epileptic encephalopathy (EIEE)	[188]
CDS	*WDR62*	Deletion/ frameshift	NM_001083961.1(WDR62) :c.3839_3855del17 (p.Gly1280Alafs)	Pathogenic	Malformations of cortical development	[189]
3′ UTR	*MECP2*	SNV	NM_004992.3(MECP2):c.*2956G > A	Uncertain significance	Rett syndrome	[190]
Promoter	*CRH*	SNV	NC_000008.11:g.66178947G > T	Pathogenic	Familial autosomal dominant nocturnal frontal lobe epilepsy	[191]
Splice site	*ATP6AP2*	SNV	NM_005765.2(ATP6AP2):c.321C > T (p.Asp107=)	Pathogenic	X-linked mental retardation and epilepsy due to inefficient inclusion of exon 4	[192]
Poly(A)	*ARSA*	SNV	NM_000487.5(ARSA):c.*96A > G	Pathogenic	Metachromatic leukodystrophy	[193]
NMD	*SNRPB*	SNV	NM_003091.3(SNRPB):c.-72C > A	Pathogenic	Cerebro-costo-mandibular syndrome	[194]
lncRNA	*ATXN8OS*	Insertion	NR_002717.2(ATXN8OS) :n.1103_1105CTG(15_40)	Pathogenic	Spinocerebellar ataxia type 8	[195]

ACMG American College of Medical Genetics and Genomics, *CDS* coding sequence, *HGVS* Human Genome Variation Society, *lncRNA* long non-coding RNA, *NMD* nonsense-mediated decay, *SNV* single-nucleotide variant, *UTR* untranslated region

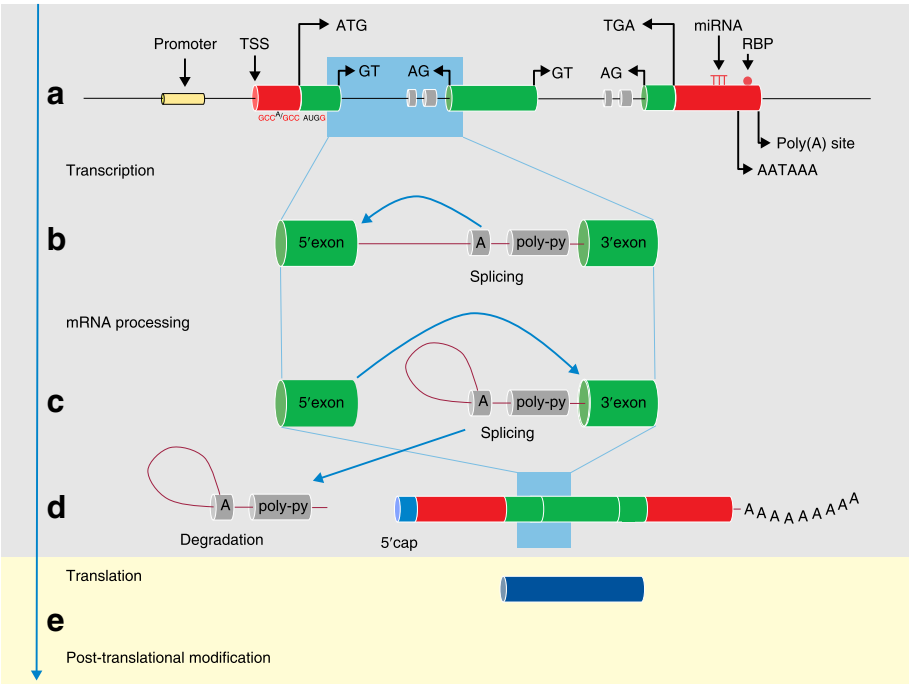

Fig. 2 The generic gene model (not to scale). **a** The exons comprise the untranslated regions (UTRs), which are shown in *red* (the 5′ UTR depicted on the *left* and the 3′ UTR depicted on the *right*) and the coding sequence (CDS), which is shown in *green*. Many important regulatory regions lie outside of the exons of a gene. Intronic regulatory regions are shown in *grey*. Promoters are illustrated as *yellow* intergenic regulatory regions, although some genes have internal transcription start sites. The transcription start site (*TSS*) is positioned at the 5′ end of the UTR, where transcription starts. The 5′ UTRs of genes contain regulatory regions. The CDS start codon is the first codon of a messenger RNA (mRNA) from which a ribosome translates. The genomic sequence around the start codon often has the consensus sequence gcc**A**cc|**AUG**|**G** (note that the important bases are highlighted here in bold, whereas the most crucial positions are −3 and +4 from the A of the AUG) [197], although, in very rare cases, a non-AUG start codon is used [198]. The stop codon, of which there are three in eukaryotes—UGA, UAG, UAA—is a nucleotide triplet sequence in an mRNA that gives the signal to terminate translation by binding release factors, causing the ribosome to release the peptide chain [199]. The 3′ untranslated region of genes contains regulatory regions. In particular, the 3′ UTR has binding sites for regulatory proteins such as RNA-binding proteins (*RBP*) and microRNAs (*miRNA*). Promoters are DNA sequences, between 100 and 1000 bp in length, where proteins that help control gene transcription bind to DNA [200]. These proteins can contain one or more DNA-binding domains that attach to a specific DNA sequence located next to the relevant gene [201]. Promoters regulate transcriptional machinery by moving it to the right place in the genome, as well as locating the 5′ end of the gene or an internal transcription start site. Approximately 40% of human genes have promoters situated in regions of elevated cytosine and guanine content, termed CpG islands [202]. A subset of promoters incorporate the variable TATA box sequence motif, which is found between 25 and 30 bp upstream of the TSS and is the position at the 5′ end of the UTR where transcription starts [203]. **b**–**d** Pre-mRNA transcribed from DNA contains both introns and exons. An RNA and protein complex called the spliceosome undertakes the splicing out of introns, leaving the constitutive exons. Intronic and exonic splice enhancers and silencers help direct this procedure, such as the branch point (*'A'*) and a poly-pyrimidine (*poly-py*) tract. The vast majority of introns have a GT sequence at the 5′ end that the branch point binds to. The intron is then cleaved from the 5′ exon (donor site) and then from the 3′ exon (acceptor site) [204] and a phosphodiester bond joins the exons, whereas the intron is discarded and degraded. During the formation of mature mRNA, the pre-mRNA is cleaved and polyadenylated. Polyadenylation occurs between 10 and 30 bp downstream from a hexamer recognition sequence that is generally AAUAAA, or AUUAAA, although other hexamer signal sequences are known [35] (as depicted in **a**). A specially modified nucleotide at the 5′ end of the mRNA, called the 5′ cap, helps with mRNA stability while it undergoes translation. This capping process occurs in the nucleus and is a vital procedure that creates the mature mRNA. **e** The translation of mRNA into protein by ribosomes occurs in the cytosol. Transfer RNAs (tRNAs), which carry specific amino acids, are read by the ribosome and then bound in a complementary manner to the mRNA. The amino acids are joined together into a polypeptide chain to generate the complete protein sequence for the coding sequence of the transcript. (*Light blue background shading* shows processes that occur in the nucleus. *Light yellow background shading* shows processes that occur in the cytosol, such as the translation of mRNAs into protein by ribosomes)

Variants in the CDS are generally the most well studied and understood area of pathogenic sequence variation. For example, approximately 700 pathogenic CDS variants have been reported in the epilepsy-associated gene *SCN1A* [34].

The 3′ UTR of a transcript can contain regions controlling regulatory proteins such as RNA binding proteins (RBPs) and microRNAs (miRNAs) (Fig. 2a). Interestingly, the 3′ UTR has been linked to overall translation efficiency and stability of the mRNA [35]. The 5′ and 3′ UTRs can also interact with each other to regulate translation through a closed-loop mechanism [36]. Important sequence motifs involved in controlling the expression of a gene include promoters, enhancers and

silencers, which are found in exonic, intragenic and intergenic regions (Fig. 2a).

A multi-exonic eukaryotic gene can produce different disease phenotypes through alternative protein isoforms that result from the use of alternative splice site/exon combinations (Fig. 3) [37]. Canonical splice sites are generally conserved at the 5′ (donor) and 3′ (acceptor) ends of vertebrate introns. The GT–intron–AG configuration is the most common, although other, rarer instances of splice sites are found, such as GC–intron–AG and AT–intron–AC [38].

Although there can be an abundant transcript that is expressed in a particular cell, the same transcript might not dominate elsewhere, and, even if a dominant transcript is identified, the transcript might not be functional [39]. Differential expression can be both tissue- and age-specific [40], can occur in response to different environmental signals [41, 42], and an exon expressed in one tissue might not be relevant to further analysis if it is not expressed in the tissue where a disease phenotype is present. For example, genes expressed in brain generally have longer 3′ UTRs than those in other tissues, and such differences could impact miRNA binding sites and other regulatory regions [43]. Studies have shown that retained introns have an important role in brain gene expression and regulation [44, 45].

Polyadenylation (poly(A)), which involves addition of the poly(A) tail, is important for nuclear export to the cytosol for translation by the ribosome and also helps with mRNA stability (Fig. 2d). Many annotated genes also have more than one poly(A) site, which can be functional in different tissues or different stages of development [42].

After translation, the polypeptide chain produced by the ribosome might need to undergo posttranslational modification, such as folding, cutting or chemical modifications, before it is considered to be a mature protein product (Fig. 2e). Noonan syndrome is believed to result from the disruption of the phosphorylation-mediated auto-inhibitory loop of the Src-homology 2 (SH2) domain during post-translational modification [46].

Transcripts that contain premature stop codons (perhaps as a result of using an alternative splice donor, splice acceptor, or inclusion/exclusion of an alternative exon, which causes a CDS frameshift) are degraded through the nonsense-mediated decay (NMD) cellular surveillance pathway (Fig. 4) [47, 48]. NMD was originally believed to degrade erroneous transcripts, but much evidence has been found to suggest it is also an active regulator of transcription [49, 50]. Several NMD factors have been shown to be important for the regulation of neurological events such as synaptic plasticity and neurogenesis [51–53].

Two other types of cellular surveillance pathways are known to exist: non-stop decay and no-go decay. Non-stop decay is a process that affects transcripts that have

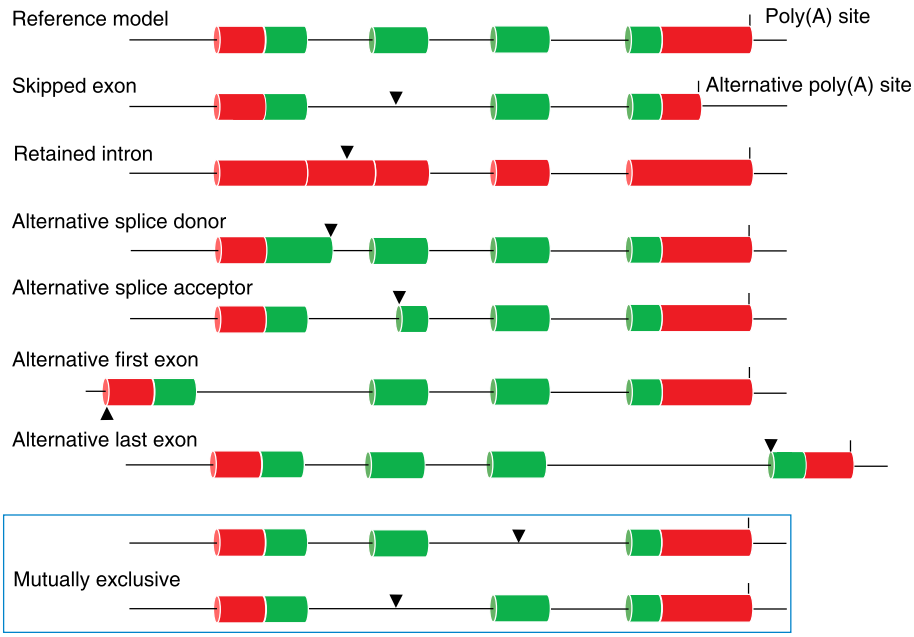

Fig. 3 Alternative splicing transcript variants. Different types of alternative splicing can give rise to transcripts that are functionally distinct from a nominal reference model. *Red* represents the untranslated region (UTR) and *green* represents the coding sequence (CDS). The retained intron is illustrated as non-coding as a retained intron is presumed to represent an immature transcript. Some transcripts can contain exons that are mutually exclusive (*boxed*). All the types of alternative exon splicing events shown here can also occur in non-coding genes. There can also be multiple alternative poly(A) features within the gene models, as seen for the skipped-exon transcript

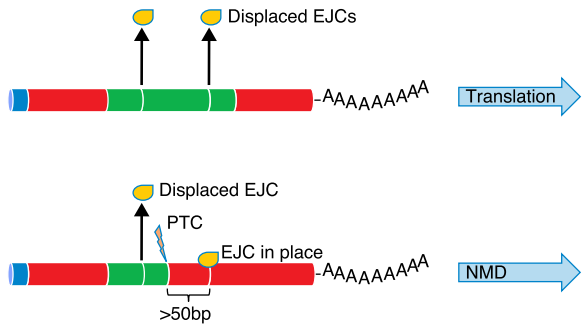

Fig. 4 The nonsense-mediated decay (NMD) pathway. Under normal cellular circumstances, exon–exon junction complexes (*EJCs*) that are in place after splicing are removed by the ribosome during the first round of translation. However, when a transcript contains a premature termination codon (*PTC*), perhaps as a result of an single-nucleotide variant (SNV), indel or inclusion of an out-of-frame exon upstream of one or more EJCs, these EJCs remain in place because the ribosome complex disassociates at the premature stop codon and thus cannot remove the downstream EJC. This triggers the NMD pathway, and the transcript is degraded

poly(A) features but do not have a prior stop codon in the CDS. The translation of such transcripts could produce harmful peptides with a poly-lysine amino acid sequence at the C-terminal end of the peptide—therefore, these transcripts are subject to degradation. Similar to NMD transcripts, either aberrant splicing or SNVs can cause the generation of these transcripts [54]. Finally, no-go decay is triggered by barriers that block ribosome movement on the mRNA [55].

The functional importance of pseudogenes

Pseudogenes are traditionally regarded as 'broken' copies of active genes. Freed of selective pressure, they have typically lost the ability to encode functional proteins through the occurrence of nonsense variations, frameshifts, truncation events, or loss of essential regulatory elements. The majority of pseudogenes fall into one of two categories: processed and unprocessed (Fig. 5, Table 2) [56].

Processed pseudogenes represent back-integration or retrotransposition of an RNA molecule into the genome sequence, and, although they generally lack introns, they frequently incorporate the remains of the poly(A) tail. Processed pseudogenes are often flanked by direct repeats that might have some function in inserting the pseudogene into the genome, and are often missing sequence compared with their parent gene (Fig. 5) [57]. By contrast, unprocessed pseudogenes are defunct relatives of functional genes that arise through faulty genomic duplication resulting in missing (parts of) exons and/or flanking regulatory regions (Fig. 5).

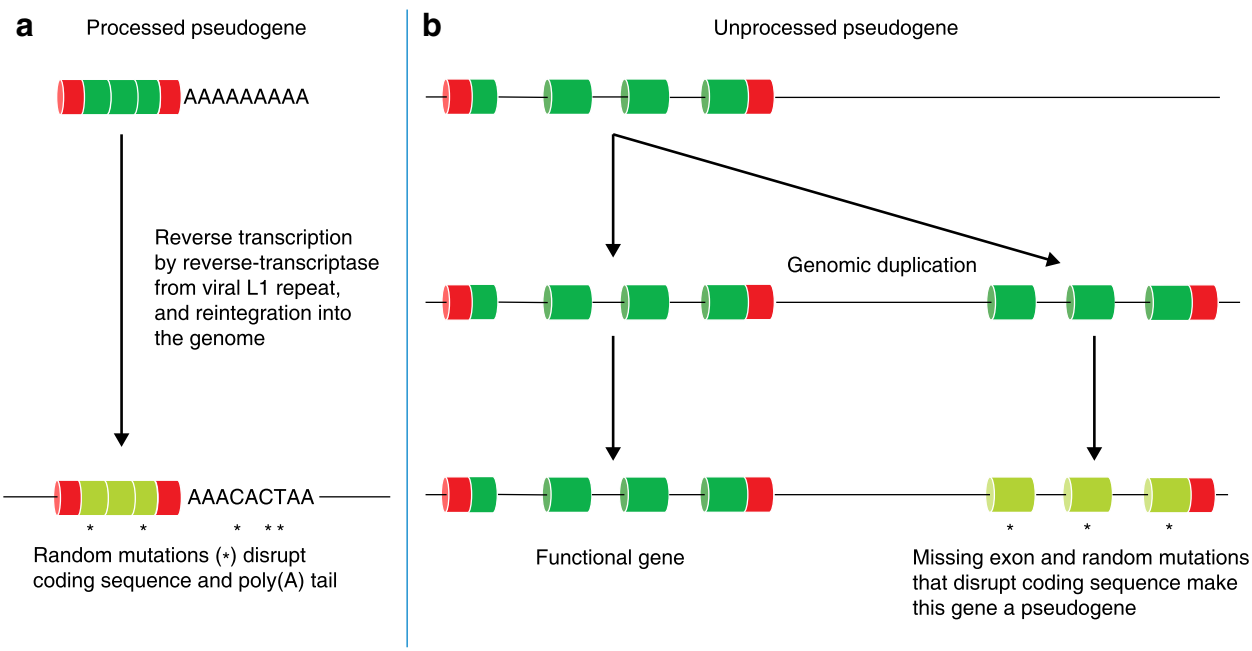

Fig. 5 The processes involved in the 'pseudogenisation' of genes. **a** Processed pseudogenes are derived from mature mRNA that is reverse-transcribed by the viral L1 repeat enzyme reverse-transcriptase and reintegrated into the genome, and will generally lack introns. Processed pseudogenes are often flanked by direct repeats that might have some function in inserting the pseudogene into the genome and they are often missing sequence compared with their parent. Often they terminate in a series of adenines, which are the remains of the poly(A) tail, which is the site of genomic integration. **b** Unprocessed pseudogenes—the defunct relatives of functional genes—arise from genomic duplication. Such duplications can be complete or partial with respect to the parent gene

Table 2 GENCODE annotation biotypes (2017)

Biotype	Description
Protein coding	Contains an ORF that has strong coding potential
Known coding	100% identical to known RefSeq protein or Swiss-Prot entry
Novel coding	Shares >60% length with known coding sequence from RefSeq, or Swiss-Prot, or has cross-species/family support or domain evidence
Putative coding	Shares <60% length with known coding sequence from RefSeq, or Swiss-Prot, or has an alternative first or last coding exon
Nonsense-mediated decay	If the coding sequence (following the appropriate reference) of a transcript finishes >50 bp from a downstream splice site, then it is tagged as NMD. If the variant does not cover the full reference coding sequence, then it is annotated as NMD if NMD is unavoidable—i.e. no matter what the exon structure of the missing portion is, the transcript will be subject to NMD
Non-stop decay	Transcripts that have poly(A) features (including signal) without a prior stop codon in the CDS—i.e. a non-genomic poly(A) tail attached directly to the CDS without a 3′ UTR; these transcripts are subject to degradation
Retained intron	Alternatively spliced transcript believed to contain intronic sequence relative to other, coding, variants
Processed transcript	Cannot assign an ORF, but is part of a coding locus
lncRNA	Long non-coding RNA—lacks protein-coding potential and is of length >200 bp
Bidirectional promoter	Transcription start sites of the lncRNA model and the protein-coding model are on opposite strands and within 200 bp of one another, or are found in the same CpG island
3-Prime overlapping	Transcription start site and/or published experimental data support independent transcription from the 3′ UTR of a coding gene
Antisense	At least one variant overlaps a protein-coding locus on the opposite strand, or evidence of antisense regulation of a coding gene has been published
lincRNA	Long intergenic ncRNA: does not overlap (neither sense nor antisense) a coding gene
Sense intronic	In an intron of a coding gene; no exonic overlap
Sense overlapping	Contains a coding gene in an intron; no exonic overlap.
Pseudogene	Matches to protein, but ORF disrupted by frameshifts and/or premature stop codons
Processed	Lacks introns and arose from retrotransposition of parent gene mRNA
Unprocessed	Can contain introns and is produced by genomic duplication
Transcribed	Locus-specific transcripts indicate transcription; these can be classified into '*processed*' and '*unprocessed*'
Translated	Locus-specific protein mass spectroscopy data suggest translation; the connection is maintained with the pseudogene biotype until the experimental community validates it as a coding gene
Polymorphic	Pseudogene owing to a single-nucleotide variant (SNV), or insertion-deletion variant (indel); but the same gene is translated in other individuals/haplotypes/strains
Unitary	Species-specific unprocessed pseudogene, without a parent gene, that has an active orthologue in another species

Data sourced from GENCODE project [196]
ncRNA noncoding RNA, *ORF* open reading frame, *UTR* untranslated region

Computational annotation of pseudogenes tends to suffer from significant false positives/negatives and can cause problems that result from the misalignment of NGS data. Specifically, identification of transcribed pseudogenes and single-exon pseudogenes can be a challenge [58]. Such difficulties were demonstrated where it was found that more than 900 human pseudogenes have evidence of transcription, indicating functional potential [58, 59]. Consequently, the ability to distinguish between pseudogenes and the functional parent gene is essential when predicting the consequence of variants.

MacArthur and colleagues [60] reported that reference sequence and gene annotation errors accounted for 44.9% of candidate loss-of-function (LoF) variants in the NA12878 genome, which belongs to the daughter from a trio of individuals belonging to the CEPH/Utah pedigree whose genomes were sequenced to high depth as part of the HapMap project [61]. The NA12878 genome sequence and transformed cells from the same individual (the GM12878 cell line) are often used as a reference in other projects [62, 63]. After reannotation of protein-coding genes harbouring 884 putative LoF variants, 243 errors in gene models were identified, 47 (19.3%) of which were updated from protein-coding to pseudogene, removing a significant source of false-positive LoF annotation [60].

Transcripts derived from the pseudogene locus *PTENP1* have been shown to regulate the parent *PTEN* locus [64]. Deletion of *PTENP1* has been reported to downregulate *PTEN* expression in breast and colon cancer [64] and melanoma [65], and downregulation of *PTENP1* through methylation of its promoter sequence in clear-cell renal cell carcinoma suppresses cancer progression [66]. Although *PTENP1* has not yet been associated with any neuronal disorders, both *PTEN* and *PTENP1* are expressed in multiple brain tissues [67, 68].

The non-coding genome

Most of the genome is non-coding, and therefore most variation occurs in non-coding regions. To understand the effect of a sequence variant in such regions, the non-coding elements need to be classified. Non-coding elements consist of *cis*-regulatory elements such as promoters and distal elements (for example, enhancers) [69] and non-coding RNAs (ncRNAs). Large collaborative initiatives, such as ENCODE [63] and RoadMap Epigenomics [70], have been tasked to create comprehensive maps of these regions. The Ensembl regulatory build [71] and Variant Effect Predictor (VEP) [72] are able to determine whether variants fall within such regions, but are not yet able to determine pathogenicity, although tools that do so are beginning to emerge, such as FunSeq [73] and Genomiser [74].

The ncRNAs are generally divided into two groups, small RNAs (sRNAs) and lncRNAs. sRNAs include miRNAs, Piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs) and other short RNAs [75]. The sRNAs can be predicted using tools such as Infernal [76] and Rfam [77], which makes the interpretation of sequence variation and consequence easier, especially when compared with the analysis of lncRNAs. However, correctly discriminating functional copies from pseudogenes remains a challenge.

Of particular interest to the study of neurological disease are microRNAs (miRNAs), which are small (approximately 20 nucleotides) ncRNAs that are involved in the regulation of post-transcriptional gene expression [78]. miRNAs can trigger transcript degradation, modify translational efficiency and downregulate gene expression by triggering epigenetic changes (DNA methylation and histone modifications) at the promoter of target genes, and are the best-understood of the ncRNAs. Studies have shown that variants in miRNA binding sites are associated with some neurological diseases, and there is evidence for a role in epilepsy, suggesting that miRNAs might be good candidates for the development of novel molecular approaches for the treatment of patients with epilepsy [79, 80]. For example, miRNA *MIR328* binds to the 3′ UTR of *PAX6* to regulate its expression. However, variation in the miRNA binding site reduces the binding affinity of *MIR328*, which in turn results in an increase in the abundance of *PAX6* transcripts, which is associated with electrophysiological features of Rolandic epilepsy [81]. The EpiMiRNA consortium is investigating the role of miRNAs in the development, treatment and diagnosis of temporal lobe epilepsy [82].

The classification of lncRNAs is increasingly used to convey functional information, despite the fact that we know relatively little about the role or mechanism of the vast majority of them (Fig. 6). The term lncRNA was itself established to distinguish longer ncRNAs from the small ncRNAs that were initially separated using an experimental threshold of >200 nucleotides, which remains the simplest definition of a lncRNA [63]. RNA sequencing (RNA-Seq) assays predict that potentially tens, if not hundreds, of thousands of lncRNA transcripts have now been identified [83], which has inevitably led to the naming of many proposed subclasses of lncRNA [84, 85]. Without any international agreement on the classification of lncRNAs, proposed subclasses have been classified based on either length, function, sequence or structural conservation, or

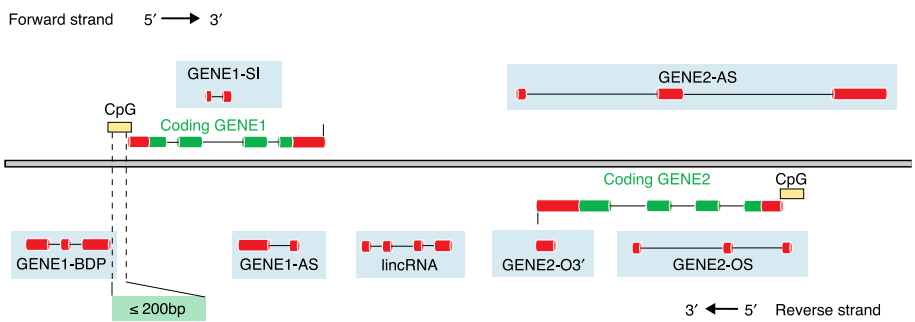

Fig. 6 Different classifications of long non-coding RNAs (lncRNAs). The classification of lncRNAs is based on their position with respect to coding genes. lncRNAs are illustrated here with only *red* exons, whereas the coding genes are shown as *red* and *green*. *AS* antisense, *BDP* bi-directional promoter, *lincRNA* long-intergenic RNA (not overlapping a protein-coding locus on either strand), *OS* overlapping sense, *O3′* overlapping 3′, *SI* sense intronic. Figure adapted from Wright 2014 [84]

association with either protein-coding genes, DNA elements, subcellular location or a particular biological state. They are hard to predict owing to their size, but also because they are expressed at low levels and lack a known tertiary structure, unlike miRNAs. A recent study by Nitsche and colleagues showed that >85% of lncRNAs have conserved splice sites that can be dated back to the divergence of placental mammals [86].

lncRNAs, such as *XIST* [87], have been studied for some time, yet little is known about the function of most. However, they are gaining interest within the scientific and medical community [63] owing to their potential involvement in disease [88, 89]. Experiments in mouse models have demonstrated that dysregulation of certain lncRNAs could be associated with epilepsy [90], and a role in gene regulation is proposed for the vast number of unstudied cases [91], which makes them interesting candidates for new targeted therapies and disease diagnostics [92]. For example, experiments in a knock-in mouse model of Dravet syndrome have shown that the upregulation of the healthy allele of *SCN1A* by targeting a lncRNA improved the seizure phenotype [93].

CNVs also play an important role in human disease and can affect multiple coding genes, resulting in dosage effects, truncation of single genes or novel fusion products between two genes. CNVs have also been shown to be pathogenic in non-coding regions [94]. Talkowski and colleagues [95] observed a CNV causing disruption in the long-intergenic non-coding RNA (lincRNA) *LINC00299* in patients with severe developmental delay, raising the possibility that lincRNAs could play a significant role in developmental disorders. More recently, Turner et al. [96] reported WGS of 208 patients from 53 families with simplex autism and discovered small deletions within non-coding putative regulatory regions of *DSCAM*, implicated in neurocognitive dysfunction in Down syndrome. These CNVs were transmitted from the mother to the male proband.

Repetitive sequences and transposable elements are known to be involved in disease and are believed to make up more than two-thirds of the human genome. They also have a strong association with genomic CNVs [97]. Long interspersed nuclear elements (LINEs) and *Alu* repeats (which are types of retrotransposons) have been associated with increased genomic instability through non-allelic homologous recombination events and can lead to pathogenic duplications and deletions [98]. *Alu–Alu* repeat recombinations within the introns of *ALDH7A1* have been associated with pyroxidine-dependent epilepsy [99]. The ability to accurately detect repetitive sequences is of great importance due to the problems they can cause during the aligning or assembling of sequence reads [100], and the human genome is commonly analysed for repeats using Repbase annotation [101] and computational algorithms, such as the hidden Markov model (HMM)-derived database Dfam [102].

Genome annotation

The ability to understand the function of a gene and how variation might affect its function is dependent upon understanding its structure, which can be elucidated by genome annotation. Genome annotation in its simplest form proceeds by ab initio gene prediction algorithms that search a genome for putative gene structures [103–105] such as signals associated with transcription, protein-coding potential and splicing [106]. Although these gene-prediction algorithms were used in the early analysis of the human genome [107, 108], they are limited in both accuracy and coverage [29]. The current automated gene-annotation tools, such as Ensembl, provide fast computational annotation of eukaryotic genomes using evidence derived from known mRNA [109], RNA-Seq data [110] and protein sequence databases [111].

Computational annotation systems are essential for providing an overview of gene content in newly sequenced genomes and those with fewer resources assigned to annotation, yet manual annotation is still regarded as the 'gold standard' for accurate and comprehensive annotation (Table 3) [112]. As part of the ENCODE project, which was established to investigate all functional elements in the human genome [113], a genome-annotation assessment project was developed to assess the accuracy of computational gene annotation compared with a manually annotated test-set produced by the Human and Vertebrate Analysis and Annotation (HAVANA) team [29]. Although the best computational methods identified ~70% of the manually annotated loci, prediction of alternatively spliced transcript models was significantly less accurate, with the best methods achieving a sensitivity of 40–45%. Conversely, 3.2% of transcripts only predicted by computational methods were experimentally validated.

Only two groups, HAVANA and Reference Sequence (RefSeq) [30], produce genome-wide manual transcript annotation. The HAVANA team is based at the Wellcome Trust Sanger Institute, UK, and provides manual gene and transcript annotation for high-quality, fully finished 'reference' genomes, such as that of human [3]. HAVANA manual annotation is supported by computational and wet lab groups who, through their predictions, highlight regions of interest in the genome to be followed up by manual annotation, identify potential features missing from annotation and experimentally validate the annotated transcripts, then provide feedback to computational groups to help improve the analysis pipelines.

The RefSeq collection of transcripts and their associated protein products is manually annotated at the National Center for Biotechnology Information (NCBI) in the USA.

Table 3 Comparison of computationally derived annotation versus manually derived annotation

Annotation procedure	Automatic annotation—for example, Ensembl	Manual annotation—for example, HAVANA
Genome analysis	Very quick	Very slow and labour intensive
Annotation consistency	Consistent	Risk of subjectivity—achieving consistency requires careful training and monitoring
Sequence quality	Flexible; can use unfinished, short-read NGS sequence, shotgun assembly	Best results on high-quality sequence, but can offer great insight into lower-quality assembly
Functional annotation	Limited, lacking comprehensive detail of manual annotation—frequently misassign related sequences—i.e. protein-coding loci and pseudogenes	Extensive use of biotypes, such as coding, pseudogene, lncRNA, NMD, etc.
Complex genomic regions	Limited in ability to represent complex structures and other nonstandard features	Superior representation and resolution of gene families and able to define CDS regions of complicated gene structures
Gene annotation	Many false-positive and false-negative calls at locus level in all gene biotypes	Better coverage of loci and alternatively spliced transcripts
Pseudogenes	Limited	Able to predict pseudogenes and differentiate from genuine coding genes
Poly(A) features	Limited	Annotates poly(A) features
Flexibility	Error prone, forces problems such as non-canonical splicing and can only look at sequences more or less in isolation	Deals with inconsistencies in data, consults literature and other databases, can compare paralogues and orthologues and rapidly integrate new sequencing technologies

CDS coding sequence, *HAVANA* Human and Vertebrate Analysis and Annotation, *lncRNA* long non-coding RNA, *NGS* next-generation sequencing, *NMD* nonsense-mediated decay

Although many RefSeq transcripts are completely manually annotated, a significant proportion are not: for example in NCBI Homo sapiens Annotation Release 106, approximately 45% of transcripts were classified as being computationally annotated [114]. Furthermore, unlike HAVANA transcripts, which are annotated on the genome, RefSeq transcripts are annotated independently of the genome and based upon the mRNA sequence alone, which can lead to difficulty mapping to the genome.

The GENCODE [58] gene set takes advantage of the benefits of both manual annotation from HAVANA and automated annotation from the Ensembl gene build pipeline by combining the two into one dataset. GENCODE describes four primary gene functional categories, or biotypes: protein-coding gene, pseudogene, lncRNA and sRNA. The adoption of further biotypes, at both the gene level and transcript level, has enriched annotation greatly (Table 2). The final gene set is overwhelmingly manually annotated (~100% of all protein-coding loci and ~95% of all transcripts at protein-coding genes are manually annotated). Computational annotation predictions of gene features are provided to give hints to manual annotators and direct attention to unannotated probable gene features, and are also used to quality control (QC) manual annotation to identify and allow correction of both false-positive and false-negative errors.

GENCODE and RefSeq collaborate to identify agreed CDSs in protein-coding genes and to try and reach agreement where there are differences as part of the collaborative Consensus CoDing Sequence (CCDS) project [115, 116]. These CDS models, which do not include 5′ or 3′ UTRs, are frequently used in exome panels alongside the full RefSeq and GENCODE gene sets that form the majority of the target sequences in exome panels.

The GENCODE gene set improves on the CCDS set as it is enriched with additional alternatively spliced transcripts at protein-coding genes as well as pseudogene and lncRNA annotation, and as such is the most detailed gene set [117]. GENCODE is now incorporated into the two most widely used commercial WES kits [118, 119], with fewer variants of potential medical importance missed [120].

To present genome annotation in a meaningful and useful manner, publicly available, web-based interfaces for viewing annotation have been provided—for example, the Ensembl Genome Browser [71] and the UCSC browser [121] (Fig. 7), both of which display the GENCODE models. The GENCODE genes are updated twice a year, whereas CCDS is updated at least once a year. All transcripts are assigned a unique stable identifier, which only changes if the structure of the transcript changes, making the temporal tracking of sequences easy.

A great deal of functionality is provided by genome browsers, such as: displaying and interrogating genome information by means of a graphical interface, which is integrated with other related biological databases; identifying sequence variation and its predicted consequence using VEP; investigating phenotype information and tissue-specific gene expression; and searching for related sequences in the genome using BLAST. Figure 7 presents by way of example the gene *KCNT1*, which is associated with early infantile

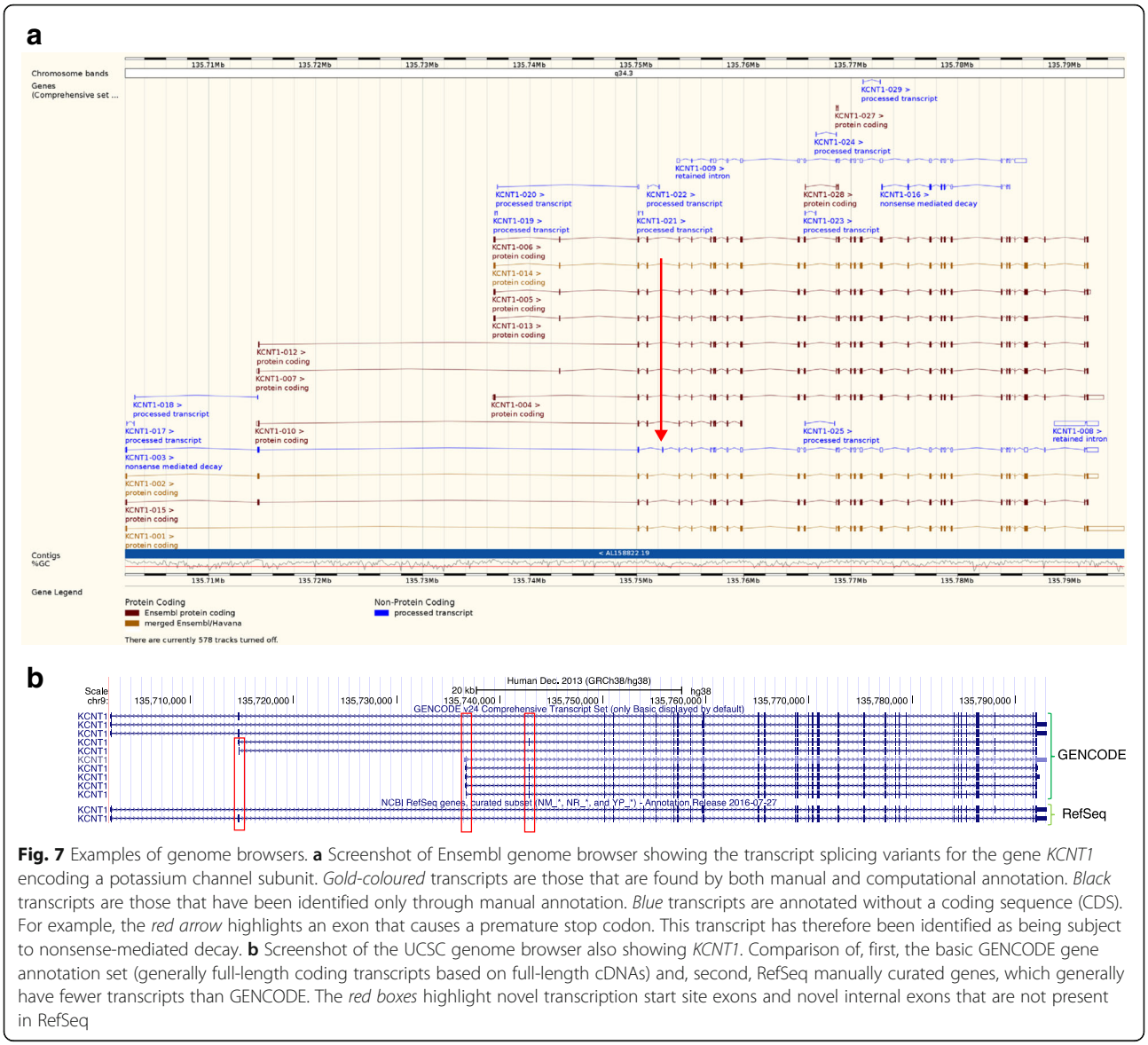

Fig. 7 Examples of genome browsers. **a** Screenshot of Ensembl genome browser showing the transcript splicing variants for the gene *KCNT1* encoding a potassium channel subunit. *Gold-coloured* transcripts are those that are found by both manual and computational annotation. *Black* transcripts are those that have been identified only through manual annotation. *Blue* transcripts are annotated without a coding sequence (CDS). For example, the *red arrow* highlights an exon that causes a premature stop codon. This transcript has therefore been identified as being subject to nonsense-mediated decay. **b** Screenshot of the UCSC genome browser also showing *KCNT1*. Comparison of, first, the basic GENCODE gene annotation set (generally full-length coding transcripts based on full-length cDNAs) and, second, RefSeq manually curated genes, which generally have fewer transcripts than GENCODE. The *red boxes* highlight novel transcription start site exons and novel internal exons that are not present in RefSeq

epileptic encephalopathies [122] displayed in both the Ensembl and UCSC genome browsers.

Using comparative genomics to confirm gene functionality

Sequence data from other organisms are essential for interpreting the human genome owing to the functional conservation of important sequences in evolution [123] that can then be identified by their similarity [124]. The zebrafish, for example, has a high genetic and physiological homology to human, with approximately 70% of human genes having at least one zebrafish orthologue. This means that the zebrafish model can provide independent verification of a gene being involved in human disease. Zebrafish also develop very quickly and are transparent, and so the fate, role and life cycle of individual cells can be followed easily in the developing organism. This makes the zebrafish a highly popular vertebrate model organism with which to study complex brain disorders [125, 126], and it has been essential for modelling disease in the DDD study [127].

Likewise, owing to a combination of experimental accessibility and ethical concerns, the mouse is often used as a proxy with which to study human disease [128, 129], and this justified the production of a high-quality, finished, reference mouse genome sequence, similar to that of the human sequence [130]. Murine behavioural traits, tissues, physiology and organ systems are all extremely similar to those of human [131], and their genomes are similar too, with 281 homologous blocks of at least 1 Mb [132] and over 16,000 mouse protein-coding genes with a one-to-one orthology to human [133]. The large number of

knockout mouse models available can be used to study many neurological diseases in patients [128], such as the Q54 transgenic mouse used to study *Scn2A* seizure disorders [134]. Recent studies in rodent models of epilepsy have identified changes in miRNA levels in neural tissues after seizures, which suggests that they could be key regulatory mechanisms and therapeutic targets in epilepsy [135]. It is therefore important that high-quality annotation for these model organisms is maintained, so that genes and transcripts can be compared across these organisms consistently [136]. With the advent of CRISPR–Cas9 technology, it is now possible to engineer specific changes into model organism genomes to assess the effects of such changes on gene function [137].

Nevertheless, model organism genomes and human genomes differ. For example, the laboratory mouse is highly inbred, whereas the human population is much more heterogeneous [138]. Furthermore, many environmental and behavioral components are known to affect disease in certain mouse strains, which are factors that are not clearly understood in human disease [139]. Although comparative genomics helps to build good gene models in the human genome and understand gene function and disease, basing predictions in clinical practice upon animal models alone might lead to misdiagnosis.

New techniques to improve functional annotation of genomic variants

NGS technologies facilitate improvements in gene annotation that have the potential to improve the functional annotation and interpretation of genomic variants. The combination of both long and short NGS reads [140] will change the scope of annotation. While short-read RNA-Seq assays may be able to produce hundreds of millions of reads and quantify gene expression, they are generally unable to represent full-length transcripts, which makes the assembly of such transcripts incredibly difficult [141]. However, the greater read lengths produced by new sequencing technologies such as PacBio and synthetic long-read RNA-Seq (SLR-Seq), which uses Illumina short-read sequencing on single molecules of mRNA, have the potential to produce sequence for complete transcripts in a single read. In addition, utilizing longer-read technologies such as that from PacBio has already been shown to improve resolution of regions of the genome with SVs [142], and emerging technologies, such as 10X genomics [143], promise further improvements. This is especially important because WES is unable to represent structural variation reliably. The importance of representing such regions through WGS has been demonstrated by numerous neurological diseases associated with SVs, including cases of severe intellectual disability [144]. Other examples of SV-induced neurological disease include Charcot–Marie–Tooth disease, which is most commonly caused by gene-dosage effects as a result of a duplication on the short arm of chromosome 17 [145], although other causes are known [146]; Smith–Magenis syndrome, caused by copy-number variants on chromosome 17p12 and 17p11.2 [147]; and Williams–Beuren syndrome, caused by a hemizygous microdeletion involving up to 28 genes on chromosome 7q11.23 [148].

Together, NGS data will also lead to the discovery of new exons and splice sites that both extend and truncate exons in a greater diversity of tissues and cell types. Whether the variants identified that are associated with novel exons or splice sites belong to protein-coding transcripts, or potential regulatory transcripts, or are transcripts likely to be targets of the NMD pathway, such technologies will permit better functional annotation of these overlapping variants. An example is the re-annotation of variants that were previously called intronic as exonic sequences. Similarly, a previously described synonymous substitution, or benign non-synonymous substitution, could affect core splice-site bases of a novel splice junction. RNA-Seq assays are able to discern expression of individual exons, allowing prioritisation of variants expressed in appropriate tissues for a disease. In the future, clinical investigation could target the genome in conjunction with the transcriptome—for example, using patient tissue as the basis for RNA-Seq assays—to identify regions where genes are expressed irregularly.

Transcriptomics datasets, such as CAGE [33], RAMPAGE [149] and polyA-seq [150], aid the accurate identification of the 5′ (for the two former) and 3′ (for the latter) ends of transcripts. This knowledge allows researchers to better annotate the functionality of a biotype, specifically enabling the addition of CDS where this was not previously possible, and enriching the functional annotation of overlapping variants. Furthermore, knowledge of termini allows the confident annotation of 5′ and 3′ UTRs that could harbor important regulatory sequences such as uORFs and miRNA target sites.

Other datasets, such as mass spectrometry (MS) [151] and ribosome profiling (RP, or Riboseq) [152], indicate translation, either by directly identifying proteins (MS) or by identifying translation on the basis of ribosomal binding to mRNA transcripts (RP), which aids the accurate identification of the presence and extent of expression of the CDS. Combining these datasets with cross-species conservation of protein coding potential found by PhyloCSF [153] allows annotators to identify previously unannotated protein-coding loci and confirm lncRNAs as lacking in protein-coding potential.

With the increasing importance of epigenetics and its role in neurological disorders [154], such as epilepsy [155], several companies are making detection of these features a priority—for example, detecting methylated

nucleotides directly, as part of their sequencing reaction [156]. Other well-described genetic marks are the DNase hypersensitivity sites that are often found in regions of active transcription [63]. However, before these marks are considered in the process of annotation, we will require better experimental datasets that validate them. To put such marks into context and aid validation, gene annotation must be as accurate and comprehensive as possible so that potential *cis* (local) and *trans* (distant) interactions can be identified. Regulatory regions such as enhancers are features that can be described as part of the extended gene and represent the next frontier for gene annotation using data such as Capture Hi-C [157] and ChIA-PET [158] to identify physical connections between regulatory regions affected by variation and the genes they regulate, which can often be located a great distance away. This could mean that variants that were previously considered to be benign could in future be reclassified as pathogenic. For example, variants in evolutionarily conserved transcription factor binding sites are believed to have a role in narcolepsy [159].

Computational and manual genome-annotation methods that have been described have relied almost exclusively on traditional transcriptional evidence to build or extend models of genes and their transcripts. While the number of sequences in public databases continues to increase, genes expressed at very low levels, or with restricted expression profiles (such as many non-coding loci), are likely to remain either under-represented or incomplete when relying on such evidence [160, 161].

New technologies and software will help assess the complexity of loci much more thoroughly through the investigation of alternative splicing/translation start sites/poly(A) sites [162], alternative open reading frames, and so on. They will also allow the revisiting of the human genome—for example, to investigate evolutionarily conserved regions and regulatory features for functionality and to identify new non-coding loci structures as well as new coding transcripts.

Conclusions

We have reviewed how important regions of the genome that harbor pathogenic sequence variation can lie outside the CDS of genes. We have discussed how researchers can better understand why an incorrect interpretation of a pathogenic variant could arise. Such reasons can range from the human reference genome being incomplete, not all exons being represented in public databases, to incorrect annotation of transcripts/exons owing to their expression in a different tissue or at a different developmental stage to the disease phenotype. Table 4 gives a summary of such examples. As such, considerable efforts continue to be made to increase the catalogue of new genes involved in diseases, such as neurological disease [127]. However, even well-studied genes should be revisited iteratively to identify novel features that previous technology could not detect. For example, a recent publication by Djemie and colleagues [163] revisited patients who had presented with Dravet syndrome, typically associated with *SCN1A* variants, but had been *SCN1A* variant-negative after clinical sequencing. By re-testing with NGS, it was possible to identify 28 variants that were overlooked with Sanger sequencing. Around 66% of the reported false-negative results were attributed to human error, whereas many of the others were a result of poor base-calling software [164].

It is important to remember that the full human transcriptome has yet to be annotated across all tissues of the human genome. Clearly, while gene panels and whole-exome sequences are a great start to getting a diagnosis, they are not perfect as they are snapshots of sequence at a particular point in time, meaning that pathogenic sequence variants that lie in yet-to-be-annotated exons will not be detected. This emphasizes the power of whole-genome sequences as, unlike exomes, they can be re-analysed again at any point in the future as new gene structures are found [165]. To identify such features, it will be important to update the annotation of disease genes using the most relevant experimental methods and tissue to help identify transcripts that might be expressed at low levels or only at certain developmental stages.

Similarly, improvements in the understanding and annotation of gene structures can lead to reclassification of variants as less pathogenic than previously believed, with implications for treatment strategies. For example, de la Hoya and colleagues demonstrated that improvements to understanding of native alternative splicing events in the breast cancer susceptibility gene *BRCA1* show that the risk of developing cancer is unlikely to be increased for carriers of truncating variants in exons 9 and 10, or indeed other alleles that retain 20–30% tumour-suppressor function, even where such variants had been previously characterized as pathogenic [166].

Accordingly, it is essential to consider multiple transcripts for pathogenic variant discovery, unlike the standard clinical approach of only considering a 'canonical' transcript, invariably based on the longest CDS but not necessarily on any expression values [167]. Such situations could result in ambiguous HGVS nomenclature when transcript IDs are not specified, and, as a result, important variants might be missed if variant analysis is only performed against the canonical transcript. For example, a variant can be classified as intronic based on the canonical transcript but could be exonic when based upon an alternatively spliced transcript. Such technical challenges illustrate the difficulties for clinicians when dealing with clinical reports containing details of identified variants (for example, HGVS identifiers) and attempting to map them accurately to function and allow variant interpretation.

Table 4 Important areas to consider for genome annotation

Genome assembly is not complete	Human assembly is still not complete and still being refined The current assembly is GRCh38, which still contains fragmented genes, and gene duplications are incorrectly represented, yet most analysis is still performed on GRCh37
Transcriptome is still incomplete	Some exons are still not represented in the human genome owing to low expression or temporal expression in tissue that has not yet been interrogated WES kits will not contain all exons WGS-negative cases should be iteratively re-analysed as new transcriptional features are revealed
Reference annotation datasets can be missing key features	Automatic annotation is fast but not as accurate as manual annotation CCDS—missing UTRs LRG—single, usually canonical, transcript—potential for missing exons; choice of transcript is arbitrary RefSeq—based on transcriptome, potential for missing exons and problems with inconsistent mapping to reference assembly Annotation does not necessarily determine which transcripts are the most likely to be functional, and the longest one might not be the major one
Non-coding genome	Long-range gene interactions are poorly understood; methods such as Capture Hi-C will provide insights into such epigenetics Previously ignored transcript biotypes such as NMD and retained intron are now known to have important regulatory roles in disease Non-coding RNAs have an important role in disease, yet they are hard to predict and their function remains largely unknown.
Biotype associations	A biotype conflict in annotation datasets will cause incorrect variant calls—for example, lncRNA variant compared with coding gene, coding gene compared with pseudogene
Transcript expression profile	Is transcript expressed in correct tissue for disease phenotype? Is transcript expressed at the right developmental time for disease phenotype?

CCDS Collaborative Consensus Coding Sequence project, *lncRNA* long non-coding RNA, *LRG* Locus Reference Genomic project, *NMD* nonsense-mediated decay, *WES* whole-exome sequencing, *WGS* whole-genome sequencing

A solution to this problem would be to identify all the high-confidence transcripts and call variants against these transcripts, highlighting variants that might have severe effects against one or more such transcripts. To improve sensitivity, these findings could be weighted by transcript expression level in the disease-relevant tissue(s) (Fig. 8). To improve sensitivity even further, RNA-Seq assays from different developmental stages could be interrogated to see whether exons are expressed at the correct developmental stage as that of the disease phenotype [63].

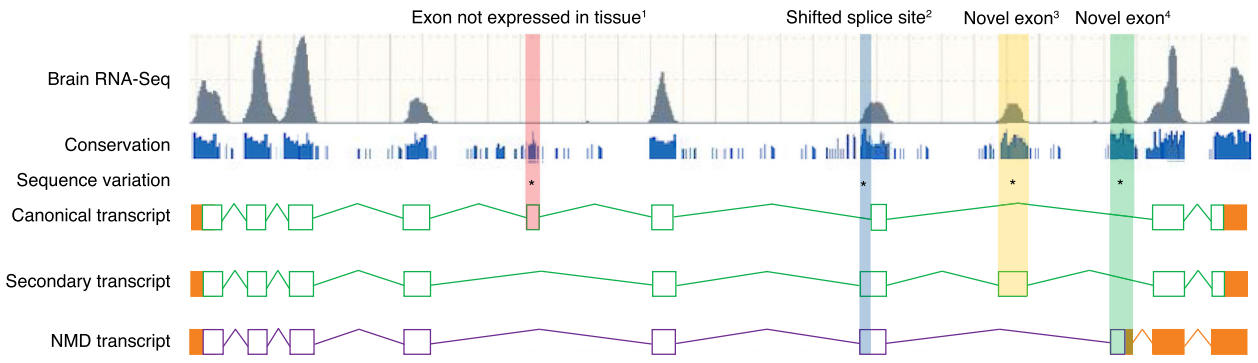

Fig. 8 The importance of multiple alternative transcripts for variant interpretation. This hypothetical example of gene '*AGENE*' expressed in brain highlights how the same variant could have different outcomes in different transcripts. We illustrate this further using hypothetical HGVS nomenclature. Note that when there are multiple transcripts for a gene, this can have an effect on amino acid numbering of variants as different transcripts can have different exon combinations, meaning that the same exon in two different transcripts can have a different translation and can also result in different lengths for the amino acid sequence. Note too that the untranslated region is represented by *orange boxes*. *Green boxes* represent the coding sequence (CDS), whereas *purple boxes* represent the CDS of the nonsense-mediated decay (*NMD*) transcript. *Lines* that join exons represent introns. *Asterisks* indicate the positions of the following hypothetical variants. (1) NM_000000001.99(AGENE):c.2041C > T (p.Arg681Ter). This variant might not be of interest to the clinician as it lies in an exon that is not expressed in brain. (2) NM_000000002.99(AGENE):c.4002 + 2451G > C. The Human Genome Variation Society (HGVS) suggests that this variant is intronic, yet, by looking across other transcripts, it is clear that the variant falls in an extended coding exon that is expressed in brain. (3) NC_000000003.99:g.66178947G > T. This variant is intronic to the canonical transcript, but falls in a well-conserved exon that is expressed in brain. (4) ENSP0000000004.1(AGENE):p.Gly276Ala. This variant falls in an exon that induces NMD. The exon is well conserved and expressed in the brain, making it potentially relevant to the clinician. Generally, NMD transcripts have been considered to be non-coding and excluded from sequence analysis. However, such exons are now known to have an important role in gene regulation. For example, Lynch and colleagues [194] reported that variation in the highly conserved exon in *SNRPB* that induces NMD can result in severe developmental disorders

Also of interest and concern is where genes thought to be implicated in a specific disease are now thought to have insufficient evidence for their role in disease. For example, the following genes were previously thought to be associated with epilepsy: *EFHC1* [168], *SCN9A*, *CLCN2*, *GABRD*, *SRPX2* and *CACNA1H* [169]. The Epilepsy Genetics Initiative (EGI) attempts to address such problems by iteratively re-analysing WES and WGS of epilepsy cases every 6 months.

The overwhelming amount of sequence variation that is generated by WES and WGS means that many variants produced will have no role in disease. Therefore, the use of databases that contain sequence variants from global sequencing projects, such as ExAC [170] and the 1000 Genomes Project [171] can help filter out common variants to help identify rare variants [60, 172]. Such databases can be used to identify those genes that are intolerant of any variation in their sequence, and, when variants in such genes are identified in patients, this could be an indicator of pathogenic sequence variation [173]. Other variant databases, such as The Human Gene Mutation Database (HGMD) [174] and ClinVar [175], provide information on inherited disease variants and on relationships between variants and phenotype. Genomic interpretation companies are now providing increasingly quick pathogenic variant interpretation turnaround times [176–179]. However, the value of such interpretation will only be as good as the gene annotation that is used for genome analysis and interpretation, demonstrating the need for continual updating and improvement of current gene sets.

Genome annotation is also increasingly seen as essential for the development of pharmacological interventions, such as drug design. Typically, drug design targets the main transcript of a gene (the choice of such a transcript is not necessarily informed by biological data, but is generally based upon the longest transcript), yet, as mentioned previously, it is now understood that certain transcripts can be expressed in different tissues, or at certain developmental times [180]. For example, the onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, and its activity is inhibited by paraneoplastic antibodies. It is encoded by *NOVA1*, which is only expressed in neurons [181]. The alternative splicing of exon 5 of the epilepsy-associated gene *SCN1A* generates isoforms of the voltage-gated sodium channel that differ in their sensitivity to the anti-epileptic medications phenytoin and lamotrigine [180]. Finally, isoform switching in the mouse gene *Dnm1* (encoding dynamin-1), as a result of alternative splicing of exon 10 during embryonic to postnatal development, causes epilepsy [182].

With new drugs having a high failure rate and associated financial implications [183–185], it is not unreasonable to suggest that identifying tissue-specific exons and transcripts through annotation has the potential to reduce such failure rates significantly. New methods of generating genomic data must therefore be adopted continually and interrogated by annotators to facilitate the translation of genomic techniques into the clinic in the form of genomic medicines.

Such advances will begin to address some of the controversies and challenges for clinicians that the fast advances in genomics bring. They will help to understand why current technology can fail to identify the pathogenic basis of a patient's disorder, or, more worryingly, why it can produce an incorrect result where the wrong variant is labelled as causative. This understanding will help clinicians to explain the advantages and limitations of genomics to families and healthcare professionals when caring for patients. The implication is that it will empower them to request reanalysis of unsolved cases as newer technology improves the annotation of gene structure and function. It will also encourage clinicians to request referral for disease modification when therapy becomes available for a clinical disease caused by specific genomic alterations.

Abbreviations

ACMG: American College of Medical Genetics and Genomics; CAGE: Cap-analysis gene expression; CCDS: Consensus coding sequence; CDS: Coding sequence; CNV: Copy-number variant; DDD: Deciphering Developmental Disorders; HAVANA: Human and Vertebrate Analysis and Annotation; HGP: Human Genome Project; HGVS: Human Genome Variation Society; indel: Insertion and deletion; lincRNA: Long-intergenic non-coding RNA; lncRNA: Long non-coding RNA; LoF: Loss-of-function; miRNA: MicroRNA; NCBI: National Center for Biotechnology Information; ncRNA: Non-coding RNA; NGS: Next-generation sequencing; NMD: Nonsense-mediated decay; ORF: Open reading frame; PacBio: Pacific Biosciences; RefSeq: Reference Sequence; RNA-Seq: RNA sequencing; sRNA: Small RNA; TSS: Transcription start site; UTR: Untranslated region; VEP: Variant effect predictor; WES: Whole-exome sequencing; WGS: Whole-genome sequencing

Acknowledgements

We thank Jane Rogers for her guidance, Eugene Bragin for his informatics input, and Imogen Steward, who is still awaiting her genetic diagnosis, but instrumental in the undertaking of this manuscript. We hope that this paper will support patients such as Imogen, now and in the future.

Funding

This work is funded by the National Institutes of Health grant U41HG007234 to the GENCODE project, and Wellcome Trust grant (WT098051) to the Sanger Institute and the European Molecular Biology Laboratory. Part of this work was undertaken at University College London Hospitals, which received a proportion of funding from the NIHR Biomedical Research Centres funding scheme. We are grateful for support from the Epilepsy Society.

Authors' contributions

All authors wrote and proof-read the manuscript and then read and approved the final manuscript. CAS managed the writing and organized the article and produced all figures and tables after mutual discussion with all authors. APJP, BAM and SMS provided clinical input and structure.

Competing interests

CAS is Technical Support Scientist at Congenica Ltd, a clinical interpretation company and one of the partners for the UK 100,000 Genomes Project. CAS's daughter has been diagnosed with West syndrome, an early infantile epileptic encephalopathy, and is currently on the 100,000 Genomes Project.

JH is Program Manager for Population Sequencing at Illumina Inc. The other authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Congenica Ltd, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK. [2]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [3]Addenbrooke's Hospital and University of Cambridge, Cambridge CB2 0QQ, UK. [4]Department of Pediatrics (Neurology), University of Texas Southwestern, Dallas, TX, USA. [5]Program in Genetics and Genome Biology and Department of Paediatrics (Neurology), The Hospital for Sick Children and University of Toronto, Toronto, Canada. [6]Department of Clinical and Experimental Epilepsy, UCL Institute of Neurology, London WC1N 3BG, UK. [7]Chalfont Centre for Epilepsy, Chesham Lane, Chalfont St Peter, Buckinghamshire SL9 0RJ, UK. [8]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [9]Illumina Inc, Great Chesterford, Essex CB10 1XL, UK.

Published online: 30 May 2017

References

1. EpiPM Consortium. A roadmap for precision medicine in the epilepsies. Lancet Neurol. 2015;14:1219–28.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. Erratum in: Nature. 2001;411:720. Szustakowki, J [corrected to Szustakowski, J]. Nature 2001 Aug 2;412(6846):565.
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.
4. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9:e1001091.
5. GENCODE. Human GENCODE version 24. 2016. http://www.gencodegenes.org/stats/current.html. Accessed 14 Feb 2017.
6. Ensembl. Ensembl Human, release 83, GRC38. 2016. http://www.ensembl.org/Homo_sapiens/Info/Annotation. Accessed 14 Feb 2017.
7. Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, et al. An SNP map of human chromosome 22. Nature. 2000;407:516–20.
8. Firth HV, Wright CF. The Deciphering Developmental Disorders (DDD) study. Dev Med Child Neurol. 2011;53:702–3.
9. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–8.
10. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74:5463–7.
11. Papandreou A, McTague A, Trump N, Ambegaonkar G, Ngoh A, Meyer E, et al. GABRB3 mutations: a new and emerging cause of early infantile epileptic encephalopathy. Dev Med Child Neurol. 2016;58:416–20.
12. Illumina. Illumina Inc. https://www.illumina.com/. Accessed 26 Apr 2017.
13. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.
14. McPherson JD. A defining decade in DNA sequencing. Nat Methods. 2014;110:1003–5.
15. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
16. 100K Genomes. Sequencing 100000 Genomes. 2014. http://www.genomicsengland.co.uk/. Accessed 14 Feb 2017.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
18. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
20. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38:1767–71.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
22. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
23. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;9:692–9.
24. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
25. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006;16:949–61.
26. Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. Predicting the functional consequences of non-synonymous DNA sequence variants–evaluation of bioinformatics tools and development of a consensus strategy. Genomics. 2013;102:223–8.
27. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–24.
28. HGVS. HGVS nomenclature. 2017. http://www.hgvs.org/mutnomen. Accessed 24 Apr 2017.
29. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, et al. EGASP: the human ENCODE genome annotation assessment project. Genome Biol. 2006;7 Suppl. 1:S2. 1–31.
30. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(Database issue):D756–63.
31. Bauters M, Frints SG, Van Esch H, Spruijt L, Baldewijns MM, de Die-Smulders CE, et al. Evidence for increased SOX3 dosage as a risk factor for X-linked hypopituitarism and neural tube defects. Am J Med Genet A. 2014;164A:1947–52.
32. Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, et al. Before it gets started: regulating translation at the 5′ UTR. Comp Funct Genomics. 2012;2012:475731.
33. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003;100:15776–81.
34. Parihar R, Ganesh S. The SCN1A gene variants and epileptic encephalopathies. J Hum Genet. 2013;58:573–80.
35. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 2000;10:1001–10.
36. Kang MK, Han SJ. Post-transcriptional and post-translational regulation during mouse oocyte maturation. BMB Rep. 2011;44:147–57.
37. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003;72:291–336.
38. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000;28:4364–75.
39. Gonzalez-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.
40. Jaffe AE, Shin J, Collado-Torres L, Leek JT, Tao R, Li C, et al. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. Nat Neurosci. 2015;18:154–61.
41. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA. 2008;14:802–13.
42. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 2013;27:2380–96.
43. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC, et al. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. Genome Res. 2013;23:812–25.
44. Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev. 2012;26:1209–23.
45. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24:1774–86.

46. Reimand J, Wagih O, Bader GD. Evolutionary constraint and disease associations of post-translational modification sites in human genomes. PLoS Genet. 2015;11:e1004919.
47. Cheng J, Maquat LE. Nonsense codons can reduce the abundance of nuclear mRNA without affecting the abundance of pre-mRNA or the half-life of cytoplasmic mRNA. Mol Cell Biol. 1993;13:1892–902.
48. Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci. 1998;23:198–9.
49. Zhao Y, Lin J, Xu B, Hu S, Zhang X, Wu L. MicroRNA-mediated repression of nonsense mRNAs. Elife. 2014;3:e03032.
50. Boutz PL, Bhutkar A, Sharp PA. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. 2015;29:63–80.
51. Nguyen LS, Jolly L, Shoubridge C, Chan WK, Huang L, Laumonnier F, et al. Transcriptome profiling of UPF3B/NMD-deficient lymphoblastoid cells from patients with various forms of intellectual disability. Mol Psychiatry. 2012;17:1103–15.
52. Adlakha YK, Saini N. Brain microRNAs and insights into biological functions and therapeutic potential of brain enriched miRNA-128. Mol Cancer. 2014;13:33.
53. Lin YS, Wang HY, Huang DF, Hsieh PF, Lin MY, Chou CH, et al. Neuronal splicing regulator RBFOX3 (NeuN) regulates adult hippocampal neurogenesis and synaptogenesis. PLoS One. 2016;11:e0164164.
54. Sundermeier T, Ge Z, Richards J, Dulebohn D, Karzai AW. Studying tmRNA-mediated surveillance and nonstop mRNA decay. Methods Enzymol. 2008; 447:329–58.
55. Shoemaker CJ, Green R. Translation drives mRNA quality control. Nat Struct Mol Biol. 2012;19:594–601.
56. Frankish A, Harrow J. GENCODE pseudogenes. Methods Mol Biol. 2014; 1167:129–55.
57. Vanin EF. Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 1985;19:253–72.
58. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.
59. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, et al. The GENCODE pseudogene resource. Genome Biol. 2012;13:R51.
60. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–8.
61. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299–320.
62. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32:246–51.
63. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
64. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465:1033–8.
65. Poliseno L, Haimovic A, Christos PJ, Vega Y Saenz de Miera EC, Shapiro R, Pavlick A, et al. Deletion of PTENP1 pseudogene in human melanoma. J Invest Dermatol. 2011;131:2497–500.
66. Yu G, Yao W, Gumireddy K, Li A, Wang J, Xiao W, et al. Pseudogene PTENP1 functions as a competing endogenous RNA to suppress clear-cell renal cell carcinoma progression. Mol Cancer Ther. 2014;13:3086–97.
67. GTEX. GTEX. 2017. http://www.gtexportal.org/. Accessed 24 Apr 2017.
68. Atlas. Expression Atlas. https://www.ebi.ac.uk/gxa/home. Accessed 12 Feb 2017.
69. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2011; 13:59–69.
70. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
71. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. 2017;45(D1):D635–42.
72. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
73. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342:1235587.
74. Smedley D, Schubach M, Jacobsen JO, Köhler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet. 2016;99:595–606.
75. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014; 15:423–37.
76. Barquist L, Burge SW, Gardner PP. Studying RNA homology and conservation with infernal: from single sequences to RNA families. Curr Protoc Bioinformatics. 2016;54:12.13.1–12.13.25.
77. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015; 43(Database issue):D130–7.
78. Ambros V. The functions of animal microRNAs. Nature. 2004;431:350–5.
79. Henshall DC. MicroRNA and epilepsy: profiling, functions and potential clinical applications. Curr Opin Neurol. 2014;27:199–205.
80. Ren L, Zhu R, Li X. Silencing miR-181a produces neuroprotection against hippocampus neuron cell apoptosis post-status epilepticus in a rat model and in children with temporal lobe epilepsy. Genet Mol Res. 2016;15(1); doi:10.4238/gmr.15017798.
81. Panjwani N, Wilson MD, Addis L, Crosbie J, Wirrell E, Auvin S, et al. A microRNA-328 binding site in PAX6 is associated with centrotemporal spikes of rolandic epilepsy. Ann Clin Transl Neurol. 2016;3:512–22.
82. Reschke CR, Silva LF, Norwood BA, Senthilkumar K, Morris G, Sanz-Rodriguez A, et al. Potent anti-seizure effects of locked nucleic acid antagomirs targeting miR-134 in multiple mouse and rat models of epilepsy. Mol Ther Nucleic Acids. 2017;6:45–56.
83. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013;154:26–46.
84. Wright MW. A short guide to long non-coding RNA gene nomenclature. Hum Genomics. 2014;8:7.
85. St Laurent G, Wahlestedt C, Kapranov P. The Landscape of long noncoding RNA classification. Trends Genet. 2015;31:239–51.
86. Nitsche A, Rose D, Fasold M, Reiche K, Stadler PF. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. RNA. 2015;21:801–12.
87. McHugh CA, Chen CK, Chow A, Surka CF, Tran C, McDonel P, et al. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature. 2015;521:232–6.
88. Liu Z, Sun M, Lu K, Liu J, Zhang M, Wu W, et al. The long noncoding RNA HOTAIR contributes to cisplatin resistance of human lung adenocarcinoma cells via downregualtion of p21(WAF1/CIP1) expression. PLoS One. 2013;8:e77293.
89. Zhang X, Weissman SM, Newburger PE. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. RNA Biol. 2014;11:777–87.
90. Lee DY, Moon J, Lee ST, Jung KH, Park DK, Yoo JS, et al. Dysregulation of long non-coding RNAs in mouse models of localization-related epilepsy. Biochem Biophys Res Commun. 2015;462:433–40.
91. Morris KV. The theory of RNA-mediated gene evolution. Epigenetics. 2015;10:1–5.
92. Vitiello M, Tuccoli A, Poliseno L. Long non-coding RNAs in cancer: implications for personalized therapy. Cell Oncol (Dordr). 2015;38:17–28.
93. Hsiao J, Yuan TY, Tsai MS, Lu CY, Lin YC, Lee ML, et al. Upregulation of haploinsufficient gene expression in the brain by targeting a long non-coding RNA improves seizure phenotype in a model of Dravet syndrome. EBioMedicine. 2016;9:257–77.
94. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102–10.
95. Talkowski ME, Maussion G, Crapper L, Rosenfeld JA, Blumenthal I, Hanscom C, et al. Disruption of a large intergenic noncoding RNA in subjects with neurodevelopmental disabilities. Am J Hum Genet. 2012;91:1128–34.
96. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am J Hum Genet. 2016;98:58–74.
97. Zhou W, Zhang F, Chen X, Shen Y, Lupski JR, Jin L. Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms. Hum Mol Genet. 2013; 22:2642–51.
98. Chen L, Zhou W, Zhang L, Zhang F. Genome architecture and its roles in human copy number variation. Genomics Inform. 2014;12:136–44.

99. Mefford HC, Zemel M, Geraghty E, Cook J, Clayton PT, Paul K, et al. Intragenic deletions of ALDH7A1 in pyridoxine-dependent epilepsy caused by Alu-Alu recombination. Neurology. 2015;85:756–62.
100. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7: e1002384.
101. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
102. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44(D1):D81–9.
103. Burge CB, Karlin S. Finding the genes in genomic DNA. Curr Opin Struct Biol. 1998;8:346–54.
104. Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. Genome Res. 2000;10:516–22.
105. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19 Suppl 2:ii215–25.
106. Mudge J, Harrow J. Methods for improving genome annotation. In: Alterovitz G, Ramoni MF, editors. Knowledge based bioinformatics: from analysis to interpretation. Chichester, West Sussex: John Wiley & Sons; 2010. p. 209–14.
107. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, et al. The DNA sequence of human chromosome 21. Nature. 2000;405:311–9. Erratum in: Nature. 2000;407:110.
108. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, et al. The DNA sequence of human chromosome 22. Nature. 1999;402:489–95. Erratum in: Nature. 2000;404:904.
109. Karsch-Mizrachi I, Nakamura Y, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2012;40(Database issue):D33–7.
110. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13:329–42.
111. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 2011;39(Database issue):D214–9.
112. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4. 1-9.
113. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447: 799–816.
114. Frankish A, Uszczynska B, Ritchie GR, Gonzalez JM, Pervouchine D, Petryszak R, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015;16 Suppl 8:S2.
115. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009; 19:1316–23.
116. Farrell CM, O'Leary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, et al. Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res. 2014;42(Database issue):D865–72.
117. Mudge JM, Frankish A, Harrow J. Functional transcriptomics in the post-ENCODE era. Genome Res. 2013;23:1961–73.
118. SeqCap. SeqCap EZ Human Exome Library v3.0. 2014. http://sequencing.roche.com/products/nimblegen-seqcap-target-enrichment/seqcap-ez-system/seqcap-ez-exome-v3.html. Accessed 12 Feb 2017.
119. Chen R, Im H, Snyder M. Whole-exome enrichment with the agilent sureselect human all exon platform. Cold Spring Harb Protoc. 2015;2015: 626–33.
120. Coffey AJ, Kokocinski F, Calafato MS, Scott CE, Palta P, Drury E, et al. The GENCODE exome: sequencing the complete human exome. Eur J Hum Genet. 2011;19:827–31.
121. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 2017; 45(D1):D626–34.
122. Barcia G, Fleming MR, Deligniere A, Gazula VR, Brown MR, Langouet M, et al. De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. Nat Genet. 2012;44:1255–9.
123. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13.
124. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
125. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503. Erratum in: Nature. 2014;505:248.
126. Kalueff AV, Stewart AM, Gerlai R. Zebrafish as an emerging model for studying complex brain disorders. Trends Pharmacol Sci. 2014;35:63–75.
127. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519:223–8.
128. Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011;474:337–42.
129. Steward CA, Gonzalez JM, Trevanion S, Sheppard D, Kerry G, Gilbert JG, et al. The non-obese diabetic mouse sequence, annotation and variation resource: an aid for investigating type 1 diabetes. Database (Oxford). 2013;2013:bat032.
130. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009;7:e1000112.
131. Hofker MH, Deursen JV. Transgenic mouse: methods and protocols. Methods in molecular biology. Totowa, NJ: Humana Press; 2003. p. 3741. xiii.
132. Pevzner P, Tesler G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A. 2003;100:7672–7.
133. MGI. MGI-Mouse Vertebrate Homology. 2017. http://www.informatics.jax.org/homology.shtml. Accessed 24 Apr 2017.
134. Kearney JA, Plummer NW, Smith MR, Kapur J, Cummins TR, Waxman SG, et al. A gain-of-function mutation in the sodium channel gene Scn2a results in seizures and behavioral abnormalities. Neuroscience. 2001;102:307–17.
135. Henshall DC, Hamer HM, Pasterkamp RJ, Goldstein DB, Kjems J, Prehn JH, et al. MicroRNAs in epilepsy: pathophysiology and clinical utility. Lancet Neurol. 2016;15:1368–76.
136. Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Group MGD. Mouse genome database 2016. Nucleic Acids Res. 2016;44(D1):D840–7.
137. Ma X, Chen C, Veevers J, Zhou X, Ross RS, Feng W, et al. CRISPR/Cas9-mediated gene manipulation to create single-amino-acid-substituted and floxed mice with a cloning-free method. Sci Rep. 2017;7:42244.
138. Leiter EH, von Herrath M. Animal models have little to teach us about type 1 diabetes: 2. In opposition to this proposal. Diabetologia. 2004;47:1657–60.
139. Roep BO, Atkinson M. Animal models have little to teach us about type 1 diabetes: 1. In support of this proposal. Diabetologia. 2004;47:1650–6.
140. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55:641–58.
141. Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10:1177–84.
142. Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, et al. Long-read sequence assembly of the gorilla genome. Science. 2016;352:aae0344.
143. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34:303–11.
144. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014;511:344–7.
145. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell. 1991;66:219–32.
146. Speevak MD, Farrell SA. Charcot-Marie-Tooth 1B caused by expansion of a familial myelin protein zero (MPZ) gene duplication. Eur J Med Genet. 2013; 56:566–9.
147. Yuan B, Neira J, Gu S, Harel T, Liu P, Briceño I, et al. Nonrecurrent PMP22-RAI1 contiguous gene deletions arise from replication-based mechanisms and result in Smith-Magenis syndrome with evident peripheral neuropathy. Hum Genet. 2016;135:1161–74.
148. Corley SM, Canales CP, Carmona-Mora P, Mendoza-Reinosa V, Beverdam A, Hardeman EC, et al. RNA-Seq analysis of Gtf2ird1 knockout epidermal tissue provides potential insights into molecular mechanisms underpinning Williams-Beuren syndrome. BMC Genomics. 2016;17:450.
149. Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 2013;23:169–80.

150. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, et al. A quantitative atlas of polyadenylation in five mammals. Genome Res. 2012;22:1173–83.
151. Zhang G, Annan RS, Carr SA, Neubert TA. Overview of peptide and protein analysis by mass spectrometry. Curr Protoc Protein Sci. 2010;Chapter 16:Unit16.1.
152. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 2014;15:205–13.
153. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011;27:i275–82.
154. Jakovcevski M, Akbarian S. Epigenetic mechanisms in neurological disease. Nat Med. 2012;18:1194–204.
155. Henshall DC, Kobow K. Epigenetics and epilepsy. Cold Spring Harb Perspect Med. 2015;5(12); doi:10.1101/cshperspect.a022731.
156. PacBio. Detecting DNA Base Modification. 2017. http://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf. Accessed 24 Apr 2017.
157. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47:598–606.
158. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem. 2009;107:30–9.
159. Guturu H, Chinchali S, Clarke SL, Bejerano G. Erosion of conserved binding sites in personal genomes points to medical histories. PLoS Comput Biol. 2016;12:e1004711.
160. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. PLoS Biol. 2011;9:e1000625. discussion e1001102.
161. Bussotti G, Leonardi T, Clark MB, Mercer TR, Crawford J, Malquori L, et al. Improved definition of the mouse transcriptome via targeted RNA sequencing. Genome Res. 2016;26:705–16.
162. Frankish A, Mudge JM, Thomas M, Harrow J. The importance of identifying alternative splicing in vertebrate genome annotation. Database (Oxford). 2012;2012:bas014.
163. Djemie T, Weckhuysen S, von Spiczak S, Carvill GL, Jaehn J, Anttonen AK, et al. Pitfalls in genetic testing: the story of missed SCN1A mutations. Mol Genet Genomic Med. 2016;4:457–64.
164. Mercimek-Mahmutoglu S, Patel J, Cordeiro D, Hewson S, Callen D, Donner EJ, et al. Diagnostic yield of genetic testing in epileptic encephalopathy in childhood. Epilepsia. 2015;56:707–16.
165. Foo JN, Liu JJ, Tan EK. Whole-genome and whole-exome sequencing in neurological diseases. Nat Rev Neurol. 2012;8:508–17.
166. de la Hoya M, Soukarieh O, López-Perolio I, Vega A, Walker LC, van Ierland Y, et al. Combined genetic and splicing analysis of BRCA1 c.[594-2A > C; 641A > G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. Hum Mol Genet. 2016;25:2256–68.
167. MacArthur JA, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res. 2014;42(Database issue):D873–8.
168. Subaran RL, Conte JM, Stewart WC, Greenberg DA. Pathogenic EFHC1 mutations are tolerated in healthy individuals dependent on reported ancestry. Epilepsia. 2015;56:188–94.
169. Helbig I, Tayoun AA. Understanding genotypes and phenotypes in epileptic encephalopathies. Mol Syndromol. 2016;7:172–81.
170. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
171. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
172. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508:469–76.
173. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.
174. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017. doi: 10.1007/s00439-017-1779-6
175. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.
176. Congenica. Congenica Ltd. 2017. https://www.congenica.com/. Accessed 24 Apr 2017.
177. Sophia-Genetics. Sophia Genetics. 2017. http://www.sophiagenetics.com/home.html. Accessed 24 Apr 2017.
178. WuXi. WuXi NextCODE. https://www.wuxinextcode.com/. Accessed 7 Apr 2017.
179. Omicia. Omicia 2016. http://www.omicia.com/. Accessed 24 Apr 2017.
180. Barrie ES, Smith RM, Sanford JC, Sadee W. mRNA transcript diversity creates new opportunities for pharmacological intervention. Mol Pharmacol. 2012;81:620–30.
181. Buckanovich RJ, Yang YY, Darnell RB. The onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, the activity of which is inhibited by paraneoplastic antibodies. J Neurosci. 1996;16:1114–22.
182. Boumil RM, Letts VA, Roberts MC, Lenz C, Mahaffey CL, Zhang ZW, et al. A missense mutation in a highly conserved alternate exon of dynamin-1 causes epilepsy in fitful mice. PLoS Genet. 2010;6. doi: 10.1371/journal.pgen.1001046
183. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9:203–14.
184. Arrowsmith J, Miller P. Trial watch: phase II and phase III attrition rates 2011–2012. Nat Rev Drug Discov. 2013;12:569.
185. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32:40–51.
186. Vengoechea J, Parikh AS, Zhang S, Tassone F. De novo microduplication of the FMR1 gene in a patient with developmental delay, epilepsy and hyperactivity. Eur J Hum Genet. 2012;20:1197–200.
187. Lemke JR, Lal D, Reinthaler EM, Steiner I, Nothnagel M, Alber M, et al. Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes. Nat Genet. 2013;45:1067–72.
188. Epi4K Consortium. De novo mutations in SLC1A2 and CACNA1A are important causes of epileptic encephalopathies. Am J Hum Genet. 2016;99:287–98.
189. Bilguvar K, Oztürk AK, Louvi A, Kwan KY, Choi M, Tatli B, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature. 2010;467:207–10.
190. Coutinho AM, Oliveira G, Katz C, Feng J, Yan J, Yang C, et al. MECP2 coding sequence and 3′UTR variation in 172 unrelated autistic patients. Am J Med Genet B Neuropsychiatr Genet. 2007;144B:475–83.
191. Combi R, Dalprà L, Ferini-Strambi L, Tenchini ML. Frontal lobe epilepsy and mutations of the corticotropin-releasing hormone gene. Ann Neurol. 2005;58:899–904.
192. Ramser J, Abidi FE, Burckle CA, Lenski C, Toriello H, Wen G, et al. A unique exonic splice enhancer mutation in a family with X-linked mental retardation and epilepsy points to a novel role of the renin receptor. Hum Mol Genet. 2005;14:1019–27.
193. Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet. 2013;14:496–506.
194. Lynch DC, Revil T, Schwartzentruber J, Bhoj EJ, Innes AM, Lamont RE, et al. Disrupted auto-regulation of the spliceosomal gene SNRPB causes cerebro-costo-mandibular syndrome. Nat Commun. 2014;5:4483.
195. Qureshi IA, Mehler MF. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nat Rev Neurosci. 2012;13:528–41.
196. GENCODE. GENCODE annotation biotypes. https://www.gencodegenes.org/gencode_biotypes.html. Accessed 24 Apr 2017.
197. Kozak M. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 1987;15:8125–48.
198. Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV, et al. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. 2011;39:4220–34.
199. Brenner S, Barnett L, Katz ER, Crick FH. UGA: a third nonsense triplet in the genetic code. Nature. 1967;213:449–50.
200. Venters BJ, Pugh BF. Genomic organization of human transcription initiation complexes. Nature. 2013;502:53–8.
201. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science. 1989;245:371–8.
202. Fatemi M, Pao MM, Jeong S, Gal-Yam EN, Egger G, Weisenberger DJ. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res. 2005;33:e176.
203. Down TA, Hubbard TJ. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 2002;12:458–61.
204. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6:386–98.