

A joint NCBI and EMBL-EBI transcript set for clinical genomics and research

<https://doi.org/10.1038/s41586-022-04558-8>

Received: 13 July 2021

Accepted: 7 February 2022

Published online: 6 April 2022

Open access

 Check for updates

Joannella Morales^{1,3}, Shashikant Pujar^{2,3}, Jane E. Loveland¹, Alex Astashyn², Ruth Bennett¹, Andrew Berry¹, Eric Cox², Claire Davidson¹, Olga Ermolaeva², Catherine M. Farrell², Reham Fatima¹, Laurent Gil¹, Tamara Goldfarb², Jose M. Gonzalez¹, Diana Haddad², Matthew Hardy¹, Toby Hunt¹, John Jackson², Vinita S. Joardar², Michael Kay¹, Vamsi K. Kodali², Kelly M. McGarvey², Aoife McMahon¹, Jonathan M. Mudge¹, Daniel N. Murphy¹, Michael R. Murphy², Bhanu Rajput², Sanjida H. Rangwala², Lillian D. Riddick², Françoise Thibaud-Nissen², Glen Threadgold¹, Anjana R. Vatsan², Craig Wallin², David Webb², Paul Flicek¹, Ewan Birney¹, Kim D. Pruitt², Adam Frankish¹, Fiona Cunningham¹ & Terence D. Murphy²✉

Comprehensive genome annotation is essential to understand the impact of clinically relevant variants. However, the absence of a standard for clinical reporting and browser display complicates the process of consistent interpretation and reporting. To address these challenges, Ensembl/Gencode¹ and RefSeq² launched a joint initiative, the Matched Annotation from NCBI and EMBL-EBI (MANE) collaboration, to converge on human gene and transcript annotation and to jointly define a high-value set of transcripts and corresponding proteins. Here, we describe the MANE transcript sets for use as universal standards for variant reporting and browser display. The MANE Select set identifies a representative transcript for each human protein-coding gene, whereas the MANE Plus Clinical set provides additional transcripts at loci where the Select transcripts alone are not sufficient to report all currently known clinical variants. Each MANE transcript represents an exact match between the exonic sequences of an Ensembl/Gencode transcript and its counterpart in RefSeq such that the identifiers can be used synonymously. We have now released MANE Select transcripts for 97% of human protein-coding genes, including all American College of Medical Genetics and Genomics Secondary Findings list v3.0 (ref. ³) genes. MANE transcripts are accessible from major genome browsers and key resources. Widespread adoption of these transcript sets will increase the consistency of reporting, facilitate the exchange of data regardless of the annotation source and help to streamline clinical interpretation.

For more than 20 years, the RefSeq and Ensembl/Gencode teams, the two major sources of human genome annotation, have provided high-quality reference gene and transcript sets. These resources are used widely for biological research and discovery, with the choice of set depending on the use case. For instance, RefSeq transcripts are typically used for variant submissions to ClinVar⁴ or for variant descriptions in publications. Conversely, large-scale research projects such as ENCODE⁵, gnomAD⁶, DECIPHER⁷ and GTEx⁸ use the Ensembl/Gencode set. Although both sets are supported by abundant evidence, the two are not identical owing to differences in curation timing, methodology and interpretation of evidence in data-poor genomic regions. Moreover, sequence differences are present because a few RefSeq transcripts do not perfectly match the reference genome sequence. No simple method has been developed thus far to determine end-to-end equivalence between entire transcripts from the two sources, and navigating these differences can therefore be challenging.

In the clinical context, no accepted standard reference sequence is available for reporting variants. Therefore, individuals or laboratories choose their own transcript, typically according to criteria such as transcript length or creation date. Additionally, resources and tools that are routinely consulted for clinical genomics often differ in their choice of preferred transcript. This can confound data interpretation and may cause errors in variant classification, potentially leading to real clinical harm. These challenges call for a transcript set that can be universally adopted across the clinical and research communities as a biologically informed standard reference for variant reporting to provide consistency across browser displays, resources and tools. Indeed, a 2018 survey⁹ conducted by Ensembl highlighted this need, with the majority of respondents expressing the desire for Ensembl/Gencode and RefSeq to agree on a primary transcript for each gene. The respondents included approximately 800 individuals, of whom around 35% were healthcare professionals or were working in clinical diagnostics.

¹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. ²National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. ³These authors contributed equally: Joannella Morales, Shashikant Pujar. [✉]e-mail: murphyte@ncbi.nlm.nih.gov

MANE collaboration

To meet community needs, we established the Matched Annotation from NCBI and EMBL-EBI (MANE) collaboration. The initial results of this effort are (1) the MANE Select transcript set, designed to include a single representative transcript for every protein-coding gene for clinical reporting and other applications, and (2) the MANE Plus Clinical set for genes at which the MANE Select transcript alone is inadequate for describing all publicly available pathogenic (P) variants. Key features of the MANE transcripts include end-to-end matching between the exons of Ensembl/Gencode and RefSeq transcript sequences, perfect alignment to the GRCh38 reference genome assembly¹⁰ (Discussion) and the use of biologically relevant criteria for transcript selection, such as transcript expression levels and conservation of the coding regions. Together, the two sets eliminate the need to choose between annotations when selecting a default transcript or when reporting variants. Access to the MANE data and detailed documentation on the MANE collaboration is available on the NCBI (<https://www.ncbi.nlm.nih.gov/refseq/MANE/>) and EMBL-EBI Transcript Archive (Tark; http://tark.ensembl.org/web/mane_project/) websites.

MANE Select

To build the MANE Select set, our joint approach involved designing independent pipelines that would each identify representative transcripts for protein-coding genes (Supplementary Methods 1 and Extended Data Fig. 1). We aimed to include all coding exons that are well expressed and show evidence of evolutionary conservation. We then developed a workflow to iteratively compare the pipeline outputs, identify transcript pairs with the same coding sequence (CDS) and exon structure, and standardize the transcript ends.

When using transcripts on GRCh38 available as of May 2018 in our initial comparison of the pipeline outputs, we determined that the Ensembl/Gencode and RefSeq selections were identical for only 14% of protein-coding genes. In particular, 73% had differences only in the untranslated regions (UTRs), either in the extent of the 5' or 3' end or in the choice of UTR exons, and 11% differed in the CDS. For the remaining 2% of genes, we observed other scenarios, such as a missing corresponding transcript in one source. For choices that differed in CDS or UTR exons, we iteratively resolved these differences through pipeline improvements, additional automated data analyses and manual curation following consensus curation guidelines (Supplementary Methods 2). Manual review was aided by quality assurance metrics that flagged discrepancies (Supplementary Table 2).

Owing to strong interest from the clinical community, we focused our manual curation efforts on a subset of clinically relevant genes ($n = 3,803$). The clinical relevance of these genes was validated by key clinical partners, including the Transforming Genomic Medicine Initiative (TGMI; <http://www.thetgmi.org>) and the Clinical Genome Resource (ClinGen)¹¹; alternatively, inclusion in repositories such as Genomics England (PanelApp¹²), Gene2Phenotype¹³, OMIM¹⁴ and ClinVar was used for validation. For genes in the American College of Medical Genetics and Genomics Secondary Findings list (ACMG SF v2.0; ref.¹⁵), we reviewed the suitability of the pipeline choice and discussed challenging genes with our clinical partners. Figure 1 illustrates the application of our key criteria, conservation and expression, when choosing a MANE Select transcript for two high-value clinical genes. As mentioned above, our goal was to select transcripts that include well-conserved and well-expressed protein-coding exons. When a coding exon did not meet either criterion (for example, in *MEN1*), a transcript excluding that exon was chosen as the MANE Select transcript. However, a coding exon displaying a signal of conservation was considered for inclusion if it passed our minimum expression threshold, even if this exon was expressed at lower levels than neighbouring exons (for example, in *TSC2*).

After selecting transcript pairs, with one transcript from each source, we determined the optimal 5' and 3' ends on the basis of the supporting evidence. We incorporated high-throughput datasets (described in Methods) to programmatically determine and automatically update the 5' and 3' ends of both the RefSeq and Ensembl/Gencode transcripts, even for some of the pairs that were initially found to be identical. Once updates were completed and perfect identity was achieved, both transcripts in the pair were tagged as MANE Select. Extended Data Fig. 2 illustrates our method to determine the transcription start site (TSS) for the MANE Select transcript of the gene *PTPRC* (HGNC:9666). Similar logic was used to compute poly(A) clusters to determine the 3' ends of transcripts. Additional details are provided in Supplementary Methods 3. The 5'-end updates of the transcripts resulted in an enrichment for motifs characteristic of eukaryotic transcription initiation¹⁶, including initiation at purines and the presence of properly positioned TATA box or initiator motifs for a subset of transcripts (Extended Data Fig. 3).

By June 2021 (MANE release v0.95), we had defined a MANE Select transcript for 97% (18,584) of protein-coding genes across the genome. This includes all ACMG SF v3.0 genes and more than 99% (3,793 of 3,803) of the subset of disease-associated genes (Extended Data Fig. 4). The outstanding clinical genes include those to be added in the next release (*KLK4*, *TOMT*) or those affected by errors in the GRCh38 chromosome sequences (*ABO*, *FUT3*, *MUC1*, *ORAI1*, *POLR2A*, *SHANK3*) or atypically complex annotation (*PEG10*). We aim to complete the set in early 2022. The vast majority of MANE Select transcripts will be stable. However, we will allow updates on the rare occasion that new data demonstrate, without ambiguity, that the MANE Select transcript requires an update or needs to be replaced with a better transcript.

MANE Plus Clinical

Although the MANE Select set serves as a variant reporting standard for the majority of genes, some clinically relevant genes require more than one transcript to report all known P or likely pathogenic (LP) variants if these variants map to alternatively spliced exons. For cases in which the MANE Select transcript alone is not sufficient to report all known variants, we defined an additional transcript: the MANE Plus Clinical transcript. After consultation with our clinical partners, we have released MANE Plus Clinical transcripts for 55 genes. Figure 2 illustrates the need for a second transcript to report the P and LP variants that map to mutually exclusive exons in the *SCN5A* (HGNC:10593) gene.

Updates to original transcript datasets

A crucial aspect of the MANE set is the fact that the Ensembl/Gencode and RefSeq transcripts (and, therefore, proteins) in a MANE pair are identical, and the identifiers can be used interchangeably. To achieve the perfect match, the vast majority of transcripts selected by both pipelines (94% for RefSeq and 94.1% for Ensembl/Gencode) underwent updates, resulting in version increments (Table 1). This includes some of the transcripts that were identical at the beginning of the project but did not conform to the UTR rules mentioned above. Most of the updates (86% for RefSeq and 88% for Ensembl/Gencode) were in the UTR. However, a small percentage of transcripts required changes to the CDS (1.8% for RefSeq and 1.5% for Ensembl/Gencode), which typically involved a change to the location of the start codon. In addition to these updates, new transcripts were created for existing annotations that were incomplete or inconsistent with the MANE criteria (2.4% for RefSeq and 1.6% for Ensembl/Gencode).

Comparison with alternative datasets

A goal of the MANE collaboration is to deliver a transcript set that can be widely adopted as a standard for reporting and for display across

Analysis

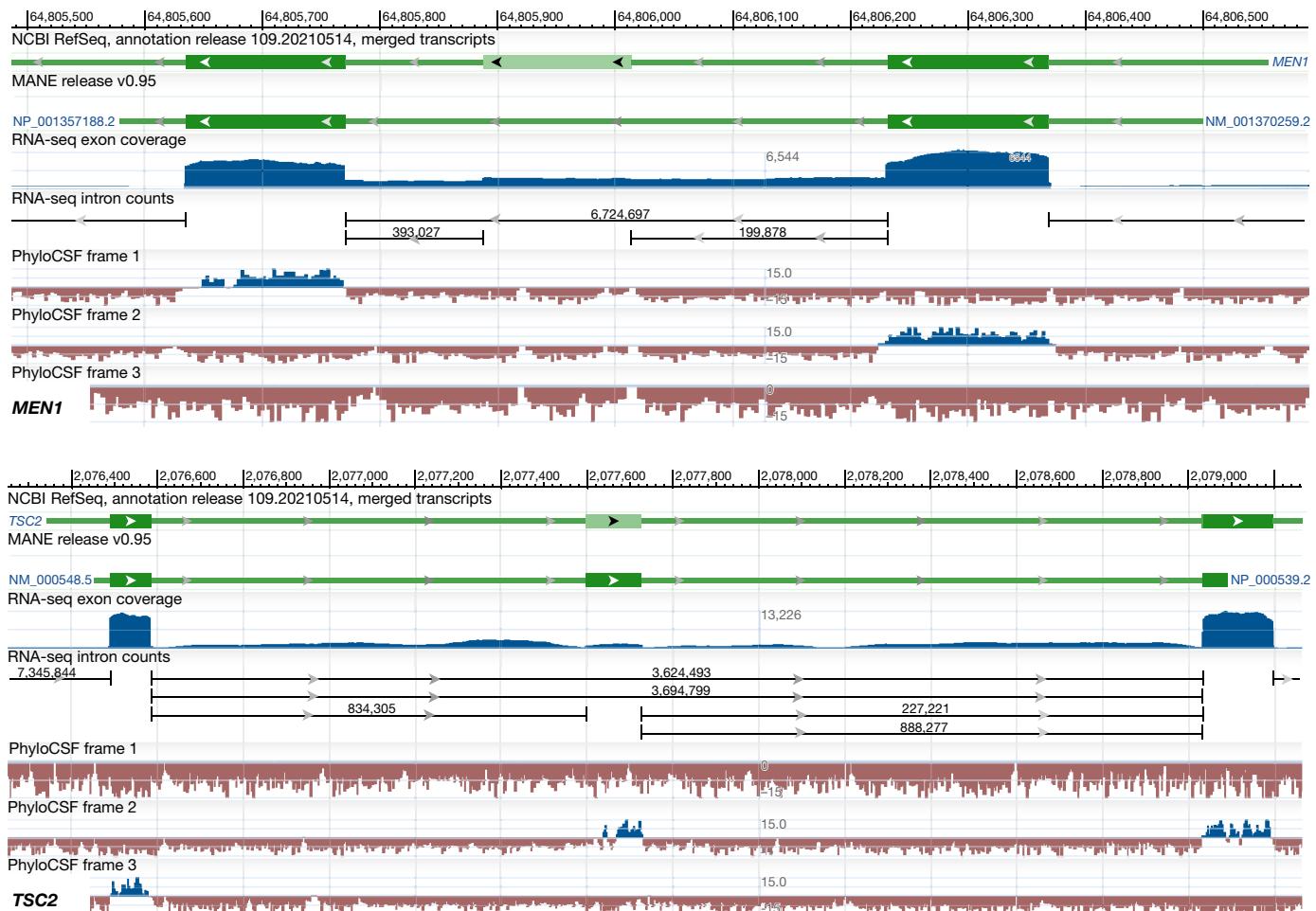


Fig. 1 | Conservation versus expression when manually curating two high-value clinical genes. Top, gene *MEN1* (HGNC:7010) tracks from NCBI GDV, as described below from top to bottom. Track 1, magnified region of the gene showing a portion of the CDS including an alternatively spliced exon (NCBI annotation release 109.20210514). Track 2, MANE v0.95 track showing the corresponding region of the MANE Select transcript (NM_001370259.2) lacking the alternatively spliced exon. Track 3, RNA-seq exon coverage (aggregate, filtered), with the numbers indicating the peak heights of the graph on a linear scale. Track 4, RNA-seq intron-spanning data from recount3, with horizontal lines depicting introns and numbers above the line indicating the number of reads. Track 5, PhyloCSF tracks. A transcript excluding the alternatively spliced exon was chosen as the MANE Select transcript owing to

low expression (tracks 3 and 4) and lack of evolutionary constraint (no positive PhyloCSF signal, as indicated by blue colour) for the alternatively spliced exon. Bottom, gene *TSC2* (HGNC:12363) tracks from GDV, as described below from top to bottom. Track 1, NCBI annotation release 109.20210514 track showing a portion of the coding region. Track 2, MANE v0.95 track showing the corresponding region of the MANE Select transcript (NM_000548.5). Track 3, RNA-seq exon coverage (aggregate, filtered). Track 4, portion of RNA-seq intron-spanning data from recount3. Track 5, PhyloCSF tracks. The MANE Select transcript includes the alternatively spliced protein-coding exon, which, despite its lower expression compared with neighbouring exons, shows evolutionary constraint of the CDS (presence of positive signal in the PhyloCSF track, as indicated by blue colour).

resources commonly used by the clinical and research communities. To assess the impact of our work, we aimed to quantify the overlap between the MANE Select set and two representative resources, gnomAD and ClinVar, which use Ensembl/GENCODE and RefSeq annotations, respectively. We chose to analyse disease-associated genes and these two resources because they reflect data used in the clinical community, represent orthogonal views of what users are exposed to or are using, and were available with sufficiently broad gene coverage to make the analysis informative. gnomAD shows Ensembl transcripts and could be perceived by users as a recommendation of a particular canonical transcript. The ClinVar submission data indicate which RefSeq transcripts are being used by submitters on the basis of unknown and likely varied criteria. The choice of datasets informed the set of genes included in the analysis. For the subset of manually curated disease-associated genes, we determined whether the canonical transcript in gnomAD (v3.1.1) and the transcript most commonly used for variant submission to ClinVar matched the MANE Select transcript accession. As shown in Fig. 3, the same accession as

that for MANE Select was used for 62.3% ($n = 2,945$) of the genes we reviewed. However, different accessions were used in one or both resources for the remaining 37.7% ($n = 1,779$) of genes (7.1% and 19.3% in ClinVar and gnomAD, respectively). This divergence demonstrates the consequence of having no standard transcript set and affirms the aims of our collaboration.

We collaborated with resources such as ExAC/gnomAD, ClinGen, ClinVar, DECIPHER and the Ensembl Variant Effect Predictor (VEP)¹⁷, all of which had different preferred transcripts, to encourage adoption of the MANE Select set, achieve standardization and ensure consistency. The interfaces of these resources now display the MANE Select transcript (Extended Data Fig. 5). In addition, UniProt is expected to update its browser in the near future to include flagged MANE Select proteins.

Access and display of MANE data

All data produced by the MANE collaboration are freely accessible in genome browsers, by bulk download and programmatically (see

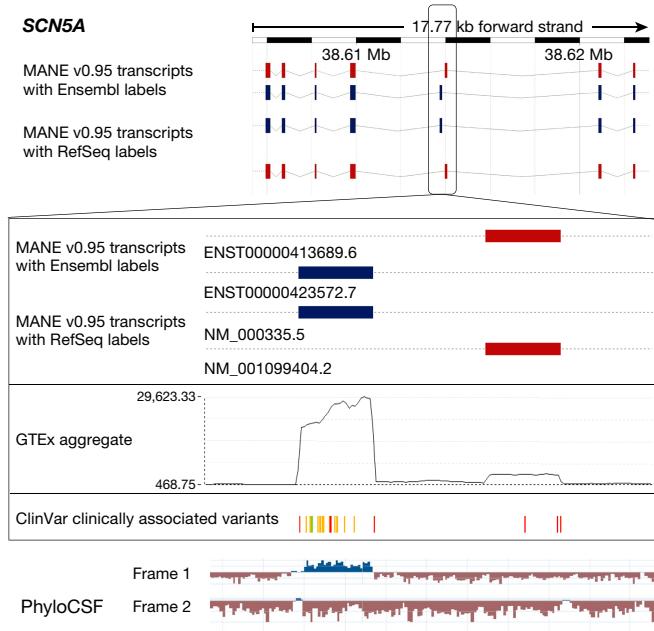


Fig. 2 | The need for a MANE Plus Clinical transcript for the SCN5A (HGNC:10593) gene. Top, Ensembl browser display of the SCN5A gene showing MANE Select (blue) and MANE Plus Clinical (red) transcripts (Ensembl/Gencode on top and RefSeq below) from MANE release v0.95. Bottom, magnified view of the portion of the gene that includes two mutually exclusive exons. The tracks are as described below, from top to bottom. Track 1, MANE v0.95 track showing the upstream MANE Select exon and downstream MANE Plus Clinical exon, shown in blue and red, respectively. Track 2, GTEx aggregate exon coverage (black wiggle plot). Track 3, ClinVar variants described as P or LP, coloured to indicate the type of variant (green, synonymous; yellow, missense; red, stop gained). Track 4, PhyloCSF tracks (one row for each frame) from NCBI GDV, with positive signal shown in blue.

links in Extended Data Table 1). A complete list of MANE transcripts with RefSeq and Ensembl identifiers in the latest MANE release (v0.95) is available in the MANE.GRCh38.v0.95.summary.txt.gz file on the FTP site (https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/release_0.95/) and Tark (<http://tark.ensembl.org/web/manelist>). As shown in Extended Data Fig. 6, the Ensembl browser displays MANE data using a custom-made track hub and labels the MANE transcripts in the transcript table within the gene-specific pages. The NCBI Genome Data Viewer (GDV)¹⁸ allows display of tracks for each MANE release and includes MANE tags in the RefSeq annotation (Extended Data Fig. 7). In addition, the University of California, Santa Cruz (UCSC) Genome Browser¹⁹ allows selection of a MANE data track in the Genes and Gene

Table 1 | Updates to RefSeq and Ensembl/Gencode transcripts.

Type of change	RefSeq	Ensembl/Gencode
No change (same exons, same CDS)	1,110 (6.0%)	1,094 (5.9%)
5'- or 3'-end change	15,968 (85.9%)	16,336 (87.9%)
New UTR, same CDS	724 (3.9%)	569 (3.1%)
Same exons, CDS changed	328 (1.8%)	285 (1.5%)
New CDS	454 (2.4%)	300 (1.6%)
Total	18,584	18,584

Comparison of RefSeq annotation release 109 (limited to NM_ transcripts) or Ensembl release 92 to MANE release v0.95.

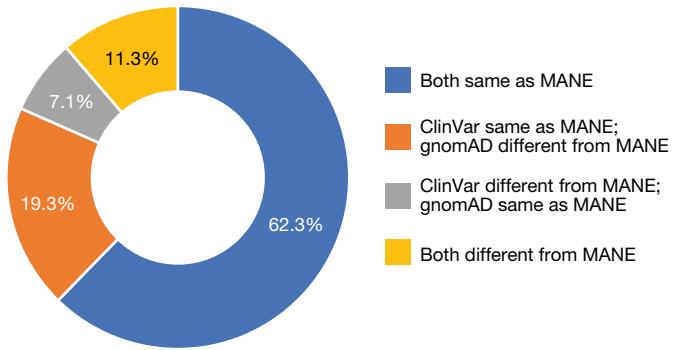


Fig. 3 | Comparison of the MANE Select dataset with gnomAD and ClinVar.

Doughnut chart showing a comparison of MANE Select transcripts with the most frequently used RefSeq transcript accession for variant submission in ClinVar and Ensembl canonical transcripts used for display in the gnomAD v3.1.1 resource.

Predictions section and exploration of the data in the Table Browser tool (Extended Data Fig. 8).

Discussion

RefSeq and Ensembl/Gencode have collaborated in the past to converge on annotation and provide joint, high-quality, evidence-based reference sets. We initiated the Consensus Coding Sequence (CCDS)²⁰ project in 2005 to provide transcript coding regions consistently annotated by the two groups. In 2008, we established the Locus Reference Genomic (LRG)²¹ project to provide stable reference sequences to report clinical variants. The MANE project goes beyond these collaborations in scope and content. It is not limited to coding regions, as in CCDS, but provides end-to-end matches between transcripts from the two sources. MANE is an improvement over LRG because, in addition to covering all protein-coding genes rather than a limited set of clinical genes, it provides transcript annotations that perfectly match the reference assembly. This is vital to reduce errors, considering that diagnostic pipelines now use whole-exome sequencing or whole-genome sequencing or will implement these methods in the near future. Therefore, NCBI and EMBL-EBI leaders of the LRG project decided to keep the LRG webpage and existing data available but have stopped expanding the LRG set. We recommend using the MANE transcript sets over those of LRG as a reference standard for clinical reporting. Existing LRG accessions now incorporate MANE transcript annotation (Extended Data Fig. 9) and will continue to be supported. Moreover, the Human Genome Variant Society (HGVS)²² now includes a recommendation to use MANE transcripts in its general and reference sequence guidelines.

Caveats and limitations

Selection of one transcript does not imply that the rich biology of the human genome can be reduced to one transcript at each locus, nor does it mean that transcripts not included in the MANE set are inferior or can be ignored. Even though the MANE set drives standardization for browser display and clinical reporting, we are not suggesting that only MANE transcripts be considered when analysing variants of potential clinical significance. For example, some disease mechanisms involve regulating expression in a tissue-specific manner or during a particular stage of development. This level of specificity and transcript diversity is not within the scope of the MANE Select set. Furthermore, when generating the MANE Plus Clinical set, we considered only P or LP exonic variants reported in ClinVar or other public resources. Given that not all laboratories make their variants freely accessible, our Plus Clinical set is a work in progress. We expect the set to increase as new variants are discovered and reported in

Analysis

public archives. Although this work has been driven by our annotation expertise, feedback from the community is encouraged. We will consider additional transcripts of clinical interest after consulting clinical experts. Enquiries about existing MANE transcripts and addition of new transcripts may be sent to mane-help@ncbi.nlm.nih.gov or mane-help@ebi.ac.uk.

The MANE sets are currently limited to protein-coding genes. We anticipate including well-supported non-coding genes in the future, particularly those with clinical relevance. In addition, a small percentage of protein-coding genes cannot currently be matched between RefSeq and Ensembl owing to errors in the GRCh38 primary reference assembly. We are collaborating with the Genome Reference Consortium (GRC) to generate patch sequences that correct errors or improve the assembly. GRC has indefinitely postponed the release of GRCh39; therefore, some protein-coding genes in MANE will have annotation on a patch. Additionally, mitochondrial genes and genes that undergo ribosomal slippage, such as *PEG10*, are presently not included in the MANE sets. However, we intend to include them in the future.

The MANE transcript sets are based on GRCh38 by design; thus, we plan to keep MANE matched only to GRCh38 for the foreseeable future to provide a unified stable clinical reporting standard. Most users are well served by a single reference genome assembly used uniformly across different resources. The most recent research data, analysis and annotation are available exclusively on GRCh38, which is supported in key clinical resources and tools such as gnomAD, ClinVar and DECI-PHER. Accordingly, RefSeq and GENCODE will continue using GRCh38 as the primary annotation reference for years to come. However, we recognize that many clinical laboratories will continue to use GRCh37 and that there is interest in new complete or nearly complete genome assemblies representing additional population diversity²³. Thus, RefSeq and Ensembl/GENCODE will develop further resources and tools to enable future pan-genomes and variation in other assemblies to be interpreted relative to MANE transcripts. For example, the RefSeq annotation of GRCh37, updated in March 2022, is available with markup for RefSeq Select transcripts, including those mapped to GRCh37 from MANE v0.95. A comparison of MANE transcripts to Ensembl/GENCODE annotation on GRCh37 is available at http://tark.ensembl.org/web/mane_GRCh37_list/. Mappings of MANE annotation from GRCh38 will be available on additional human assemblies in the future from both RefSeq and Ensembl/GENCODE. However, because MANE transcripts are planned to be generated only on GRCh38, those mapped to other assemblies may have sequence differences (for example, for 5% of genes in GRCh37), which need to be accounted for when generating HGVS expressions. We therefore recommend broad adoption of GRCh38 in the clinical community to take full advantage of MANE, improve consistency in variant identification and promote the exchange of clinical variant data. Failure to transition could cause discordance in variant identification²⁴, making variant interpretation vulnerable to outdated or incomplete genome annotation and severely limiting the exchange of clinical variant reports.

Future plans

We expect to finalize the MANE Select set in early 2022 (or finish as close to 100% of genes as possible given the limitations mentioned above) and to iteratively extend the MANE Plus Clinical set as new P variants are discovered. We are working with UniProt to align its set with MANE Select to provide access to a wealth of protein-based annotation in a consistent manner. Genes not currently in the MANE set include those for which the annotation differs between Ensembl/GENCODE and RefSeq owing to locus complexity and lack of evidence. In addition, genes needing genome patches and those annotated in only one of the two sets will be manually reviewed.

In the long term, we aim to produce a new set to include additional high-value transcripts, including those for the non-coding

genome, such as transcripts that carry exclusive, well-conserved exons that utilize alternative promoters or that have different termini. We will work on this set once we have mature workflows to integrate long transcriptomic data and data arising from rapid technical advances in the wider transcriptomics and proteomics fields. We are also considering the development of a set to label genes and transcripts relevant for human diseases. As a starting point, we plan to use the sets of genes defined by groups that are actively assessing gene–disease validity, such as the global Gene Curation Coalition (<https://thegencc.org/>).

In summary, as a result of our efforts to converge on the annotation of human protein-coding genes, our collaboration initiative between RefSeq and Ensembl/GENCODE delivers a joint transcript set to standardize clinical genomics and research. This set of one transcript per gene can be used as a default for tools and resources and as a reference set for clinical reporting and research. Universal adoption of this high-value set will promote consistency in reporting, limit clinical harm caused by errors in interpretation, increase the bidirectional exchange of data and help drive improvements in human health and diagnostics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-022-04558-8>.

- Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
- O’Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
- Miller, D. T. et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1381–1390 (2021).
- Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
- ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Firth, H. V. et al. DECI-PHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
- GTEX Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Morales, J. et al. The value of primary transcripts to the clinical and non-clinical genomics community: survey results and roadmap for improvements. *Mol. Genet. Genomic Med.* **9**, e1786 (2021).
- Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
- Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
- Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
- Thormann, A. et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
- Amberger, J. S. & Hamosh, A. Searching Online Mendelian Inheritance in Man (OMIM): a knowledgebase of human genes and genetic phenotypes. *Curr. Protoc. Bioinformatics* **58**, 1.2.1–1.2.12 (2017).
- Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
- Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).
- McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
- Rangwala, S. H. et al. Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). *Genome Res.* **31**, 159–169 (2021).
- Lee, C. M. et al. UCSC Genome Browser enters 20th year. *Nucleic Acids Res.* **48**, D756–D761 (2020).
- Pujar, S. et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res.* **46**, D221–D228 (2018).
- MacArthur, J. A. L. et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.* **42**, D873–D878 (2014).

22. den Dunnen, J. T. Describing sequence variants using HGVS nomenclature. *Methods Mol. Biol.* **1492**, 243–251 (2017).
23. Miga, K. H. & Wang, T. The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
24. Li, H. et al. Exome variant discrepancies due to reference genome differences. *Am. J. Hum. Genet.* **108**, 1239–1250 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution

and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2022

Analysis

Methods

MANE Select workflow

To produce the MANE Select set, (1) both annotation groups developed pipelines to choose a representative transcript; (2) the two pipeline choices were compared; (3) the matched choices were updated to adjust the ends; and (4) when the two pipeline choices did not match, they were binned into multiple categories (Supplementary Table 1) to be resolved by pipeline refinements or manual review (Supplementary Methods 2). Although the two pipelines are described in detail in Supplementary Methods 1 and Extended Data Fig. 1, the key features are outlined here. The Ensembl pipeline takes into account evidence of functional potential, including transcript expression levels (Intropolis²⁵ and recount3; ref.²⁶) and evolutionary constraint of the coding region (Phylogenetic Codon Substitution Frequencies, PhyloCSF²⁷). Other factors are CDS length and concordance with the APPRIS²⁸ principal isoform and the UniProt/Swiss-Prot²⁹ canonical isoform. The pipeline assigns a score for each component from which a composite score is derived. The transcript with the highest composite score is selected as the Ensembl choice, although some length exceptions apply. The RefSeq Select pipeline uses a hierarchical list of parameters, with prior use in clinical reporting and conservation of the coding region (PhyloCSF) at the top. Each parameter is assigned a binary score, and the RefSeq Select transcript is chosen on the basis of a composite score reflecting the ranked choice of the individual parameters.

Defining UTRs

To standardize the 5' and 3' ends of the transcripts, we used high-throughput cap analysis of gene expression (CAGE) data from the FANTOM consortium³⁰ and poly(A)-seq data from multiple studies^{31–37}, respectively. For the 5' ends, we imported the CTSS TotalCounts data included for 2,006 runs of CAGE sequence data on the HelicoScope platform from 1,829 distinct samples mapped to the GRCh38 assembly (BioProject, PRJDB1099). The FANTOM data were reprocessed to combine CAGE clusters found in close proximity (within 50 nucleotides of each other) on the same strand and to re-analyse the TotalCounts data in the region of each merged cluster to find the maximum peak. The TSS was then recalculated to be the 5'-most peak in the merged cluster with a signal of at least 50% of the maximum peak. This criterion is referred to as the 'longest strong' rule. The goal of the reprocessing is to determine a frequently used TSS that is representative of the overall data rather than that with the absolute maximum tag counts. In this way, we maximize the coverage of commonly observed 5'-UTR bases (and any sequence- or structure-based features they contain). The reprocessed CAGE tracks are available from NCBI GGD as RefSeq-processed FANTOM CAGE peaks tracks. To update the transcripts to the calculated longest strong TSS, we used an automated process that identified CAGE clusters overlapping the first exons of transcripts or those within 500 nucleotides of the first nucleotide. Alternatively, we updated them manually when the genes required additional review. We followed a similar logic for the 3' end and poly(A)-seq clusters (Supplementary Methods 3).

Comparison of transcript ends with genomic TSS signatures

We scanned the genomic sequence for the following TSS signatures: (1) enrichment of purines (A or G), which is characteristic of RNA polymerase II transcription initiation, and (2) TATA box motifs at about –30 and initiator³⁸ motifs at –1 relative to the TSS. We performed a comparison with two datasets, transcripts at the beginning of this project as well as predating bulk CAGE-based transcripts and those in the current MANE set. We used HOMER³⁹ to analyse nucleotide frequencies, the FIMO⁴⁰ tool from the MEME suite to scan for motifs using a position weight matrix (PWM) from JASPAR⁴¹ for analysis of TATA boxes and a PWM from ref.³⁸ for analysis of the initiator motif. The 200-nucleotide sequence centred on each TSS was scanned using FIMO, and the position of the highest scoring match to the PWM was recorded using a

P-value threshold of 0.01. Additional details are provided in Supplementary Methods 3 and Extended Data Fig. 3.

MANE Plus Clinical workflow

The starting point for the MANE Plus Clinical set was the list of known P and LP variants available in the ClinVar 20200513 release. All P and LP variants were considered, regardless of their review status ('star' designation). We identified transcripts that contained conserved coding exons not represented in the MANE Select set and that overlapped these P or LP variants. This set of additional transcripts was manually reviewed to ensure the same high degree of quality as for the transcripts in the MANE Select set.

RefSeq and Ensembl/Gencode transcript updates

The annotation comparison logic used for the MANE workflow (Supplementary Methods 4) was adapted to compare transcripts from the early RefSeq and Ensembl/Gencode annotation sets with those of the most recent MANE release. The first comparison was carried out using the human RefSeq109 and Ensembl 92 annotation sets. For each MANE Select transcript, the comparison dataset was checked for transcript and CDS annotations that were completely identical; differed only in the extent of the 5' and 3' UTRs; differed in the CDS but had the same transcript splice pattern, indicating a change in start codon; or cases in which a transcript lacked an equivalent splice pattern. The comparisons were performed independently of transcript identifiers; in some cases, a transcript was indicated as 'new' when it was an update of an existing transcript but exons were added or removed. The comparisons did not consider sequence changes or the removal of poly(A) tails from some RefSeq transcripts, which resulted in additional updates.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The datasets generated during the current study are available on the NCBI FTP site (https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/) and the Tark webpage (http://tark.ensembl.org/web/mane_project/). Source data are provided with this paper. The datasets analysed during the current study can be accessed using the following resources. All Ensembl/Gencode annotation builds used in the comparison of RefSeq and Ensembl/Gencode transcripts for determination of transcript matches in the MANE analysis are available in the release 96–105 directories on the Ensembl FTP site (http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.gtf.gz). All RefSeq annotation builds used in the comparison of RefSeq and Ensembl/Gencode transcripts for determination of transcript matches in the MANE analysis are available at https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Homo_sapiens/annotation_releases/. The Ensembl canonical transcripts used for the comparison of gnomAD versus ClinVar versus MANE were from Ensembl release 103. These can be accessed using the Ensembl Perl API for release 103 with the following call on the gene: http://www.ensembl.org/info/docs/Doxxygen/core-api/classBio_1_1EnsEMBL_1_1Gene.html. Alternatively, the same data are available through the Ensembl REST API by using the following lookup endpoint: <https://jan2020.rest.ensembl.org/documentation/info/lookup>. The aggregated CTSS TotalCounts CAGE data and the CAGE clusters as computed by the FANTOM consortium were imported from http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks/hg38_fair+new_CAGE_peaks_phase1and2.bed.gz and <https://fantom.gsc.riken.jp/5/datahub/hg38/reads/>. The poly(A)-seq data used to generate the poly(A) clusters and to determine the poly(A) sites were from multiple studies listed in refs.^{31–37}. The data are available in study accessions SRP041182, SRP003483, SRP007359 and SRP133500 in

the NCBI Sequence Read Archive (SRA; <https://www.ncbi.nlm.nih.gov/sra>) and at PolyASite 2.0 (<https://www.polyasite.unibas.ch/>). APPRIS data are available at <https://appris.bioinfo.cnio.es/#/downloads>, which is updated for every Ensembl/Gencode release. These data are based on Ensembl releases 95–104. The PhyloCSF data used to identify conserved sequences were imported from <https://data.broadinstitute.org/compbio1/PhyloCSFtracks/>. Intron support data from Snaptron/recount3 were imported from <http://snaptron.cs.jhu.edu/data/>. Source data are provided with this paper.

Code availability

The analysis code used for this study is largely integral to the RefSeq and Ensembl/Gencode curation databases and was not designed for use in isolation or with other annotation datasets. The most critical aspect of the analysis was the tens of thousands of working hours spent in curator review to produce the final dataset. HOMER v4.11 is available at <http://homer.ucsd.edu/homer/>. FIMO v5.3.2 is available at https://meme-suite.org/meme/meme_5.3.2/doc/fimo.html. HISAT 2.2.1 is available at <http://daehwankimlab.github.io/hisat2/>.

25. Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).
26. Wilks, C. et al. Recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
27. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
28. Rodriguez, J. M. et al. APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.* **46**, D213–D217 (2018).
29. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
30. Noguchi, S. et al. FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
31. Wang, R., Zheng, D., Yehia, G. & Tian, B. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Res.* **28**, 1427–1441 (2018).
32. Zheng, D. et al. Cellular stress alters 3'UTR landscape through alternative polyadenylation and isoform-specific degradation. *Nat. Commun.* **9**, 2268 (2018).
33. Fontes, M. M. et al. Activity-dependent regulation of alternative cleavage and polyadenylation during hippocampal long-term potentiation. *Sci. Rep.* **7**, 17377 (2017).
34. Li, W. et al. Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. *BMC Biol.* **14**, 6 (2016).
35. Yang, Y. et al. PAF complex plays novel subunit-specific roles in alternative cleavage and polyadenylation. *PLoS Genet.* **12**, e1005794 (2016).

36. Li, W. et al. Systematic profiling of poly(A)ⁿ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.* **11**, e1005166 (2015).
37. Derti, A. et al. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183 (2012).
38. Vo Ngoc, L., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C. & Kadonaga, J. T. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.* **31**, 6–11 (2017).
39. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
40. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
41. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

Acknowledgements We would like to thank B. Aken, the Transforming Genomic Medicine Initiative Consortium (G. Black, S. Ellard, H. Firth, D. Fitzpatrick, M. Hurles, H. Rehm, J. Ware and C. Wright), the Clinical Genome Resource (H. Rehm and S. Harrison), DECIPHER (H. Firth and J. Foreman), the FANTOM Consortium (K. Hideya), PanelApp (E. McDonough), ClinVar (M. Landrum), OMIM (A. Hamosh), the CRUCS team within Ensembl (B. Flint, S. Hunt, E. Perry, S. Trevanion, A. Winterbottom and A. Yates), I. Armean, S. Bodda and F. Martin. Ensembl and Ensembl VEP are registered trademarks of EMBL. The work performed by the authors at EMBL-EBI was supported by the Wellcome Trust (WT200990/Z/16/Z, WT200990/A/16/Z, WT108749/Z/15/Z), the National Human Genome Research Institute of the National Institutes of Health under award number 2U41HG007234 and EMBL. The work performed by the authors at NCBI is supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions F.C. and T.D.M. served as joint last authors. F.C., A.F., T.D.M., J.M., S.P. and J.E.L. provided research design, management and leadership. E.B., F.C., K.D.P. and P.F. provided management support and secured funding. R.B., A.B., E.C., C.D., C.M.F., T.G., D.H., M.H., T.H., J.J., V.S.J., M.K., J.E.L., K.M.M., A.M., J.M., J.M.M., M.R.M., T.D.M., S.P., S.H.R., L.D.R., F.T.N., G.T., A.R.V. and D.W. performed manual curation and annotation updates. R.F., J.M.G., D.N.M., L.G., A.A., C.W., O.E. and V.K.K. developed the Select pipelines and processed outputs. J.M., S.P., J.E.L., R.B., C.D., M.K., A.M., G.T., A.A., V.K.K., O.E., C.W. and T.D.M. engaged in data analysis, data interpretation and pipeline quality control. J.M., S.P. and T.D.M. created the figures. J.M. and S.P. drafted the manuscript with significant review from F.C., T.D.M., A.F. and J.E.L. All authors reviewed the manuscript before submission.

Competing interests E.B. is a paid consultant for Oxford Nanopore Technologies and Dovetail, Inc. P.F. is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. All other authors declare no competing interests.

Additional information

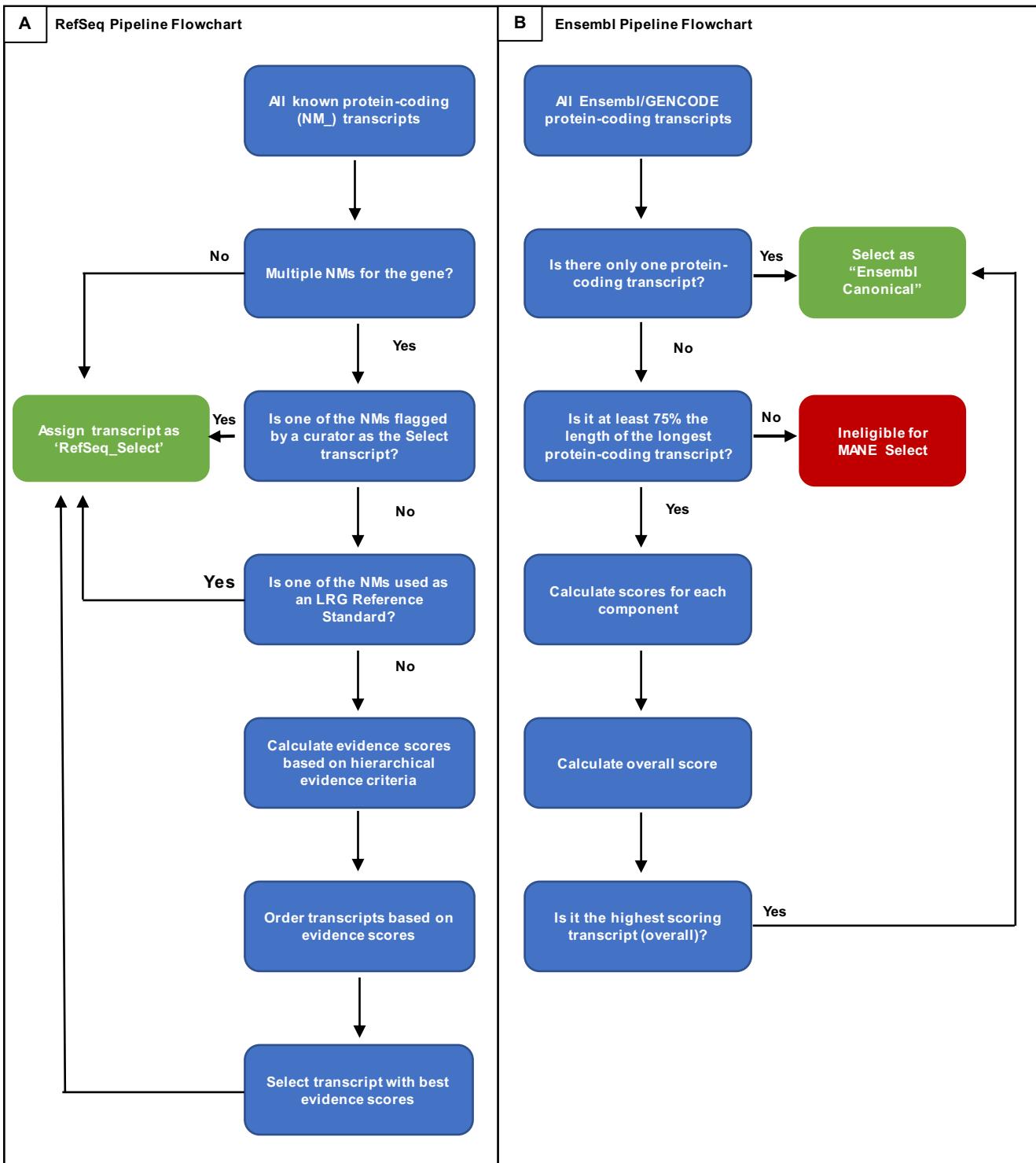
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-04558-8>.

Correspondence and requests for materials should be addressed to Terence D. Murphy.

Peer review information *Nature* thanks Marina DiStefano, Sharon Plon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

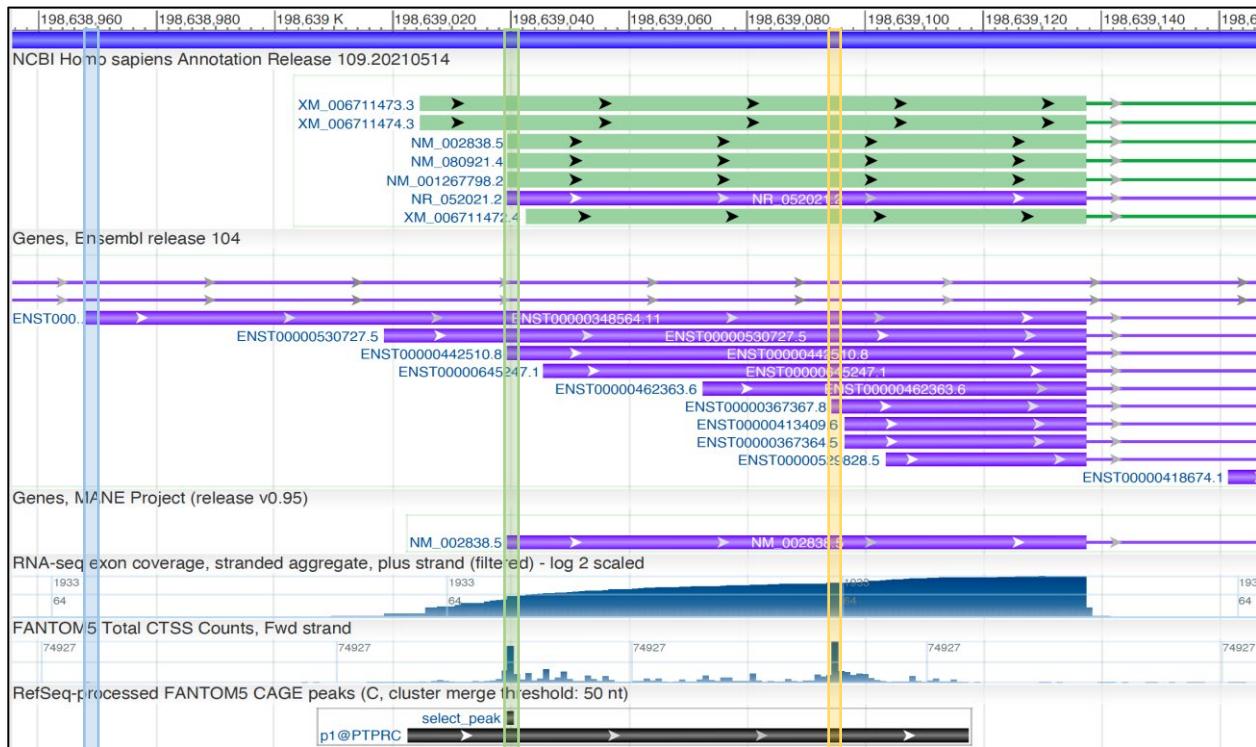
Reprints and permissions information is available at <http://www.nature.com/reprints>.

Analysis



Extended Data Fig. 1 | The Select pipelines. **a**, The RefSeq Pipeline picks the Select transcript based on a set of hierarchically scored criteria described in the Methods section and in more detail in Supplementary Method 1. **b**, The Ensembl pipeline assigns Ensembl Canonical to the transcript with the highest

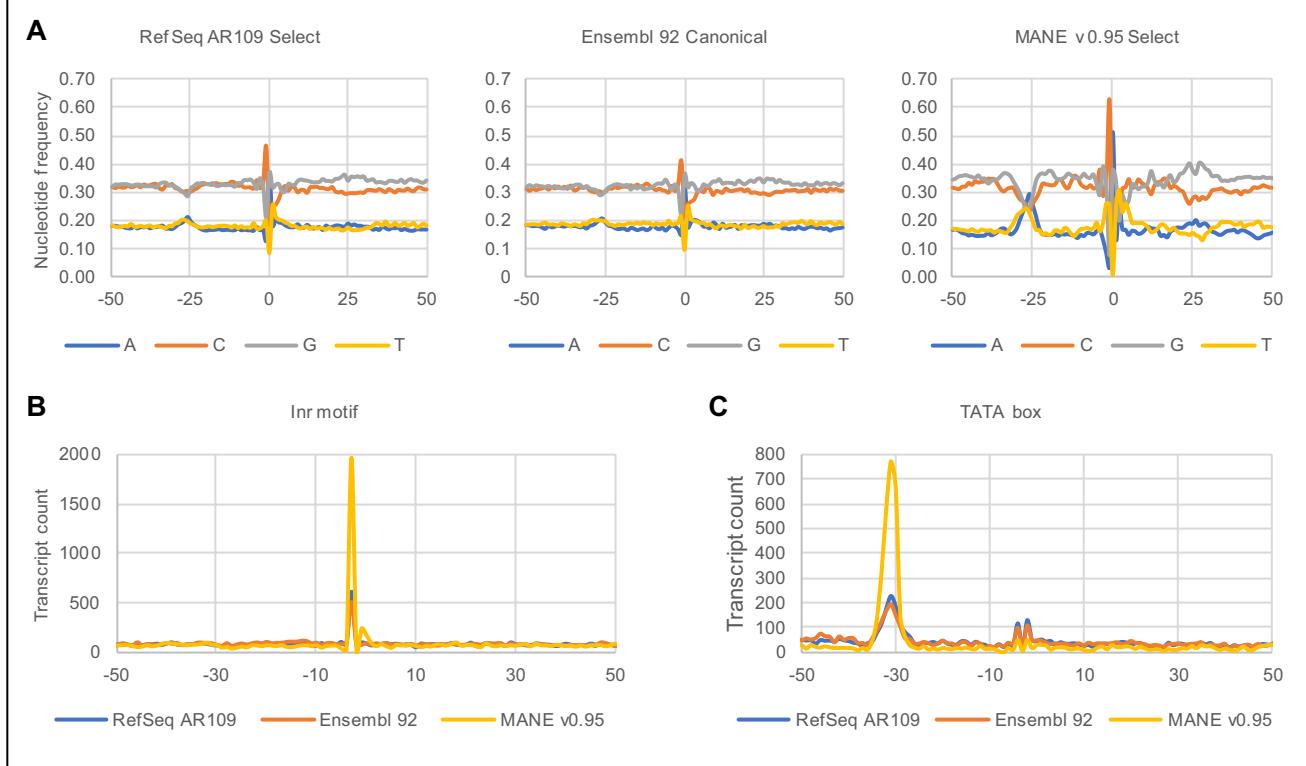
score, which is a sum of the component scores for each criteria (e.g., conservation, expression, APPRIS choice, UniProt choice, length). Details are listed in Supplementary Method 1.



Extended Data Fig. 2 | MANE collaboration UTR definition. Graphic display of the 5' terminal UTR exon of the gene *PTPRC* (HGNC:9666) in NCBI GDV to illustrate how we defined the 5' end of the transcript. Annotation tracks (top to bottom) show transcripts in RefSeq Annotation Release 109_20210514, transcripts in Ensembl Release 104 and the MANE Select (v0.95) track. The longest 5' UTR among the RefSeq and Ensembl/GENCODE annotation sets is flagged at the first base with a blue vertical box. The “FANTOM Total CTSS Counts” track displays histograms representing CAGE tag counts at each base

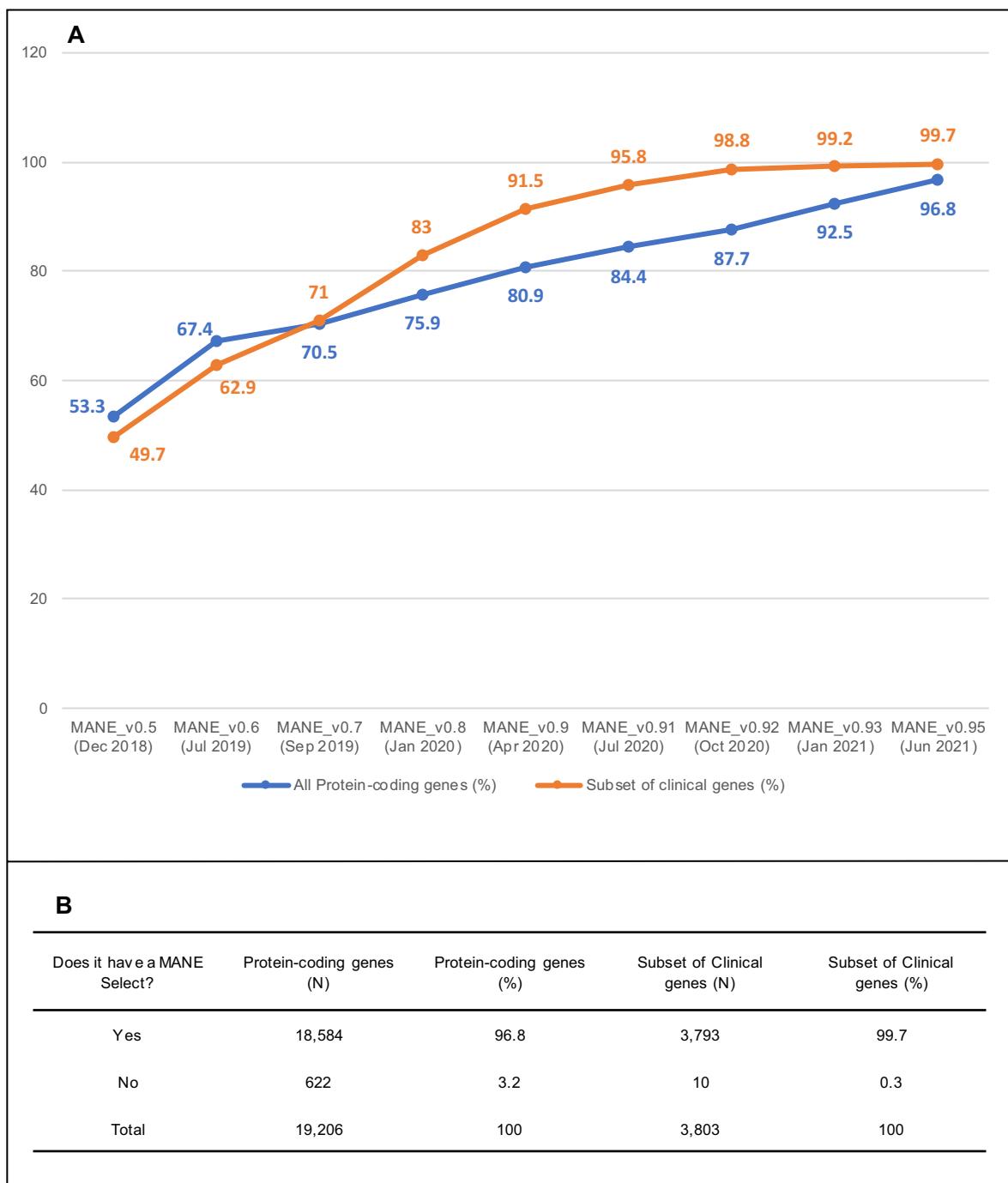
position. The strongest CAGE peak (the most abundant start site or the base position with the absolute maximum CAGE tag count) is highlighted with a yellow vertical box. The “RefSeq Processed CAGE” track at the bottom displays the start site (highlighted with a green vertical box) selected by the UTR algorithm. Details of how the UTR algorithm works are covered in the Methods and provided in Supplementary Method 3: UTR algorithm. A similar logic was used to compute polyA clusters and determine the 3' ends of transcripts.

Analysis



Extended Data Fig. 3 | Frequency of TSS signatures in RefSeq, Ensembl, and MANE transcripts. **A**) Frequency of A, C, G, T nucleotides at each position (y-axis) relative to the transcription start site (x-axis). MANE transcripts show an enrichment of Cat -1, and purine (A or G) at +1. **B**) Count of transcripts with a best Inr motif (y-axis) placed relative to the TSS (x-axis). The peak of Inr motifs

at -3 corresponds to the core CA motif located at -1 to +1. **C**) Count of transcripts with a TATA-box (y-axis) placed relative to the transcription start site (x-axis). The peak of TATA-box motifs at -31 corresponds to the core TATAAA box motif located at -28 to -23 upstream of the TSS. Details of the methods are available in Supplementary Methods 1.



Extended Data Fig. 4 | MANE Select coverage over time. (A) Graphical display of the percentage of all protein-coding genes (blue) and of the subset of clinical genes (orange) that have a defined MANE Select transcript per each MANE project release over time. (B) Number of genes that have a defined MANE Select

transcript (MANE v0.95). The list includes 101 genes that will require the MANE Select to be defined using an ALT or PATCH (rather than the GRCh38 Primary Assembly). It does not include an additional set of 345 genes that require review due to conflicting gene types between RefSeq and Ensembl/Gencode.

Analysis

PKP2 12:32790755-32896777

Reverse strand gene: plakophilin 2
Also known as: ENSG00000057294

<https://decipher.sanger.ac.uk/gene/PKP2/transcripts>

Function: "May play a role in junctional plaques." **Source:** UniProt

DECIPHER holds 2 sequence variants in this gene, in 2 open-access patients

Overview Matching patient variants 25 Matching DDD research variants 0 Phenotypes Phenotype b

Transcripts for PKP2: 1 to 7 of 7

Transcript	Location	Size
ENST00000340811.9 - PKP2-202 - NM_001005242.3 MANE Select	12 32790755 32896777	106.02 kb
ENST00000070846.10 - PKP2-201	12 32790854 32896777	105.90 kb

NM_001005242.3(PKP2):c.2259C>A (p.Tyr753Ter)

Allele ID: 858255
Variant type: single nucleotide variant
Variant length: 1 bp
Cytogenetic location: 12p11.21
Genomic location: 12: 32796207 (GRCh38) GRCh38 UCSC
12: 32949141 (GRCh37) GRCh37 UCSC

ClinVar

HGVS:

Nucleotide	Protein	Molecular consequence
NC_000012.11:g.32949141G>T		
NC_000012.12:g.32796207G>T		
NM_001005242.3:c.2259C>A MANE SELECT	NP_001005242.2:p.Tyr753Ter	nonsense

... more HGVS

Protein change: Y753*, Y797*

Other names:

gnomAD browser gnomAD v3.1 ▾ PKP2

gnomAD v3.1 released!

PKP2 plakophilin 2

https://gnomad.broadinstitute.org/gene/ENSG00000057294?dataset=gnomad_r2_1

Genome build GRCh38 / hg38
Ensembl gene ID ENSG00000057294.15
MANE Select transcript ENST00000340811.9 / NM_001005242.3
Ensembl canonical transcript ENST00000070846.10
Other transcripts ENST00000546498.1, ENST00000549461.1, and 3 more
Region 12:32790755-32896777
References Ensembl, UCSC Browser, and more

Extended Data Fig. 5 | Commonly used resources that have adopted the MANE Select in their browsers and display. Top panel: A screenshot of the gene page of *PKP2* (HGNC:9024) in the DECIPHER database (<https://www.deciphergenomics.org/>). The transcript table on the gene page shows the MANE Select label with the RefSeq and Ensembl identifiers (marked by a red box). Middle panel: A ClinVar variant display (<https://www.ncbi.nlm.nih.gov/clinvar/variation/870075/>) page for the gene *PKP2* (allele ID 858255). The HGVS

table in this page includes the RefSeq component of the MANE Select (indicated by red box). Bottom panel: A display page from the Genome Aggregation Database gnomAD v3.1. The MANE Select pair, along with the RefSeq and Ensembl identifiers, are displayed at the top of the page (indicated by red box). We note that UniProt, another commonly used resource, will update their browser soon to include flagged MANE Select proteins.

A

ow/hide columns (4 hidden) Filter

Name	Transcript ID	CCDS	UniProt Match	RefSeq Match	Flags
SCNSA-205	ENST00000423572.7	CCDS46797	Q14524-2	NM_000335.5	MANE Select v0.93 Ensembl Canonical GENCODE basic APPRIS ALT2 TSL:1
SCNSA-202	ENST00000333535.8	CCDS46796	Q14524-1	-	GENCODE basic APPRIS P4 TSL:1
SCNSA-203	ENST00000413689.6	CCDS46799	H9KVD2	NM_001099404.2	MANE Plus Clinical v0.93 GENCODE basic APPRIS ALT2 TSL:5
SCNSA-204	ENST00000414099.6	CCDS46798	E9PG18	-	GENCODE basic TSL:1
SCNSA-208	ENST00000455624.6	CCDS54570	E9PHB6	-	GENCODE basic TSL:1
SCNSA-207	ENST00000450102.6	CCDS54569	K4DIA1	-	GENCODE basic APPRIS ALT2 TSL:1
SCNSA-206	ENST00000449557.6	-	A0A0A0MT39	-	GENCODE basic APPRIS ALT2 TSL:5
SCNSA-201	ENST00000327956.6	-	A3EY21	-	TSL:1 CDS 3' incomplete
SCNSA-209	ENST00000464652.1	-	-	-	TSL:1
SCNSA-211	ENST00000491944.1	-	-	-	TSL:1
SCNSA-210	ENST00000476683.1	-	-	-	TSL:3

B

101.63 kb Forward strand →

MANE v0.95 transcripts with Ensembl labels

MANE v0.95 transcripts with RefSeq labels

Genes (Comprehensive set from GENCODE 38)

38.56Mb ENST00000413689.6

38.58Mb ENST00000423572.7

38.60Mb NM_000335.5

38.62Mb NM_001099404.2

38.64Mb

< SCNSA-204 protein coding

< SCNSA-205 protein coding

< SCNSA-203 protein coding

< SCNSA-202 protein coding

C

Transcript Archive | ReleaseSets | ReleaseStats | MANE Project | Search & Compare | REST-Tark | Help | Home | Privacy Notice | Search..

D

set Human genes (GRCh38.p13)

Filters

MANE Select transcript: Only
MANE Plus clinical: Only

Attributes

Gene stable ID
Gene stable ID version
Transcript stable ID
Transcript stable ID version

Dataset
[None Selected]

MANE Transcripts Note: "MANE Select" – representative transcripts that are matched between RefSeq and Ensembl

Show 25 entries

COPY CSV Excel PDF Column visibility Filter results:

Ensembl StableID	RefSeq StableID	MANE TYPE	Gene (Click to Search)
ENST00000263100.8	NM_130786.4	MANE SELECT	A1BG
ENST00000373997.8	NM_014576.4	MANE SELECT	A1CF
ENST00000318602.12	NM_000014.6	MANE SELECT	A2ML
ENST00000299698.12	NM_144670.6	MANE SELECT	A2ML1
ENST00000442999.3	NM_01080438.1	MANE SELECT	A3GALT2

GENCODE basic annotation Only Excluded

APPRIS annotation Only Excluded

Ensembl Canonical Only Excluded

MANE Select transcript Only Excluded

MANE Plus clinical Only Excluded

Extended Data Fig. 6 | Display of MANE data in Ensembl. (A) In Ensembl's Gene page, the Ensembl/GENCODE transcript(s) in the MANE set is highlighted with the "MANE Select" or "MANE Plus Clinical" flags, visible in the last column of the transcript table. The identical RefSeq transcript is highlighted in the same table, in the column titled "RefSeq Match". (B) Graphical representation is visible in the Location page after configuring the view by adding the custom-made MANE Project track hub (<https://ftp.ncbi.nlm.nih.gov/refseq/MANE/trackhub/hub.txt>). (C) The list of MANE transcripts can be accessed and downloaded from Ensembl's Transcript Archive (Tark) MANE Project page

(<http://tark.ensembl.org/web/manelist>) and programmatically using APIs available in the REST API page (http://tark.ensembl.org/api/#/transcript/transcript_manelist_list), or Ensembl's REST API e.g. <https://rest.ensembl.org/overlap/id/ENSG00000128573?feature=mane;content-type=text/xml>.

(D) MANE data can also be downloaded from Ensembl BioMart (<https://www.ensembl.org/biomart/martview/c24cb3213fe65da552fc8b755c2910c>) by choosing the 'Human Genes (GRCh38.p13) dataset and the 'MANE transcripts' filter.

Analysis

A **Genome Data Viewer**

Genes, MANE Project (Release v0.95)

(H) MANE v0.95 transcripts with RefSeq identifiers as primary labels

(H) MANE v0.95 transcripts with Ensembl identifiers as primary labels

NCBI Homo sapiens Annotation Release 109 20210514

SCN5A

Was this helpful?

B

[SCN5A – sodium voltage-gated channel alpha subunit 5](#)

Homo sapiens (human)
Also known as: CDDC2, CMD1E, CMPD2, HB1, HB2, HBBB, HH1, ICCD, IVF, LQT3, Nav1.5, PFHB1, SSS1, VF1
Gene ID: 6331
RefSeq transcripts (9) RefSeq proteins (9) RefSeqGene (1) PubMed (640)
Orthologs Genome Viewer BLAST Download

C ***Homo sapiens sodium voltage-gated channel alpha subunit 5 (SCN5A), transcript variant 2, mRNA***

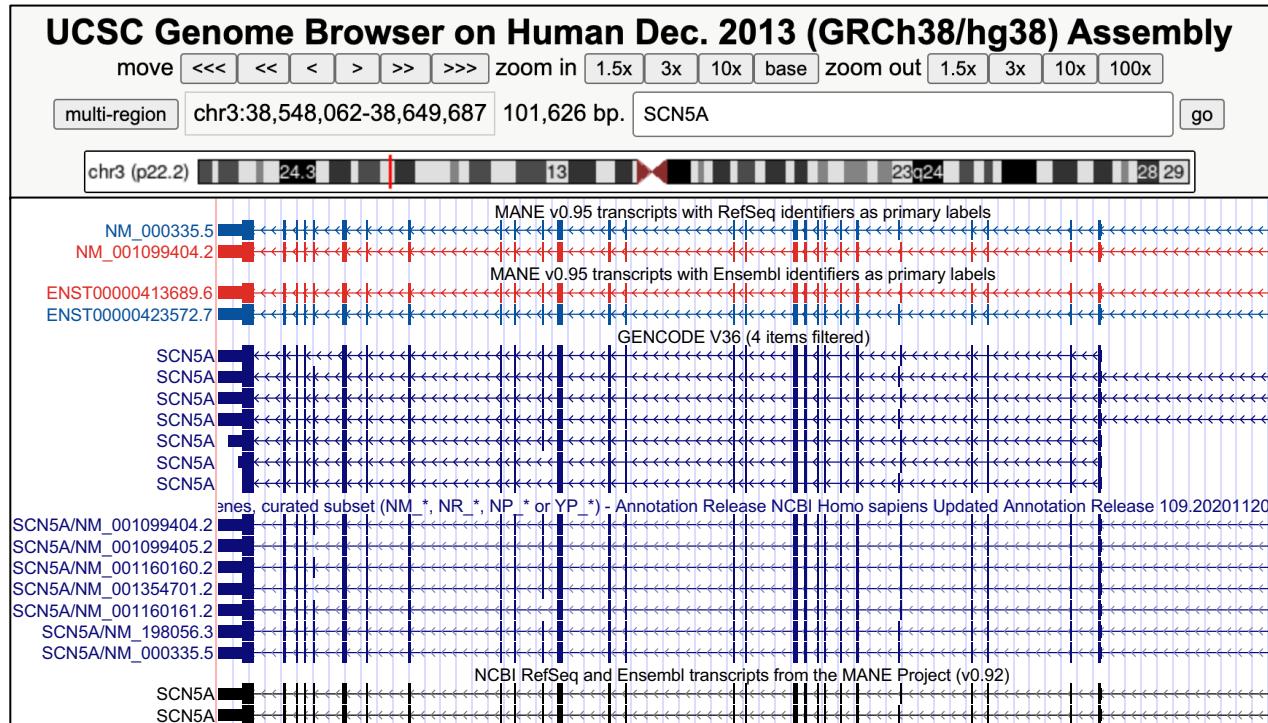
NCBI Reference Sequence: NM_000335.5
[FASTA](#) [Graphics](#)

Go to:

LOCUS	NM_000335	8516 bp	mRNA	linear	PRI 19-APR-2021
DEFINITION	Homo sapiens sodium voltage-gated channel alpha subunit 5 (SCN5A), transcript variant 2, mRNA.				
ACCESSION	NM_000335 XM_001131636				
VERSION	NM_000335.5				
KEYWORDS	RefSeq; MANE Select.				
SOURCE	Homo sapiens (human)				
ORGANISM	Homo sapiens				
	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo				

Extended Data Fig. 7 | Access to MANE data in NCBI resources. (A) Genome Data Viewer (GDV). The MANE track (green, at the top) shows RefSeq transcripts assigned as MANE Select and MANE Plus Clinical for the gene *SCN5A* (HGNC:10593). The middle section shows RefSeq and Ensembl identifiers included in the MANE sets, available by adding the MANE track hub using the ‘Configure Track Hubs’ menu. The bottom section shows a portion of RefSeq annotation release 109.20210514. (B) The gene search results page (shown here for the gene *SCN5A*), reached by searching for any human protein-coding gene in <https://www.ncbi.nlm.nih.gov/gene/>, flags the MANE Select in the expanded transcript list. (C) A portion of the transcript record of NM_000335.5, the MANE Select for *SCN5A*. The MANE Select tag (boxed) is included in the ‘KEYWORDS’ section. The keyword can be used in Nucleotide and Protein

database queries to extract a list of MANE Select transcripts. For example: PALM[gene] AND MANE Select[keyword]. The entire list of MANE Select transcripts can be obtained using the Entrez query “*Homo sapiens*[organism] AND MANE_select[keyword]”. MANE data can also be parsed from the annotation files available in the NCBI RefSeq FTP page (https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20210514/GCF_000001405.39_GRCh38.p13/) using the “MANE Select” tag attribute (tag=MANE Select in GFF3, or tag “MANE Select” in GTF), in the rows associated with the mRNA, CDS and exon features. In addition, column 9 also contains the matching Ensembl transcript identifier as an external database reference (Dbxref). Rows in the annotation files associated with the CDS feature contain the MANE Select tag, along with the matching Ensembl protein identifier.



Extended Data Fig. 8 | Access to MANE Data in UCSCbrowser. The MANE data are accessible in UCSC's Genome Browser as a data track in the Genes and Gene Predictions section (bottom of figure). MANE data can also be viewed in this browser by adding the track hub (<https://ftp.ncbi.nlm.nih.gov/refseq/MANE/>

trackhub/hub.txt), which displays the RefSeq and Ensembl identifiers of the MANE Select separately (top of figure), as shown in this display of the SCN5A (HGNC:10593).

Analysis

A

LRG_1431
Gene: CYP3A5

PUBLIC
This LRG has been made public (finalised) i.e. the fixed reference sequences will not change.

HGNC Gene Symbol (Identifier): CYP3A5 (HGNC:2638) | Made Public: 03 June 2020 | Last Update: 31 March 2021

Fixed reference sequences in this record

Number of sequences: Genomic: 1 / Transcript: 1 / Protein: 1

Genomic			Transcript			Protein			
Name	Length	Source	Name	Length	Source	Name	Length	Source	CCDS
LRG_1431	38,805 nt	NG_007938.2	t1	1,720 nt	NM_000777.5 ENST00000222982.8	p1	502 aa	NP_000768.1 ENSP00000222982.4	CCDS5672.1

MANE Select

LRG_1431 transcript

Transcript t1

Transcript identifier: LRG_1431t1
Comment(s):

- This transcript is identical to the RefSeq transcript NM_000777.5
- This transcript is identical to the Ensembl transcript ENST00000222982.8

B

LRG_6
Gene: ATP1A2

PUBLIC
This LRG has been made public (finalised) i.e. the fixed reference sequences will not change.

HGNC Gene Symbol (Identifier): ATP1A2 (HGNC:800) | Made Public: 17 March 2010 | Last Update: 31 March 2021

Fixed reference sequences in this record

Number of sequences: Genomic: 1 / Transcript: 1 / Protein: 1

Genomic			Transcript			Protein				
Name	Length	Source	Name	Length	Source	MANE type	Name	Length	Source	CCDS
LRG_6	34,834 nt	NG_008014.1	t1	5,436 nt	NM_000702.2	-	p1	1,020 aa	NP_000693.1 ENSP00000354490.3	CCDS1196.1

MANE Select

Transcript t1

Transcript identifier: LRG_6t1
Comment(s):

- This transcript is identical to the RefSeq transcript NM_000702.2 but with the polyA tail removed.
- The coding sequence of this transcript is identical to the coding sequence of Ensembl transcript ENST00000361216.8. The two transcripts differ at the 5'UTR. [Details](#)
- The MANE Select transcript for the gene ATP1A2 is NM_000702.4.

Extended Data Fig. 9 | MANE transcript display in LRG records. Screenshots of the LRG records for the genes *CYP3A5* (HGNC:2638) (http://ftp.ebi.ac.uk/pub/databases/Irgex/LRG_1431.xml) and *ATP1A2* (HGNC:800) (http://ftp.ebi.ac.uk/pub/databases/Irgex/LRG_6.xml) displaying MANE transcript annotations. As illustrated in this figure, if the LRG and MANE Select transcripts are identical (Panel A, LRG_1431 for *CYP3A5*), the MANE Select flag is displayed

in the Fixed Reference Sequence and Transcript sections of the LRG. In the event that the LRG transcript is not the MANE Select (Panel B, LRG_6 for *ATP1A2*), there will be no flag in the Fixed reference section but the MANE Select transcript will be listed in the Transcript section for the user's information.

Extended Data Table 1 | Links to access MANE data from Ensembl, NCBI and UCSC

Browser	Browser Access and Display	Bulk Download	Programmatic Access
NCBI	<p>Genome Data Viewer: https://www.ncbi.nlm.nih.gov/genome/gdv/ Open browser view for a gene using a gene symbol in the 'Search in Genome' box and use 'Tracks>Configure Tracks>Genes and Products>NCBI Genes' menu path to load MANE data in the browser view.</p> <p>Gene view: https://www.ncbi.nlm.nih.gov/gene/?term=human+TSC2 Expand 'Reference Sequences' in the above example to see the MANE Select accession. Replace TSC2 in the URL with NCBI or HGNC gene symbol of your choice.</p>	<p>NCBI FTP site: https://ftp.ncbi.nlm.nih.gov/efseq/MANE/MANE_human/current/</p>	<p>E-utils search nucleotide or protein database for "MANE Select[keyword]"</p>
Ensembl	<p>Ensembl Browser Location Tab: https://www.ensembl.org/Homo_sapiens/Location/View?g=GNE In 'Region in Detail', click on 'Custom tracks' on the left-hand menu. Click on 'Track Hub Registry Search'. Then type 'MANE' in the 'Text search' box. It should come back with 1 track hub. Click on 'Attach this hub'. Then exit the window. Click on "Configure this page" on the left-hand menu to select the MANE track. Click on the desired track and select the track style from the "Change track style" menu. Then exit the window and the track should appear in the browser.</p> <p>Gene Page: https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=GNE MANE transcripts and the equivalent RefSeq IDs are displayed in the transcript table.</p>	<p>Transcript Archive (Tark): http://tark.ensembl.org/web/access_mane_data/</p> <p>BioMart: http://www.ensembl.org/biomart/martview</p>	<p>REST-Tark (Transcript Archive RestAPI): http://tark.ensembl.org/api/#/transcript/transcript_manelist_1st</p> <p>Ensembl RestAPI: https://rest.ensembl.org/overlap/pid/ENSG00000128573?feature=mane;content-type=text/xml</p> <p>Perl API http://www.ensembl.org/info/docs/Doxygen/core-api/classBio_1_1EnsEMBL_1_1Transcript.html using the mane_transcript call.</p>
UCSC	<p>Genome Browser: https://genome.ucsc.edu/cgi-bin/hgTracks?hideTracks=1&knownGene=pack&refSeqComposite=pack&mane=full&db=hg38&position=chr16.2048020-2088718 Under the 'Genes and Gene Predictions' heading, click on the dropdown below 'MANE select v0.95' to select the preferred display mode.</p>	<p>FTP site: http://hgdownload.soe.ucsc.edu/gbdb/hg38/mane/</p>	<p>Table Browser https://genome.ucsc.edu/cgi-bin/hgTables</p> <p>Data Integrator https://genome.ucsc.edu/cgi-bin/hgIntegrator</p>

Examples are provided for browser access in each genome browser. Columns 2 and 3 have links for bulk download of the data and for programmatic querying.

Corresponding author(s): Dr. Terence Murphy

Last updated by author(s): 12/15/2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for data collection

Data analysis The analysis code used for this study is largely integral to the RefSeq and Ensembl/GENCODE curation databases, and not designed for use in isolation or on other annotation datasets. The most critical aspect of the analysis is from tens of thousands of person-hours in curator review to produce the final dataset. HOMER v4.11 is available from <http://homer.ucsd.edu/homer/>. FIMO v5.3.2 is available from https://meme-suite.org/meme/meme_5.3.2/doc/fimo.html. HISAT 2.2.1 is available from <http://daehwankimlab.github.io/hisat2/>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The datasets generated during the current study are available from NCBI's FTP site (https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/) and can also be accessed from EMBL-EBI's Transcript Archive (Tark) page (http://tark.ensembl.org/web/mane_project/).

The datasets analyzed during the current study can be accessed using the following resources:

Ensembl/Gencode annotation: All Ensembl/Gencode annotation builds used in the comparison of RefSeq and Ensembl/Gencode transcripts for determining transcript matches in MANE analysis are available in release 96-105 directories on the Ensembl FTP site (e.g. http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.gtf.gz)

RefSeq Annotation: All RefSeq annotation builds used in the comparison of RefSeq and Ensembl/Gencode transcripts for determining transcript matches in MANE analysis are available at https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Homo_sapiens/annotation_releases/

Ensembl Canonical Transcripts: The Ensembl canonical transcripts used for the comparison between gnomAD vs ClinVar vs MANE were from Ensembl release 103. These can be accessed using the Ensembl Perl API for release 103 using this call on the gene: http://www.ensembl.org/info/docs/Dxygen/core-api/classBio_1_1EnsEMBL_1_1Gene.html

Alternatively, the same data are available via the Ensembl REST API, using the lookup endpoint: <https://jan2020.rest.ensembl.org/documentation/info/lookup>.

CAGE: Aggregated “CTSS TotalCounts” CAGE data and the CAGE clusters as computed by the FANTOM consortium was imported from http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks/hg38_fair+new_CAGE_peaks_phase1and2.bed.gz and <https://fantom.gsc.riken.jp/5/datahub/hg38/reads/>

Poly A: PolyA-seq data used to generate polyA clusters and determine polyA sites were from multiple studies listed in references 31-37. The data are available in study accessions SRP041182, SRP003483, SRP007359, SRP133500 in NCBI’s Sequence Read Archive (SRA: <https://www.ncbi.nlm.nih.gov/sra>) and from PolyASite 2.0 (<https://www.polyasite.unibas.ch/>).

APPRIS: APPRIS data is available at <https://appris.bioinfo.cnio.es/#/downloads>. It is updated for every Ensembl/Gencode release. APPRIS data is based on Ensembl releases 95 - 104.

PhyloCSF: PhyloCSF data used to identify conserved sequences were imported from <https://data.broadinstitute.org/compbio1/PhyloCSFtracks/>.

Recount3: Intron support data from Snaptron/recount3 was imported from <http://snaptron.cs.jhu.edu/data/>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No experiments where sample size is relevant were conducted for this study
Data exclusions	No experiments where data exclusion is applicable were conducted for this study
Replication	No experiments where replication is applicable were conducted for this study
Randomization	No experiments where randomization is applicable were conducted for this study
Blinding	No experiments where blinding is applicable were conducted for this study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | | |
|-------------------------------------|--|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Human research participants |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

Methods

- | | |
|-------------------------------------|---|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |