D916–D923 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 3 December 2020*
doi: 10.1093/nar/gkaa1087

GENCODE 2021

Adam Frankish[1], Mark Diekhans[2], Irwin Jungreis[3,4], Julien Lagarde[5], Jane E. Loveland[1], Jonathan M. Mudge[1], Cristina Sisu[6,7], James C. Wright[8], Joel Armstrong[2], If Barnes[1], Andrew Berry[1], Alexandra Bignell[1], Carles Boix[3,4,9], Silvia Carbonell Sala[5], Fiona Cunningham[1], Tomás Di Domenico[10], Sarah Donaldson[1], Ian T. Fiddes[2], Carlos García Girón[1], Jose Manuel Gonzalez[1], Tiago Grego[1], Matthew Hardy[1], Thibaut Hourlier[1], Kevin L. Howe[1], Toby Hunt[1], Osagie G. Izuogu[1], Rory Johnson[11,12], Fergal J. Martin[1], Laura Martínez[10], Shamika Mohanan[1], Paul Muir[13,14], Fabio C. P. Navarro[6], Anne Parker[1], Baikang Pei[6], Fernando Pozo[10], Ferriol Calvet Riera[1], Magali Ruffier[1], Bianca M. Schmitt[1], Eloise Stapleton[1], Marie-Marthe Suner[1], Irina Sycheva[1], Barbara Uszczynska-Ratajczak[15], Maxim Y. Wolf[16], Jinuri Xu[6], Yucheng T. Yang[6,17], Andrew Yates[1], Daniel Zerbino[1], Yan Zhang[6,18], Jyoti S. Choudhary[8], Mark Gerstein[6,17,19], Roderic Guigó[5,20], Tim J. P. Hubbard[21], Manolis Kellis[3,4], Benedict Paten[2], Michael L. Tress[10] and Paul Flicek[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA, [3]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar St, Cambridge, MA 02139, USA, [4]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA, [5]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, Barcelona, E-08003 Catalonia, Spain, [6]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [7]Department of Bioscience, Brunel University London, Uxbridge UB8 3PH, UK, [8]Functional Proteomics, Division of Cancer Biology, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK, [9]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA, USA, [10]Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, [11]Department of Medical Oncology, Inselspital, University Hospital, University of Bern, Bern, Switzerland, [12]Department of Biomedical Research (DBMR), University of Bern, Bern, Switzerland, [13]Department of Molecular, Cellular & Developmental Biology, Yale University, New Haven, CT 06520, USA, [14]Systems Biology Institute, Yale University, West Haven, CT 06516, USA, [15]Centre of New Technologies, University of Warsaw, Warsaw, Poland, [16]Department of Biomedical Informatics at Harvard Medical School, 10 Shattuck Street, Suite 514, Boston, MA 02115, USA, [17]Program in Computational Biology & Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA, [18]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA, [19]Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA, [20]Universitat Pompeu Fabra (UPF), Barcelona, E-08003 Catalonia, Spain and [21]Department of Medical and Molecular Genetics, King's College London, Guys Hospital, Great Maze Pond, London SE1 9RT, UK

Received September 21, 2020; Revised October 21, 2020; Editorial Decision October 22, 2020; Accepted October 24, 2020

ABSTRACT

The GENCODE project annotates human and mouse genes and transcripts supported by experimental data with high accuracy, providing a foundational resource that supports genome biology and clinical genomics. GENCODE annotation processes make use of primary data and bioinformatic tools and analysis generated both within the consortium and externally to support the creation of transcript structures and the determination of their function. Here, we present improvements to our annotation infrastructure, bioinformatics tools, and analysis, and the

*To whom correspondence should be addressed. Tel: +44 1223 492581; Fax: +44 1223 494494; Email: flicek@ebi.ac.uk

advances they support in the annotation of the human and mouse genomes including: the completion of first pass manual annotation for the mouse reference genome; targeted improvements to the annotation of genes associated with SARS-CoV-2 infection; collaborative projects to achieve convergence across reference annotation databases for the annotation of human and mouse protein-coding genes; and the first GENCODE manually supervised automated annotation of lncRNAs. Our annotation is accessible via Ensembl, the UCSC Genome Browser and https://www.gencodegenes.org.

INTRODUCTION

GENCODE produces widely-used reference genome annotation of protein-coding and non-coding loci including alternatively spliced transcripts and pseudogenes for the human and mouse genomes and makes these annotations freely available for the benefit of biomedical research and genome interpretation. The GENCODE consortium develops, maintains and improves targeted tools, analysis and primary transcriptomic and proteomic data in support of gene and transcript annotation. These resources support updates to genes in all functional classes or biotypes, including (i) the discovery of new features such as novel protein-coding genes and long non-coding RNA (lncRNA) genes; (ii) the extension of existing annotation including the identification of novel alternatively spliced transcripts at protein-coding and lncRNA loci and (iii) the continuous critical reappraisal of existing annotation that may result in removal or reclassification of protein-coding genes that lack evidence of protein-coding potential given all data now available. GENCODE defines genes in terms of their transcriptional and functional overlap. The functional information implicit in the CDS of protein-coding gene supports decision making and provides high confidence in the interpretation of protein-coding genes. For lncRNAs, the lack of analogous knowledge makes representation of complex lncRNA loci difficult and we are working with lncRNA community and other reference annotation databases to improve their annotation.

Among other achievements, over the last two years we have developed a manually supervised automated annotation pipeline and an annotation triage tool to leverage the volume of data generated by current transcriptomics experiments while ensuring that the resulting annotated transcript models maintain the quality of expert human annotation. We have completed the first pass manual annotation of the mouse reference genome based on experiences on completing the human annotation in 2013 and have used whole genome PhyloCSF (1) analysis to generate ranked lists of candidate coding regions for investigation by expert human annotators. To support research responding to the COVID-19 pandemic, we have reviewed and improved the annotation for a set of protein-coding genes associated with SARS-CoV-2 infection and immediately released the results using a trackhub (2). We worked with the RefSeq (3) and UniProt (4) reference annotation databases toward achieving annotation convergence by ensuring that when a protein-coding gene or protein is present in one resource, it will be represented in the others or there will be an explanation why not. We are part of the Matched Annotation from NCBI and EMBL-EBI (MANE) project to define a single representative 'MANE Select' transcript for all protein-coding genes and ensure its structure and sequence is identical in both the Ensembl/GENCODE and RefSeq genesets. We annotated new human protein-coding genes based on improved analyses and experimental validation using mass spectrometry. We have also improved the annotation of lncRNAs via the discovery of novel loci and novel transcripts at existing loci primarily based on incorporating long transcriptomic sequence data generated using the CLS protocol (5).

GENE ANNOTATION INFRASTRUCTURE

We have made several key improvements to our processes and tools used for manual gene annotation.

The Ensembl/GENCODE geneset is a merge of the manual gene annotation created by the Ensembl-HAVANA team (methods and validation described in 6–8) and the automated annotation produced by the Ensembl Genebuild team (9,10). Historically, these data were produced separately and stored in independent and structurally different databases before being merged into a single set for release. To speed data release and reduce complexity, we have now moved all manual annotation and computational annotation into a single database for human (and another for mouse). In addition to continuing the support of manual annotation, this transition allows manual annotators to directly 'bless', update or remove computationally annotated models. Most significantly, new genes and transcripts released early via the GENCODE update trackhub will be assigned their Ensembl (ENSX) formatted stable IDs at their creation, having previously been given an interim ID (OTTX format).

Long-read transcriptomic sequencing methods including those from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) produce data volumes that require change to our manual annotation process. In response, we developed the TAGENE pipeline to support greater automation of transcript model creation based on long-read datasets generated both within GENCODE and by other groups. TAGENE implements filtering and merging of long transcriptomic datasets before clustering putative transcripts into loci (both existing and novel) and applying further filters based on other transcriptomic datasets, including RNA-seq supported introns and existing GENCODE annotation (Figure 1). The clustering and final filtering steps are applied following multiple iterations of manual review until a point is reached where the false positive rate for the addition of spurious models is $<0.1\%$.

To support higher throughput manual annotation we have developed a web-based gene annotation triage tool (Kestrel; manuscript in preparation). This software allows manual annotators to rapidly visualise, browse through and, via connection to the Annotrack annotation issue-tracking database (11), record decisions for large numbers of gene annotations and QC flags. It has been specifically designed in mind for 'quick decision' cases such as high throughput checking the validity of transcript models cre-

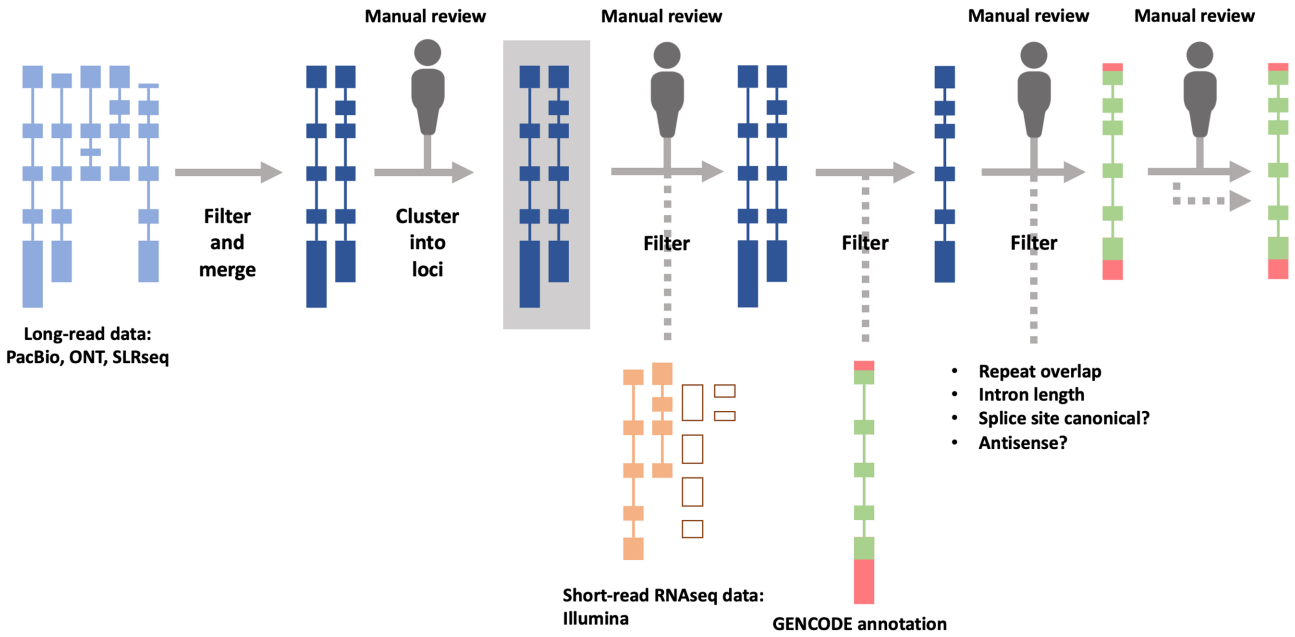

Figure 1. Schematic of the TAGENE workflow to add long transcriptomic data to GENCODE annotation. Points in the workflow where manual review is applied are indicated.

ated by the TAGENE pipeline. Kestrel is complementary to our set of high quality annotation tools in Zmap, Blixem and Dotter, which were initially developed for the clone-by-clone annotation approach used for the first pass annotation of the human and mouse reference genomes. Kestrel's streamlined functionality is often all that is required to answer emerging manual annotation questions and thus faster than our traditional workflow.

GENE ANNOTATION UPDATES

The GENCODE consortium has improved and extended the annotation of the human and mouse reference genomes and makes the annotation publicly available (see Table 1 for annotation statistics from the most recent GENCODE releases).

Since June 2018, ∼37 000 genes (∼32 000 human and 5000 mouse) and ∼63 000 transcripts (∼55 000 human and ∼8000 mouse) have either been created or updated in the GENCODE geneset (see Table 2 for a breakdown of new and updated genes and transcripts by functional biotype). During this period we have completed the first pass annotation of the mouse reference genome and conducted a number of tightly focussed annotation projects including the human and mouse olfactory receptor repertoire (12) and a re-annotation of developmental and epileptic encephalopathy-associated genes (13).

Although a number of protein-coding genes in both human and mouse have been added, removed or had their biotype changed over the past two years, the total number of genes is stable. Similarly, the number of pseudogenes of protein-coding genes is broadly stable for human, although our ability to better identify unitary pseudogenes has led to an increase in this specific class. In mouse, an increase in pseudogene count reflects the completion of manual annotation for all chromosomes. LncRNAs continue to show the largest increases in number, particularly in human where our efforts have been concentrated.

PROTEIN-CODING GENES

In response to the SARS-CoV-2 pandemic, we have applied our annotation resources to human genes with potential links to viral infection and COVID-19 disease primarily by investigating whether existing annotation for these genes can be improved. Our list of genes for reannotation comes from several sources including recently published drug repurposing studies identifying host proteins associated with other related coronaviruses (14) and human proteins found to physically associate with SARS-CoV-2 viral proteins in the cell (15). We also included genes curated by UniProt (4) and the Human Cell Atlas project (16) as well as interferon-stimulated genes with known antiviral activity (17). These efforts added previously unannotated alternatively-spliced transcript models and updated existing GENCODE transcript models, in particular 'partial' models that were incomplete at their 5′ and/or 3′ ends that could be extended to full length. All annotation takes advantage of long transcriptomic datasets and RNAseq data that was unavailable at the time of initial annotation. To date we have updated the annotation for 280 genes, adding ∼3700 novel transcripts and updating a further ∼850.

GENCODE has been actively collaborating with other reference annotation databases to try to achieve convergence on the annotation of protein-coding genes in human and mouse. The MANE project aims to create a single agreed transcript for every human protein-coding gene that has a 100% match for sequence and structure (splicing, UTR and CDS) in both the Ensembl/GENCODE and RefSeq (3) annotation sets. The project is driven by two in-

Table 1. Total numbers of genes and transcripts in the GENCODE 35 (Human) and GENCODE M25 (Mouse) releases by gene functional biotype

			Protein-coding	LncRNA	Pseudogene	sRNA	IG/TR
Human	**GENCODE 35**	**Genes**	19954	17957	14767	7569	645
		Transcripts	154580	48684	18664	7569	666
Mouse	**GENCODE M25**	**Genes**	21859	13197	13741	6108	700
		Transcripts	102241	18856	14522	6108	864

Table 2. Numbers of genes and transcripts that have been added to or updated in GENCODE Human and Mouse annotation since June 2018

		Human			Mouse		
		New	Updated	New and updated	New	Updated	New and updated
Genes	**Protein-coding**	131	17995	18126	845	1584	2429
	LncRNA	1965	7678	9643	670	282	952
	Pseudogene	75	4152	4227	676	266	942
	Total	2171	29825	31996	2191	2132	4323
Transcripts	**Protein-coding**	11334	21406	32740	4323	968	5291
	LncRNA	19042	2807	21849	1171	73	1244
	Pseudogene	247	259	506	794	137	931
	Total	30623	24472	55095	6288	1178	7466

dependent pipelines, one from each centre, followed by extensive investigation and discussion by expert human annotators where the pipelines do not agree. The latest release of MANE v0.91, gives an overall coverage of 84% of all protein-coding genes.

We have been working extensively to improve the interoperability of the existing annotations with UniProt. **G**enome **I**ntegration with **F**unc**T**ion and **S**equence (GIFTS) is a joint project between Ensembl and the EMBL-EBI component of the UniProt project and is currently available for human and mouse proteins https://www.ebi.ac.uk/gifts/. GIFTS calculates mappings and pairwise alignments between Ensembl transcripts that have a protein translation with their corresponding UniProt protein entries. Unmapped UniProt proteins are investigated by annotators from both teams and edited where necessary. We have investigated 1044 unmapped human (716) and mouse (328) proteins from UniProt and identified cases where the GENCODE annotation needs to be updated (2 human, 49 mouse), and proteins that appear invalid in their putative genomic context (640 human, 54 mouse).

We continue to analyse publications external to the GENCODE consortium reporting additional protein-coding genes in the light of GENCODE criteria. For example, we examined the novel protein-coding genes reported in the CHESS gene annotation set (18), adding five protein-coding genes, 16 pseudogenes and 37 lncRNAs. A recent survey of heart ORFs (19), has so far resulted in the annotation of 12 additional human protein-coding genes.

GENCODE annotation makes substantial use of comparative genomics to help identify regions on the genome with protein-coding potential. For example, we have used Cactus to create a 600-way vertebrate whole genome alignment incorporating data from the 200 Mammals and Bird 10K projects as the basis of a single base-pair resolution map of evolutionary selection (20). We will directly use these alignments within the PhyloCSF phylogenetic analysis tool (1). The PhyloCSF pipeline has also been run on the each new release of the human and mouse genome annotations to facilitate the discovery of additional novel coding genes, novel pseudogenes, and novel coding sequence (21). We have automated our process to generate updated lists of PhyloCSF Candidate Coding Regions (PCCRs), which are then examined by manual annotators. In human, PCCRs are part of the standard annotation workflow. In mouse, a targeted review of unannotated PCCRs analogous to that previously undertaken in human has led to the identification of 64 novel protein-coding genes, 376 novel coding exons in preexisting protein-coding genes, and 202 pseudogenes including 56 unitary pseudogenes. PhyloCSF has also been used to identify candidate ribosomal stop codon readthrough events in human and mouse (22,23). Following manual review of these and several others identified experimentally, 14 and 11 genes with stop codon readthrough events have been annotated in human and mouse, respectively (Figure 2).

GENCODE annotation utilises proteomics data to supplement transcriptomic and evolutionary evidence of protein-coding functionality and we have continued to both generate experimental MS data and use publicly available data sets to aid the identification and annotation of protein-coding genes. Our data generation focus is on elements of the proteome that are missed by standard proteomics approaches including the use of 155 novel synthetic peptides targeting distinct and unique peptides mapping to putative coding genes, newly discovered protein coding genes that require validation, and pseudogenes that have shown strong peptide evidence in previous experiments. These peptides are compiled into a reference spectral library, which is used to validate their existence in our experimental proteomics data and large public MS datasets. For example transcriptomic, conservation, and ribosome profiling data combined with experimental peptide evidence supported the discovery and validation of an alternate protein isoform originating from a non-ATG start site in the gene POLG (24), and highlighted a novel class of unannotated protein-coding features that are now under active investigation.

To support the automated analysis of proteomics data for genome annotation we collaborated with the PRIDE (25) proteomics repository at EMBL-EBI to build a reprocess-

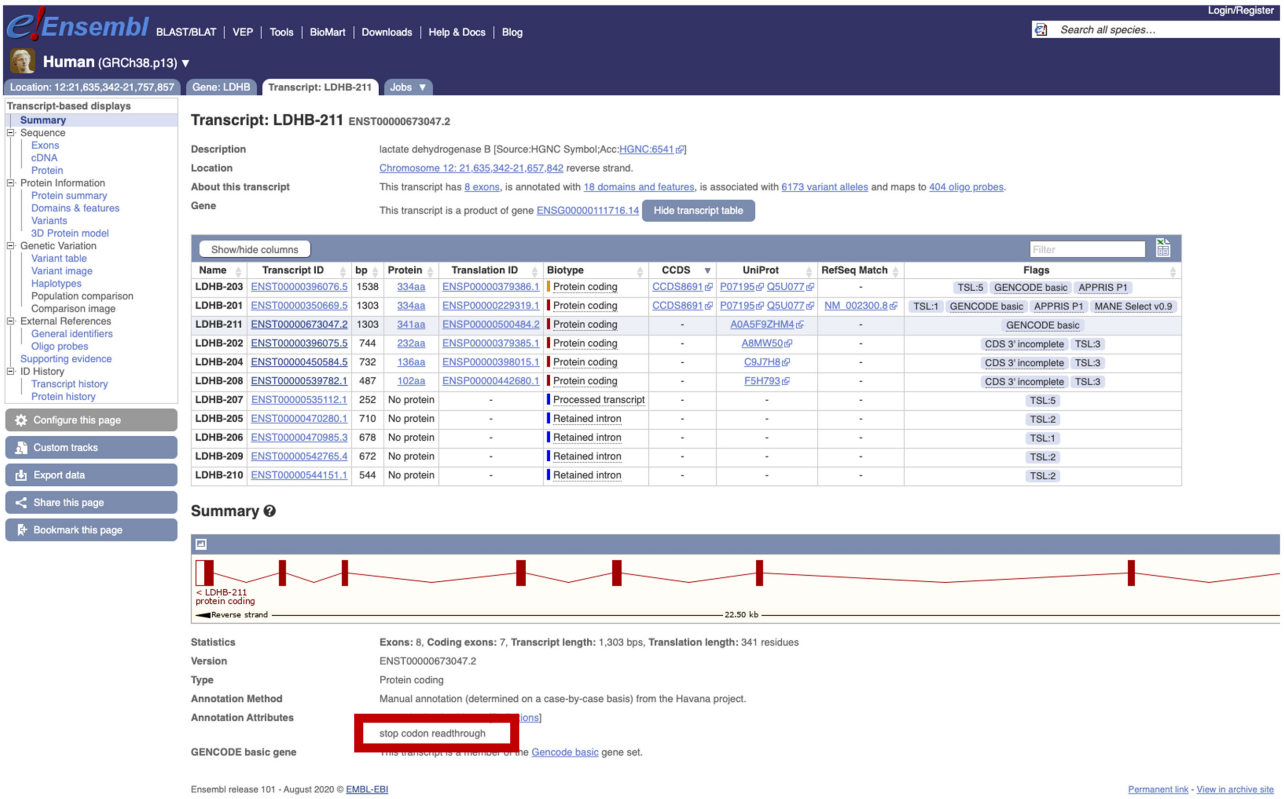

Figure 2. Screenshot from the Ensembl genome browser of the transcript view page for the gene LDHB, which contains a transcript (ENST00000673047, LDHB-211) with an annotated stop-codon readthrough event. The location of the annotation attribute flagging the stop-codon readthrough is highlighted by the red box.

ing and peptide-to-genome mapping pipeline for public proteomics.

Finally, we developed a pipeline based on UniProt (4), APPRIS (26), PhyloCSF (1), Ensembl gene trees (10), RNA-seq, MS and variation data to identify annotated protein-coding genes with weak or no support. This method enables us to scrutinise currently annotated protein-coding genes in the human and mouse gene set for misclassified gene models. To date we have flagged as potential non-coding genes more than 2475 human and 1807 mouse genes that were annotated as protein-coding. These are then reviewed in an iterative and ongoing process by expert manual annotators and retained, removed or reclassified based on their current supporting evidence. To date, ∼1000 human protein-coding genes have been reviewed and 119 removed or reclassified. A complementary approach has also been developed to identify missing and partially complete gene models in the human genome and submit to manual review.

LncRNAS

We have made improvements to the Capture Long Sequencing (CLS) lab protocol (5), including a 5′ cap selection step ('CapTrap') (27), which increases the proportion of sequenced full-length transcripts and the use of Spiked-in RNA Variant Control Mixes (SIRVs). Applying CLS, we have generated long transcriptomic data targeting a variety of suspected lncRNA-producing genomic loci in both human and mouse. Focusing primarily on unannotated regions such as GWAS sites, putative enhancers, and non-GENCODE lncRNA catalogs (e.g. miTranscriptome (28), NONCODE (29), FANTOM CAT (30)). In total we have produced more than 36 million ONT reads and 2 million PacBio Sequel (PBS) reads identifying thousands of potential novel loci (∼1600 in Human, ∼4500 in mouse) in currently unannotated genomic regions for review and inclusion in the Ensembl/GENCODE geneset. Long transcriptomic sequence data produced within GENCODE and from public data archives has been run through our TAGENE workflow and the results of this first set of analysis released to the public in GENCODE 31 (June 2019). These initial results have already made a significant difference to the coverage of lncRNAs in GENCODE, with the addition of 1711 novel loci and 17 858 transcripts, an 11% and 60% increase compared to the previous release respectively.

PSEUDOGENES

Our pseudogene annotation has benefited from the analysis of new datasets. For example, using RNA-seq datasets from ENTEx-pseudogene expression in various human tissues we have developed a computational framework to accurately quantify the expression level of pseudogenes, and identify actively transcribed pseudogenes in each tissue. We have also used our pseudogene annotation in 16 closely re-

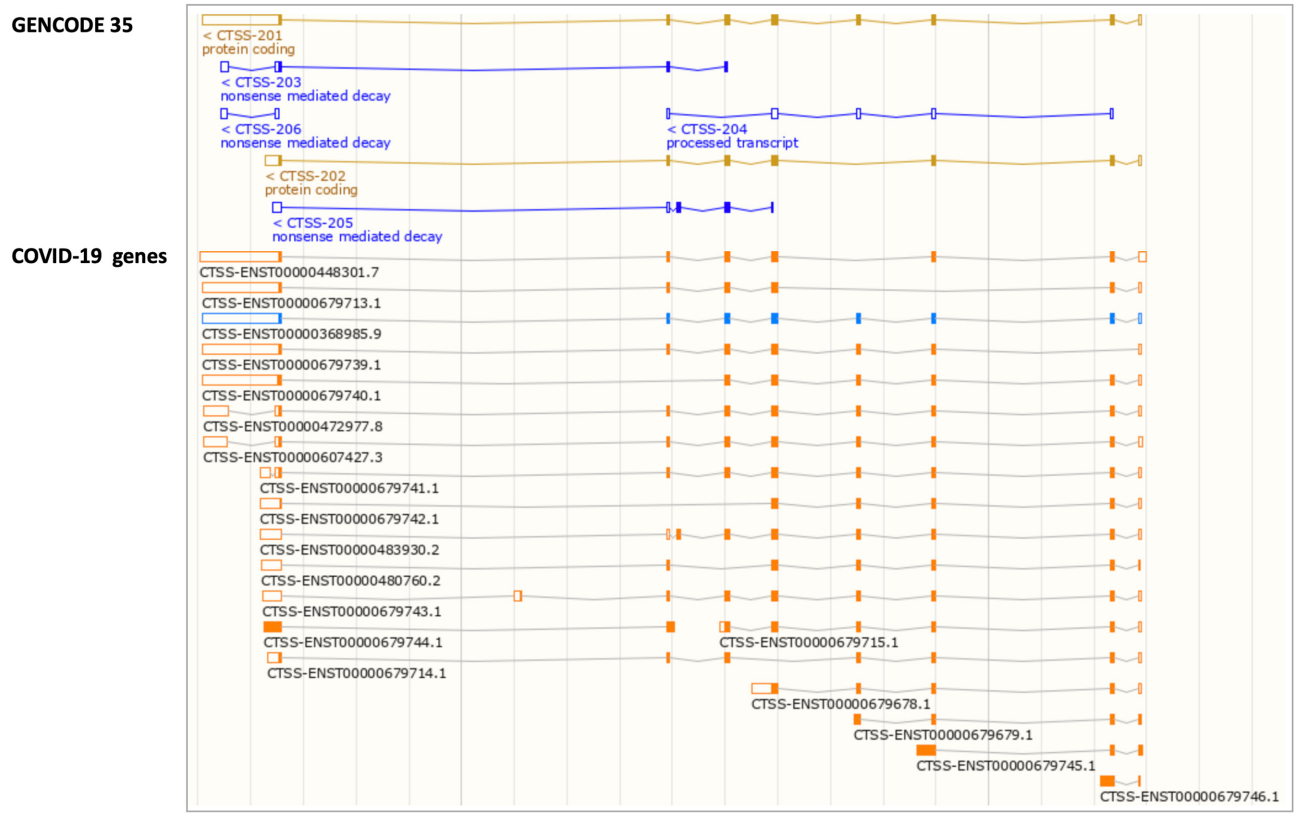

Figure 3. A screenshot from the Ensembl genome browser of the location view for the CTSS gene. The Comprehensive annotation from GENCODE 35 is shown in the upper panel and the updated annotation in the COVID-19 genes trackhub is shown in the lower panel. Transcript models that are unchanged with respect to release Ensembl 101 are coloured blue, whereas new models or pre-existing models that have been modified are shown in orange.

lated mouse strains from the Mouse Genomes Project (31) to create orthology relationships for the conserved annotations and the identification of patterns of pseudogene gain and loss between strains (32) and give a prototype for work annotating human pseudogenes leveraging variation across the human population.

DATA ACCESS

GENCODE gene sets are currently updated up to four times each year for both human and mouse. Each release is versioned and made available immediately upon release from Ensembl (6) and https//www.gencodegenes.org with release on the UCSC Genome Browser (33) normally following shortly thereafter. The current human release is GENCODE 35 (August 2020) and the current mouse release is GENCODE M25 (April 2020). Additional information and previous releases can be found at https//www.gencodegenes.org.

GENCODE is the now the standardised default human and mouse annotation for both the Ensembl and UCSC genome browsers following a transition of UCSC's mouse annotation in April 2019. Data is presented through all of the standard interfaces from both resources.

To expedite public access to updated annotation between releases, all annotation changes are made freely available within 24 h via the 'GENCODE update' Track Hub, which can be accessed at both the Ensembl and UCSC genome browsers. In the Ensembl browser, the hub has been added to the Track Hub Registry (accessed via the 'Custom tracks' section), and can be connected to by searching for 'GENCODE update'. Alternatively, the data can be added as a custom track in both Ensembl and UCSC browsers (http://ftp.ebi.ac.uk/pub/databases/gencode/update_trackhub/hub.txt). Additionally, a trackhub of updates to genes associated with COVID-19 can be accessed in the same way (http://ftp.ebi.ac.uk/pub/databases/gencode/covid19_trackhub/hub.txt). In the 'COVID-19 genes' track data view, transcript models that are unchanged with respect to release GENCODE 35/Ensembl 101 are coloured blue, whereas new models or pre-existing models that have been modified are shown in orange (Figure 3). We also offer BED and gtf files for these annotations.

We have made available the public 'Synonymous Constraint' track hub in the UCSC Genome Browser that shows protein-coding regions under synonymous constraint, indicating an overlapping function, and synonymous accelerated regions, indicating a high mutation rate (https://data.broadinstitute.org/compbio1/SynonymousConstraintTracks/trackHub/).

Supported GENCODE annotation is available on the GRCh38 human reference assembly and the GRCm38 mouse reference assembly. Selected human releases are mapped back to the GRCh37 assembly and made available from UCSC and https://www.gencodegenes.org as a service

to the community. The resulting mapping are not manually checked and may have errors especially in complicated regions of the human genome. We recommend use of the GRCh38 annotations if possible.

Training about the GENCODE annotation and its use is available from the Ensembl and UCSC training team and user support is available from the Ensembl and UCSC helpdesks.

CONCLUSION

The GENCODE consortium leverages the best available data, analysis and tools to continually improve the gene annotation of the human and mouse reference genomes. We have developed new methods and workflows to take advantage of the increasing quality and volume of data, and in particular long transcriptomic data, while maintaining the specificity afforded by expert human oversight. We expect our ability to use new data to improve our coverage of novel genes and alternatively spliced transcripts will allow us to move towards a more complete representation of all gene features of known functional classes as we monitor the emergence of new functional features that may require annotation such as alternative translations of known coding genes, non-canonical translations in, for example, lncRNAs and mRNA with multiple functions.

FUNDING

National Human Genome Research Institute of the National Institutes of Health [U41HG007234]; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health; Wellcome Trust [WT108749/Z/15/Z, WT200990/Z/16/Z]; European Molecular Biology Laboratory; Swiss National Science Foundation through the National Center of Competence in Research 'RNA & Disease' (to R.J.); Medical Faculty of the University of Bern (to R.J). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. Paul Flicek is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

REFERENCES

1. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–82.
2. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
3. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745
4. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515
5. Lagarde,J., Uszczynska-Ratajczak,B., Carbonell,S., Pérez-Lluch,S., Abad,A., Davis,C., Gingeras,T.R., Frankish,A., Harrow,J., Guigo,R. *et al.* (2017) High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet.*, **49**, 1731–1740.
6. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**, S4.
7. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
8. Howald,C., Tanzer,A., Chrast,J., Kokocinski,F., Derrien,T., Walters,N., Gonzalez,J.M., Frankish,A., Aken,B.L., Hourlier,T. *et al.* (2012) Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.*, **22**, 1698–1710.
9. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., García Girón,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database (Oxford)*, **2016**, baw093.
10. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
11. Kokocinski,F., Harrow,J. and Hubbard,T. (2010) AnnoTrack–a tracking system for genome annotation. *BMC Genomics*, **11**, 538.
12. Barnes,I.H.A., Ibarra-Soria,X., Fitzgerald,S., Gonzalez,J.M., Davidson,C., Hardy,M.P., Manthravadi,D., Van Gerven,L., Jorissen,M., Zeng,Z. *et al.* (2020) Expert curation of the human and mouse olfactory receptor gene repertoires identifies conserved coding regions split across two exons. *BMC Genomics*, **21**, 196.
13. Steward,C.A., Roovers,J., Suner,M.M., Gonzalez,J.M., Uszczynska-Ratajczak,B., Pervouchine,D., Fitzgerald,S., Viola,M., Stamberger,H., Hamdan,F.F. *et al.* (2019) Re-annotation of 191 developmental and epileptic encephalopathy-associated genes unmasks de novo variants in SCN1A. *NPJ Genom. Med.*, **4**, 31.
14. Zhou,Y., Hou,Y., Shen,J., Huang,Y., Martin,W. and Cheng,F. (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.*, **6**, 14.
15. Gordon,D.E., Jang,G.M., Bouhaddou,M., Xu,J., Obernier,K., White,K.M., O'Meara,M.J., Rezelj,V.V., Guo,J.Z., Swaney,D.L. *et al.* (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, **583**, 459–468.
16. Rozenblatt-Rosen,O., Stubbington,M.J.T., Regev,A. and Teichmann,S.A. (2017) The Human Cell Atlas: from vision to reality. *Nature*, **550**, 451–453
17. Schoggins,J.W. and Rice,C.M. (2011) Interferon-stimulated genes and their antiviral effector functions. *Curr. Opin. Virol.*, **1**, 519–525.
18. Pertea,M., Shumate,A., Pertea,G., Varabyou,A., Breitwieser,F.P., Chang,Y.C., Madugundu,A.K., Pandey,A. and Salzberg,S.L. (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, **28**, 208.
19. an Heesch,S., Witte,F., Schneider-Lunitz,V., Schulz,J.F., Adami,E., Faber,A.B., Kirchner,M., Maatz,H., Blachut,S., Sandmann,C.L. *et al.* (2019) The translational landscape of the human heart. *Cell*, **178**, 242–260.
20. Armstrong,J., Hickey,G., Diekhans,M., Fiddes,I.T., Novak,A.M., Deran,A., Fang,Q., Xie,D., Feng,S., Stiller,J. *et al.* (2020) Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**, 246–251.
21. Mudge,J.M., Jungreis,I., Hunt,T., Gonzalez,J.M., Wright,J.C., Kay,M., Davidson,C., Fitzgerald,S., Seal,R., Tweedie,S. *et al.* (2019) Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res.*, **29**, 2073–2087.
22. Jungreis,I., Chan,C.S., Waterhouse,R.M., Fields,G., Lin,M.F. and Kellis,M. (2016) Evolutionary dynamics of abundant stop codon readthrough. *Mol. Biol. Evol.*, **33**, 3108–3132.
23. Loughran,G., Jungreis,I., Tzani,I., Power,M., Dmitriev,R.I., Ivanov,I.P., Kellis,M. and Atkins,J.F. (2018) Stop codon readthrough generates a C-terminally extended variant of the human vitamin D receptor with reduced calcitriol response. *J. Biol. Chem.*, **293**, 4434–4444.
24. Khan,Y.A., Jungreis,I., Wright,J.C., Mudge,J.M., Choudhary,J.S., Firth,A.E. and Kellis,M. (2020) Evidence for a novel overlapping

coding sequence in POLG initiated at a CUG start codon. *BMC Genet.*, **21**, 25.
25. Perez-Riverol,Y., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M., Pérez,E. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
26. Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A. and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.*, **46**, D213–D217.
27. Carninci,P., Kvam,C., Kitamura,A., Ohsumi,T., Okazaki,Y., Itoh,M., Kamiya,M., Shibata,K., Sasaki,N., Izawa,M. *et al.* (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327–336.
28. Iyer,M.K., Niknafs,Y.S., Malik,R., Singhal,U., Sahu,A., Hosono,Y., Barrette,T.R., Prensner,J.R., Evans,J.R., Zhao,S. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
29. Fang,S., Zhang,L., Guo,J., Niu,Y., Wu,Y., Li,H., Zhao,L., Li,X., Teng,X., Sun,X. *et al.* (2018) NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.*, **46**, D308–D314.
30. Hon,C.C., Ramilowski,J.A., Harshbarger,J., Bertin,N., Rackham,O.J., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T.M., Severin,J. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, **543**, 199–204.
31. Lilue,J., Doran,A.G., Fiddes,I.T., Abrudan,M., Armstrong,J., Bennett,R., Chow,W., Collins,J., Collins,S., Czechanski,A. *et al.* (2018) Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.*, **50**, 1574–1583.
32. Sisu,C., Muir,P., Frankish,A., Fiddes,I., Diekhans,M., Thybert,D., Odom,D.T., Flicek,P., Keane,T.M., Hubbard,T. *et al.* (2020) Transcriptional activity and strain-specific history of mouse pseudogenes. *Nat. Commun.*, **11**, 3695.
33. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.