Specific Aims: Provide aims that address the overall goals of the project, including all the components (1 page)

Overall Specific Aims

The overall goal of the GENCODE consortium is to annotate all evidence-based gene features in the human and mouse genomes with high accuracy and release these annotations for the greatest possible benefit to biomedical research and genome interpretation.

Aim 1: Extend the human and mouse GENCODE gene sets to as near completion as possible given current experimental technology

This aim will focus on the incorporation of additional tissue-specific isoforms, novel features and extending partial and incomplete transcript annotations to full length as the primary methods to increase the quality and completeness of the protein-coding and non-coding annotation. Key well-established technologies for this aim include manual annotation and functional validation based on protein, cDNA, EST, RNA-seq and mass spectrometry data as well as core informatics methods for gene annotation, coding potential and quality control. We will incorporate RNA capture and long read transcriptome data and other relevant technologies for discovery and annotation. Human and mouse annotation benefit strongly from shared methods with the primary initial goals of extending human partial models to full length and completing the initial full pass of mouse for coding, non-coding and pseudogene annotation.

Aim 2: Population based genome annotation

The overall goal of this aim is to ensure that any transcript expressed in a human individual will be present in the reference annotation set. We will apply a similar goal to a set of key mouse strain genomes. GENCODE will also actively annotate the increasing number of alternative haplotypes that are a part of the genome assemblies maintained and distributed by the Genome Reference Consortium. We will extend our methods for automatic discovery/prioritization of variable transcripts from population transcriptomics datasets such as GTEx. Finally, as graph genome representations mature, GENCODE will pilot methods to annotate graph genomes and present its annotation on a graph representation of the genome. The tools generated in this aim are likely to be useful for creating, in effect, a personalized GENCODE.

Aim 3: Extend annotation to a definition of the gene that include core regulatory regions and tissue specific enhancers from selected datasets

This aim will begin as a pilot project that seeks to integrate and annotate data types that directly connect transcripts to relevant regulatory regions and thus annotate a more comprehensive definition of what a gene is. We will proceed as a series of pilot activities GENCODE using a combination of computational and manual approaches within and focused on data generated to initially identify regulatory activity such as polymerase recruitment, transcription initiation, epigenomes, cis-regulatory interactions and physical interactions. The results of the pilot project will inform our decision about how the most informative of these datasets will be incorporated into the GENCODE annotations.

Aim 4: Distribute GENCODE annotation and engage with community annotation efforts

We will maintain current popular distribution channels for GENCODE data including the GENCODE web site and the Ensembl and UCSC Genome Browsers, while developing support for distribution of GENCODE annotation via Global Alliance for Genomics and Health (GA4GH) compatible APIs. We will establish new mechanisms for prioritizing genes for manual annotation and for community input. Our goal is to firmly establish GENCODE as the standard annotation set used in all research and clinical genomics efforts.

Significance

The sequencing of the human genome and the resulting reference human assembly is one of the greatest scientific achievements of the 21st century. We are now apparently entering the promised new era in medicine where genomics will play a much larger and possibly game changing role.

As we have sequenced and analyzed the genomes of more and more people, a better understanding of a 'normal' genome has emerged, and determining the range of normal is potentially an important part of defining what it means to have a genetic disease. Indeed, the variety of the genome has surprised many. We have discovered that structural and copy number variation is pervasive and consequential [1], we have found that everyone's genome contains a significant number of protein truncating or loss of function mutations[2] and we are only beginning to understand the spectrum of functional sequence changes that occur in and modify disease causing pathways[3,4]. At the same time, the genome sequences of the most commonly used mouse strains facilitate the most effective use of these key models for large-scale knockout analysis[5] and disease-specific research[6].

Highly accurate genome annotation is a vital foundation to these studies and other efforts including, for example, CRISPR/Cas9 based genome editing. Accurate genome annotation is critical to the planned large-scale initiatives to sequence humans for research and clinical care including the US national Precision Medicine Initiative and the UK's 100,000 Genomes Project. The size and scope of these efforts brings into sharp relief the resources required for them function effectively and deliver the promised results[7,8]. Specifically, the annotation of the genome is the primary interpretation substrate for both genomic medicine and genome research, and every error in the annotation will eventually lead to an error in interpretation. Many of these interpretation errors will be inconsequential, some will not.

Overall goals

The objective of the GENCODE consortium is to create a foundational reference genome annotation. Our overall goal is to identify and classify all gene features in the human and mouse genomes with high accuracy and based on defined biological evidence, and then to release these annotations for the benefit of biomedical research and genome interpretation. GENCODE focuses on protein-coding and non-coding loci including their alternatively spliced isoforms and pseudogenes. To achieve this, we will continue our successful approach of leveraging computational and experimental methods to identify new genes and new transcript isoforms, directing manual annotation to regions requiring expert annotation. The GENCODE consortium has established workflows and years of demonstrated achievement. While GENCODE's goal is resource generation, to remain efficient and current we will adopt improved sequencing technologies such as PacBio and provide annotation on the full reference genomes as they move toward a graph structure. We will use our expertise to extend GENCODE genes into their regulatory regions.

GENCODE today

The GENCODE annotation is highly used in both large-scale and small projects. GENCODE is the default human and mouse gene set at Ensembl and the default human gene set for the UCSC Genome Browser (UCSC also provides the mouse GENCODE set and plans to switch to it as default). GENCODE is the gene set used for major projects including the Exome Aggregation Consortium (EXAC), GTEx[9], 1000 Genomes Project, TCGA, ICGC and ENCODE. GENCODE engages directly with the Mouse Genome Informatics (MGI) resource at the Jackson Laboratory and with NCBI as part of the

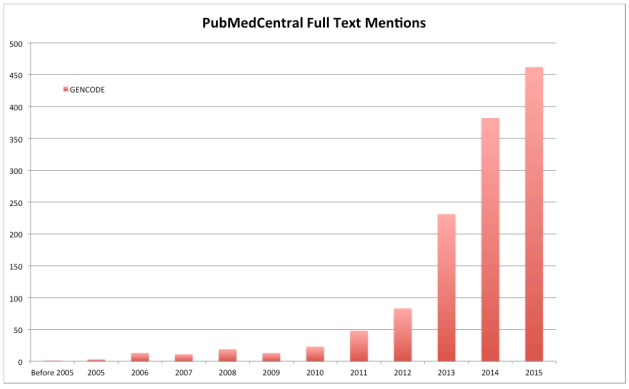


Figure 1: Number of times per year that the text "GENCODE" appears in PubMedCentral (PMC). The full text search of only the articles that are in PMC undercounts usage because only a fraction of papers are contained in PMC. Note that before the GENCODE project, there is apparently only one mention in PMC.

Consensus Coding Sequence (CCDS) project and one result of this is the comprehensive MGI mouse gene set used by the International Mouse Phenotyping Consortium (IMPC).

The growth of GENCODE usage has been dramatic over the past four years (Figure 1), and Google scholar has more than 2200 citations for the main GENCODE papers. These numbers are an underestimate of the true usage of GENCODE: many papers using the GENCODE annotation do not formally cite it and other cite the data source as a genome browser (Ensembl or UCSC) instead of the GENCODE project.

There have been several evaluations conducted by independent groups comparing the GENCODE genes to other gene sets for various purposes, and they have universally recommend the use of GENCODE as the best and most comprehensive human annotation[10,11]. We have also done specific comparisons and published our results[12]. These efforts have helped us understand exactly how the GENCODE annotation is used and catalyzed improvements such as the introduction of GENCODE Basic, a subset of the complete GENCODE annotation containing only full-length transcripts and described in more detail below, which addressed a concern that the number of GENCODE transcripts may make RNA-seq analysis more complicated.

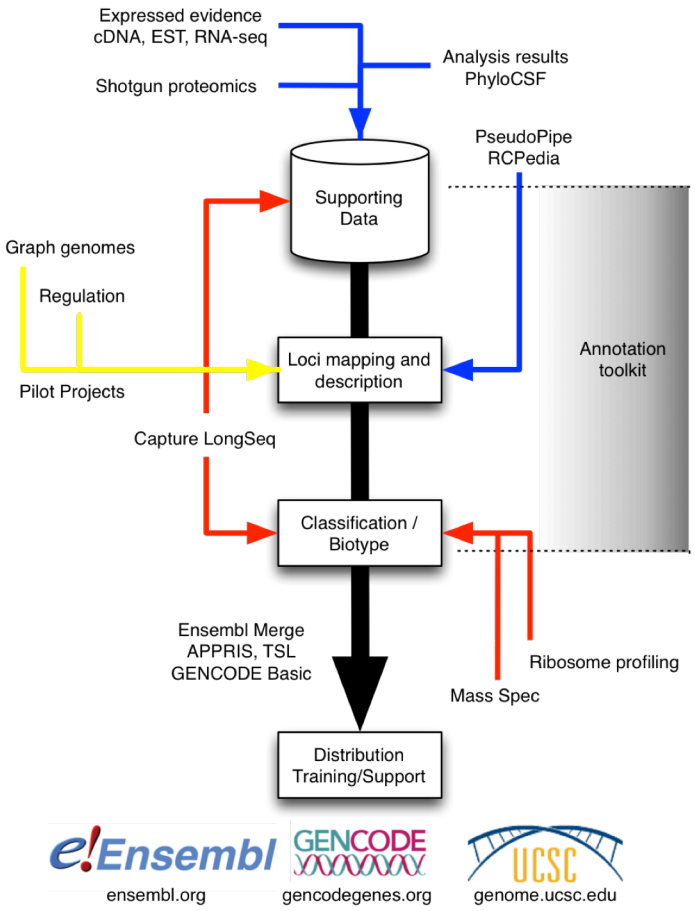


Figure 2: A simplified schematic of the core GENCODE processes, data and analysis flow. Full integrative uses of the various GENCODE and external data sets are described in the text.

Despite the large strides made by the GENCODE consortium and others since the completion of the human genome sequence, the identification and representation of the genes and transcripts they encode remain incomplete (see Progress Report). All classes of gene loci are affected including protein-coding genes, pseudogenes, long non-coding RNAs (lncRNAs) and small RNAs. The deficit manifests at multiple levels: complete absence of annotation, partial annotation, underannotation and misannotation. A gene locus may be completely absent where no transcripts associated with it are annotated. Given the relative stability in the total protein-coding gene and pseudogene numbers for recent GENCODE releases, it is likely that the majority of unannotated loci are lncRNAs. Partial annotation may occur where either alternatively spliced (AS) transcripts are absent from a locus which has some representation or where transcript annotation is not extended to its full length, almost certainly because it is based on non-full-length evidence such as ESTs. Underannotation occurs where a transcript is annotated with the correct structure but has suboptimal functional annotation. Given our role as a reference annotation resource, GENCODE do not include unsupported features. For example if a transcript in a protein-coding locus starts at a novel internal exon, no CDS is added due to uncertainty over whether the true start of the transcript has been found. Misannotation occurs where incorrect structural or functional annotation is present. This can be attributable to error, although GENCODE's extensive QC seeks to minimize this, but is more likely to be caused by the absence of a required

orthogonal dataset at the time of annotation. For example, a locus may be initially classified as a lncRNA until mass spectrometry data generated later demonstrates protein-coding potential.

We have now entered an era where technological improvements in transcriptomics and proteomics offer the possibility of complete annotation of all gene loci. For example, the emergence and improvement of third generation sequencing technologies such as PacBio and Synthetic Long Read RNA-seq (SLRseq)[13] together with the extension of recent techniques based on second generation short-read sequences such as ribosome profiling (Ribo-seq)[14], Cap Analysis of Gene Expression (CAGE)[15], RNA annotation and mapping of promoters for analysis of gene expression (RAMPAGE)[16], polyAseq[17] and shotgun proteomics will individually allow us to reduce the degree of error and to target incompleteness. However, GENCODE's strength is its ability to integrate multiple different data-types to achieve the best possible annotation of gene and transcript structure and function. Going forward, it is by building on this established expertise and utilizing multiple orthogonal datasets in combination that we will be able to shrink the gaps in annotation. Furthermore, GENCODE will be able to generate the data it needs to enrich gene annotation for clinically important genes and resolve annotation problems.

Innovation

Advancing toward a fully complete and correct genome annotation requires integration of a diverse set of evidence data and the application of well-established and proven processes (Figure 2). Computational methods are extremely rapid, consistent and informative, but to date no automatic approach is able to achieve the depth of integration provided by an experienced and trained manual annotator, especially in biologically complex regions. That a manual approach must be employed for at least part of the process is hardly surprising: while the practice of medicine has seen tremendous automation over the last half century, a future of automated computer diagnosis for every patient remains distant.

Since its inception, GENCODE has developed into a combination of well-established and conservative core procedures supplemented by targeted investigations ("pilot projects") into the value of new technologies, new data and new sources of evidence. For example, short read RNA-seq data was found to be largely inappropriate for exact isoform definition, but very useful for supporting intron locations in transcripts defined by other evidence (see QC procedures section). These pilots are a major source of innovation in the project and critical for ensuring that GENCODE remains up-to-date and in line with relevant technologies (see below). Over the course of this proposal we will follow major directions in genomics including graph-based genome representations, long-read transcriptome sequencing, connecting genes and regulatory regions affecting their transcription, and identifying genes that are not present on the current reference assembly. These pilots will determine whether and how each of these technologies contribute to the GENCODE reference annotation and, as appropriate, will be integrated into the core GENCODE processes.

Approach

The GENCODE consortium convened thirteen years ago with the aim of annotating gene regions for the ENCODE project and has resulted in an invaluable resource that is widely used (see above and in the Management, Dissemination and Training section). This enduring collaboration has **four fundamental components: (1) a comprehensive gene annotation pipeline leveraging manual annotation; (2) an integrated approach to pseudogene identification and classification; (3) a set of computational methods to evaluate and enhance gene annotation; and (4) complementary experimental pipelines for validation and functional annotation**. These fundamental components work in concert through various defined feedback loops to ensure that the right information is used in the right part of the project at the right time. The individual components and their integrated connections will be leveraged for the continued annotation of human and mouse and extended as appropriate based on the outcomes of the pilot projects. For all activities the focus and overall goal of GENCODE is the annotation of all evidence-based gene features at high accuracy.

Our experimental validation pipeline, pilot projects and resulting annotation approaches will be designed specifically to further the goals of GENCODE. However, the experimental design, methods and software are more widely applicable and can be leveraged for annotation groups working on other organisms. For example, we have shared expertise and annotation tools with annotators working on zebrafish, pig, rat and other species.

Over the last four years, GENCODE completed a full first pass manual annotation of the human genome, conducted extensive QC on the annotation, including extensive experimental validation, and investigated promising novel data types and datasets (see Progress Report). Going forward, the annotation of the human genome sequence will follow a similar path of testing new data types and extending the existing data types into new cell-lines, tissues, and developmental stages generated within the GENCODE consortium, by select other collaborators and deposited in the public repositories. GENCODE will develop annotation strategies to utilize them optimally and integrate them into our workflows to identify missing features and improve and update the existing annotation. Combining multiple novel datasets will allow us to formalize our guidelines for data integration, while the large volumes of new data with direct relevance to gene annotation will require continued development of new methods to prioritize data for use in annotation.

The GENCODE annotation of the mouse reference genome is less complete than that of the human reference genome due to an initial concentration of effort on human. As such, mouse will benefit from continued traditional manual annotation, both chromosome-by-chromosome and from targeted lists of genes and gene families, to ensure consistency with human annotation and support comparative analysis between the two species. However, we will also be able to rapidly adopt the updated methods piloted in human in order to retain as similar standards of annotation as possible for the two genomes, given the likely differences in the experimental datasets that are produced for them. In particular, human has much more experimental data, but mouse has access to tissues and developmental datasets unavailable in human research.

Comprehensive gene annotation pipeline

The GENCODE gene sets for human and mouse comprise a core of manual annotation for protein-coding, long non-coding RNA and pseudogene loci[18-20]. These are supplemented by Ensembl GeneBuild annotation as described below for small ncRNA genes, novel transcripts, and mouse genes in regions that are not yet manually annotated[21]. Experienced human annotators from the Human and Vertebrate Analysis and Annotation (HAVANA) team use the ZMap/Otter annotation toolkit (described below) to define transcript structure and function by integrating a large number of orthogonal data types, computational predictions of genic features and literature (Figure 2). Transcript structures are predominantly determined based on refined alignments of transcriptomic data generated by first, second and third generation sequencing technologies. Transcription start sites are identified using CAGE and RAMPAGE data, while polyAseq performs the same role for transcription termination sites. The protein-coding potential of transcripts is investigated using protein homologies from reference databases, cross-species conservation as defined by PhyloCSF[22] and PhastCons[23], and alignment of shotgun proteomics and Ribo-seq data. Loci where no transcripts show evidence of protein-coding potential are classified as lncRNAs. Pseudogene transcript models are annotated based on support from protein homologies and the identification of disabling mutations or retrotransposition and lack of locus-specific transcription. Manual annotators are partly guided by computational approaches that report new data or highlight inconsistencies. These include Ensembl, PhyloCSF and predictions of pseudogenes and retrotransposed loci (see below). Where a data type is not available to be displayed locally in ZMap, annotators access via UCSC, Ensembl or specialized browsers such as Zenbu[24].

Integrated approach to pseudogene identification and classification

The Yale group has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and HAVANA groups, Yale have developed a variety of methods to identify pseudogenes[20,25,26]. These include PseudoPipe, which takes as input all known protein sequences in

the genome and uses homology search to identify disabled copies of functional paralogs (referred to as pseudogene parents). Based on their formation mechanism pseudogenes are classified into 2 main types: processed and unprocessed. A second method, RCPedia, focuses on the annotation of retrotransposed (processed) pseudogenes[27]. The UCSC retrocopy annotation pipeline and Ensembl GeneBuild will supply additional supporting evidence for the annotated pseudogene models. These pipelines and extended versions of them will be used to essentially complete the annotation of pseudogenes in the mouse genome including annotation of both the mouse reference and the recently available mouse strain assemblies. The pseudogene collection will be characterized by expression activity across mouse tissues and in the mouse strain collection and further classified to identify unitary and polymorphic pseudogenes across the strains. Finally, these methods will be extended to support the annotation pseudogene variability across human individuals and, in doing so, help to refine our understanding of the boundary between protein-coding genes and pseudogenes.

Computational methods to evaluate and enhance gene annotation

The Ensembl GeneBuild provides an automated, independent method to identify and annotate all protein-coding genes, small and long non-coding RNA genes, and pseudogenes. Ensembl gene annotation is high quality, as judged by community assessments of computational annotation methods, and is based on a well-established core data flow that integrates alignments of expressed protein, cDNA and other biological sequences[28]. The primary data used to inform gene annotation are: protein sequences from UniProt, full-length mRNA and transcriptome sequences from ENA, and Rfam resources for small non-coding RNA genes. The Ensembl GeneBuild is merged with the HAVANA manual annotation to create the full GENCODE gene set and is especially valuable for filling in transcripts that may be expressed in difficult to access human tissues and for regions of the mouse genome that have not yet had comprehensive manual annotation. Ensembl is also the sole source for small noncoding RNA genes in both human and mouse.

The Ensembl RNA-seq pipeline[29] provides identification of transcribed regions and in particular has provided the basis for much of the lincRNA annotation in GENCODE, as well as providing an additional level of support for regions that otherwise have limited or inconclusive data from other sequencing technologies. A particular advantage of the RNA-seq pipeline is that it provides tissue-specific expression information.

Additional computational methods have proved highly valuable for evaluating, classifying and prioritizing gene annotations and these serve both as input to inform the main annotation pipeline as well as important information that is added to the transcripts of the final GENCODE set. Specifically, we use the current version of PhyloCSF to help identify the thousands of estimated novel protein-coding exons that remain unannotated within existing human protein-coding loci. Simultaneously, PhyloCSF will be updated to be more effective at finding protein-coding loci that have non-typical signatures of conservation. The CNIO isoform annotation pipeline (APPRIS)[30] and UCSC's Transcript Support Level (TSL) method add valuable and complementary information about the depth and type of experimental support of the transcripts in the final GENCODE set and will continue to be developed.

Experimental validation

Complementary experimental approaches at CRG, CNIO and WTSI will be used to discover, verify and validate various annotations within the GENCODE project. Specifically, CRG will use the targeted annotation of known and novel RNA transcripts by "Capture Long-Seq" (CLS): RNA capture followed by PacBio third generation sequencing[31]. This approach enables us to focus new transcript discovery on a candidate genomic locus and achieve complete or almost complete transcript models for each region (see Progress Report). CLS will be deployed to both complete existing annotation and to map new transcribed loci for a series of complex human and mouse tissues in both adult and embryonic time points. In addition, CNIO and WTSI will use high resolution tandem mass spectrometry shotgun proteomics to validate protein-coding potential for newly annotated transcripts and novel protein-coding genes from our core annotation pipeline as well as confirm transcribed pseudogenes, identify

alternative isoforms and nonsense-mediated targets. This will be extended to include targeted analysis approaches to be able to specifically focus on genes or features of interest.

Adapting to advances in genomics

Over the past decade, as both experimental technology and computational methods have advanced, new types of evidence supporting genome annotation have become widely available. In some cases, these represent incremental changes in accuracy or efficiency, while in other cases fundamentally new data types are used to assay existing or newly discovered phenomenon. For example, RNA-seq has become a primary method for transcriptome quantification and ribosome profiling, while noisy[32,33], provides insight into which mRNA transcripts are being actively translated. New data will likely also lead to new computational approaches for comparing genomes, integrating data or assessing its quality and consistency across experiments.

GENCODE has responded to these advances by a careful process of evaluation in pilot mode to determine whether and how a specific technology can benefit its annotation goals. For example, proteomics data and computational methods have been instrumental in decisions identifying and classifying protein-coding loci (see Progress Report). Once these annotation processes have been evaluated and the most efficient approach determined, they are made part of the main annotation production and become part of GENCODE's established procedures. As described above, we will follow this pilot project style approach to determine how best to incorporate new data types and changing understandings about the extent to which regulatory regions are, in fact, actually a part of the genes that they regulate. We will report progress on our pilot projects to the GENCODE Scientific Advisory Board and seek their input on required future directions throughout the duration of the grant. We anticipate that this approach will ensure that GENCODE is as valuable as possible for genome interpretation well into the future.

Meeting the community's needs

Most members of the community want to be able to access, view, download and use the GENCODE annotation as part of their work. Our deep integration with the Ensembl and the UCSC Genome Browser helps to make this easy in a wide variety of ways that suit many existing and common workflows. However, changes anticipated in this application including the growth of population transcriptomic data and a possible wider adoption of graph-based genome representations will require significant development in both visualization methods and data access tools for the genome browsers to continue to serve data to in an efficient and meaningful way. Although the browsers are funded almost totally outside of this proposal as part of the core UCSC and Ensembl funding, we describe these below for completeness and to give a sense of the synergy between GENCODE and other genomics resources.

Finally, because the GENCODE gene sets are foundational to so much of genomics, GENCODE has established mechanisms for responsive engagement with the community and many people use them. We commonly accept specific feedback, conduct targeted reannotation and answer questions concerning GENCODE tools and processes both for individual user requests and in more formal training sessions. In fact, we have made significant changes and additions to GENCODE over the past four years in response to community requests including the creation of the GENCODE Basic set and the display and distribution of specific supporting and QC data alongside the transcript annotations.

We will also describe steps to support and appropriately credit annotation done by experts outside the GENCODE consortium. We anticipate that most external annotation that we incorporate will arise from researchers whose work focuses on no more than a few loci. To help these experts contribute annotation, will we intended to conduct workshops and training as requested. Of course, researchers may prefer to email us with reference to a paper containing updated annotation and, when this happens, these will be accepted and prioritized as a targeted annotation request by the GENCODE team.

GENCODE

A comprehensive knowledge of the location, structure, and expression of genes in the human genome is central to our understanding of human biology and the mechanisms of disease. Similarly for mouse, a comprehensive high quality gene set will aid in the design of experiments and the interpretation of the effects of gene knockouts and resulting phenotypes. Also, since mouse is used as a model of human, knowledge of its genes and their relationship to human genes will help inform human gene function.

The GENCODE consortium has assembled a team of world experts in a variety of fields related to gene annotation to create and distribute this gold standard. We have been collaborating since 2003, and have expertise in: gene and transcript isoform identification, pseudogene evolution, sequence conservation, gene expression, proteomics and post-translational modifications, gene regulatory elements, development and maintenance of the infrastructure required to create genome annotation at scale, and demonstrated community engagement and leadership.

Progress report

Since April 2013 we have made eight GENCODE releases for human (referred to by version numbers v17 to v24) and nine for mouse (M1-M9). A history of GENCODE releases and supporting information is available at gencodegenes.org.

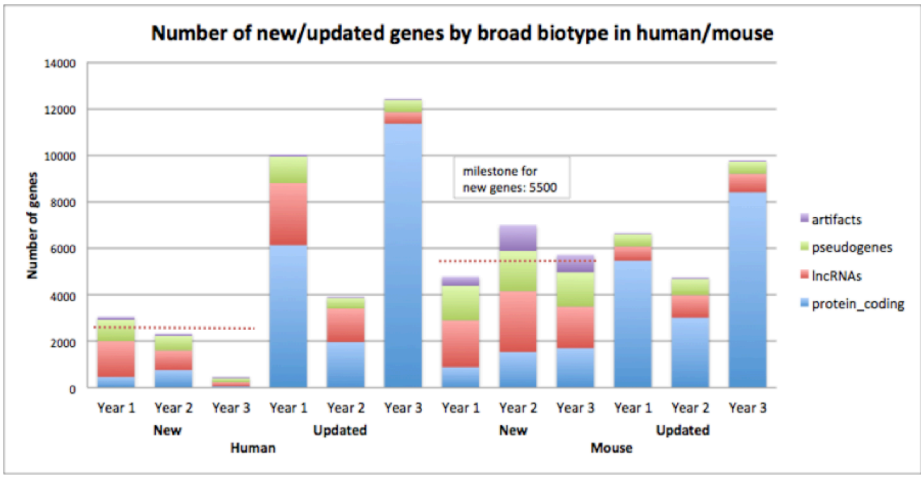


Figure 3: Annotation completed since the start of the grant against our milestones.

Comprehensive annotation of the human and mouse gene sets Human updates have focused on targeted lists of features identified as requiring manual annotation, while mouse updates have focused on chromosomes lacking comprehensive manual annotation, with an emphasis on extending the annotation of pseudogene and lncRNAs. Manual annotation has been enhanced by integration of novel data types such as polyAseq, FANTOM CAGE[34] and Ribo-seq to correctly identify transcription start sites, re-evaluate 3' UTR extensions and investigate translation. ENCODE 454[35] and PacBio[36,37] data are beginning to be incorporated to improve the annotation of lncRNAs. Figure 3 shows annotation completed since the start of the grant against our milestones, categorized by protein-coding, lncRNA and pseudogene.

The number of human protein-coding genes decreased significantly from v19 to v21 due to reanalysis done by the CNIO group in collaboration with HAVANA[38]. More recently, protein-coding gene numbers have increased through two analyses. First, a reannotation of putative coding features generated by PhyloCSF, which, aided by the CodAlignView visualization tool, added more than 100 novel protein-coding loci. Second, a reprocessing of three large-scale publicly available human proteomics datasets, made up of over 54 million mass spectra[39-41], found evidence to confidently support the addition of only 16 novel proteins to GENCODE and evidence for alternative splicing in 867 genes[42] that are under further investigation.

The review of changes to the human and mouse protein coding set continues to utilize the forum provided by CCDS collaboration between WTSI, EBI, HGNC, MGI and RefSeq. Around 50 protein-coding genes are re-examined by the teams each month and discussed.

Pseudogene analysis We conducted systematic analyses of human pseudogenes focusing on large pseudogene families[43-45] and particular types of pseudogenes such as unitary[46] and polymorphic pseudogenes[47]. The latter are peculiar pseudogenes with a dual behavior – the sequence is disabled in the reference genome but in some individuals, it encodes a functional gene. Using the RNA-seq data from the 16 tissues in Human BodyMap[48], we investigated the expression pattern of pseudogenes and found that only 3% of transcribed pseudogenes are expressed in all the 16 tissues, while the other pseudogenes show different degrees of tissue specificity. More than 50% of them are transcribed in one tissue only.

Experimental validation To systematically assess the quality of the GENCODE human gene set, we experimentally verified the structure of transcripts rated as novel or putative using RT-PCR-Seq. We tested 1,243 exon-exon splice junctions not supported by ENCODE or GTEx RNA-seq data, confirming support for 53%. In an equivalent analysis for 3,148 exon-exon junctions in mouse we confirmed 49% of targeted exon-exon structures.

To assess the completeness of lncRNA annotation using 3' and 5' RACE followed by 454 sequencing (RACE-seq), we initiated a pilot experiment targeting 400 GENCODE-annotated lncRNA loci lacking CAGE and GIS-PET support for 5' and 3' ends in seven different human tissues. This led to the addition of approximately 2,600 previously unknown alternatively spliced transcripts with nearly 48% of the 5'-extended transcripts overlaping a CAGE cluster, and 51% of 3' extended transcripts containing polyA sites. Together this indicated that RACE-seq is successful in identifying TSS and TTS. A later effort modified the RACE-seq protocol with PacBio long read sequencing and preliminary analysis of a pilot experiment targeting 541 loci from the UK Genetic Testing Network (UKGTN) in human testis and brain showed presence of approximately 8% novel canonical splice junctions in targeted genes.

To extend this workflow we used RNA Capture-LongSeq, a methodology based on RNA capture coupled with PacBio long read sequencing[31]. We targeted the entire set of GENCODE annotated lncRNAs in human and mouse plus a substantial number of other genomic elements, including other ncRNA classes (miRNAs, snoRNAs and snRNAs), enhancer elements and ultraconserved elements. Using RNA from 4 common tissues in both species (brain, heart, testis and liver), two ENCODE cell lines (HeLa and K562), and two mouse fetal time points (embryo 7d and 15d), we generated Circular Consensus (CCS) PacBio reads that led to the discovery of almost 100,000 completely novel, high-quality canonical introns in human, and more than 50,000 in mouse. These data increased the number of splice junctions in the lncRNA target annotations by 173% in human and 133% in mouse. We developed a novel analysis method to combine the CCS reads with polyA maps with FANTOM CAGE data, which has contributed to confident annotation of full-length lncRNA transcripts. All together there are 65,000 full-length human transcript models (42% novel) and 45,000 full-length mouse transcript models (33% novel). In other words, with unique transcripts, we have increased the GENCODE intergenic lncRNA annotations 136% in human and 140% in mouse. The median spliced length of lncRNA transcript annotations ranges from 600 to 1,120 bp. These have been passed on to HAVANA for manual assessment. Inspection of the data confirms that we have discovered a wealth of new structures even in deeply studied and functionally validated lncRNA loci such as *XIST*, *Jpx* and *MIAT*.

Ensuring quality The annotation described above is further analyzed and tagged before public release. Together with Ensembl, UCSC has produced a set of Transcript Support Levels (TSL) based on whole transcript support from sequences from the International Nucleotide Sequence Database Collaboration (INSDC). Both BLAT and Exonerate alignments of the INSDC mRNAs and ESTs are utilized and annotations are assigned one of five levels, where level 5 is the lowest and indicates no support. In GENCODE v24, 215,072 of the 218236 transcripts have a TSL annotated, including 45,610 (21%) at TSL level 1 and 35,244 (16%) at TSL level 5. These levels are provided to HAVANA to prioritize review or removal of transcript models as appropriate and can also be viewed in Ensembl's Transcript Table

on each Gene Summary page or highlighted in the UCSC Genome Browser and used for filtering the annotation. Ensembl and HAVANA also collaborate to improve the quality of GENCODE, particularly by identifying and removing low quality models generated through automatic annotation.

Annotation of the new human reference assembly Following the release by the Genome Reference Consortium of the GRCh38 reference human genome, GENCODE v20 was created by merging Ensembl's full re-annotation of the new assembly with HAVANA's mapping of GENCODE v19 to the new assembly, including manual annotation of the 261 alternative loci that are part of GRCh38. To address the slow migration of researchers from GRCh37 to GRCh38 and provide those still using GRCh37 access to improvements in the GENCODE gene annotation, UCSC and HAVANA developed a methodology for mapping the GENCODE gene set to previous assemblies. We have made releases of GENCODE v23 and v24 mapped to GRCh37 and these can be downloaded from gencodegenes.org.

Improving usability of GENCODE Ensembl added support for the GENCODE Basic gene set as well as TSL and APPRIS tags from GENCODE v21 and M3. From GENCODE v22, the UCSC Browser adopted GENCODE as their primary human gene set to replace "UCSC genes." Other improvements include updates to the cross-referencing system to UniProt to increase accuracy, adding functionality to the Ensembl BioMart (www.ensembl.org/biomart/martview) so that APPRIS data can be queried, and a simplification of Sequence Ontology terms linked to each gene/transcript.

Summary of GENCODE data types and plans

GENCODE is the reference annotation for the human and mouse genomes. As just described, huge strides have been made by the consortium, including the completion of first-pass manual annotation of the human genome. However, there is still much more to be done for both human and mouse. Many of the annotated human genes are necessarily incomplete because the sequence data available at the time of annotation was incomplete. In particular, the CDS of over 30% of the annotated protein-coding transcript isoforms are known to be 5-prime or 3-prime incomplete. New transcriptome data types such as SLRseq and RAMPAGE (including unpublished data available from the ENCODE portal) are able to identify many novel splice junctions that have yet to be annotated, including which tissue they are functional in and which regulatory mechanisms control their expression. Non-protein-coding genes are poorly understood in comparison to protein-coding genes: many loci are still missing from the annotation set and those that are there tend to suffer from underannotation and misannotation (see **Partial annotation and underannotation** section below). Beyond the coding and non-coding genes, GENCODE creates reference pseudogene annotation. Pseudogenes have long been considered nonfunctional elements, however recent studies indicate that they can be transcribed, translated and can play key regulatory roles. In particular some pseudogenes regulate the expression of functional protein-coding genes by serving as a source of siRNAs, antisense transcripts, microRNA binding sites, or competing mRNAs[49,50]. These data types, including completing full first-pass manual annotation of the reference mouse genome assembly, are the focus of **Aim 1**.

The pseudogenization process is also closely linked to loss-of-function (LOF) events such as premature truncation of proteins, disruption of splicing and loss-of-functional or structural domains[2,47,51]. Pseudogenes thus provide valuable opportunities to study the dynamics and evolution of gene functions. The annotation of pseudogenes is important in the analysis of personal genomes as they provide a means to avoid errors in genotyping assays and variant calling. In addition, we know from the 1000 Genome Project that the current reference human genome is unable to describe the full complexity of variation observed across all human populations and, therefore, unable to represent the full complexity of transcripts and gene elements present in human populations. Efforts are underway in the Genome Reference Consortium (GRC) to expand the definition of the reference human genome to include genomic sequence for all haplotypes and gene alleles. GRC have already committed to supporting the genomes of a collection of 16 representative strains, thus effectively replacing the linear genome with a "graph-like" structure of 16 separate haplotypes. GENCODE already annotate the full reference genome for human and mouse, including all available alternate sequences. As this reference

genome expands, GENCODE will continue to provide annotation appropriate to these new genomic sequences. This is the focus of **Aim 2**.

In addition to genomic mutations that impair gene product function, many diseases are actually caused by the faulty regulation of an otherwise healthy gene product Therefore, as our goal of completely annotating all transcript isoforms nears completion, the next step is to identify the regions regulating each gene. We will pilot the annotation of tissue-specific gene regulatory regions in our **Aim 3**.

The GENCODE data production and curation processes

Comprehensive gene annotation pipeline

The manual gene annotation process remains central to the GENCODE project. Manual annotation of protein-coding, long non-coding RNA and pseudogene loci for the GENCODE human and mouse gene sets is carried out according to the guidelines of the HAVANA group[52]. Historically the HAVANA group produced transcript models largely based on the alignment of EST and cDNA sequences from the INSDC and protein sequence data from UniProt. These sequences were aligned to the individual BAC clones that make up the reference genome sequence using BLAST[53], with a subsequent splice-aware realignment of transcriptomic data by Est2Genome[54]. The core GENCODE process (Figure 2) builds on this established method, and starts with a diverse set of data types that have either been aligned to the reference genome assembly or calculated via one of the several comprehensive annotation pipelines viewed in the ZMap annotation interface (http://www.sanger.ac.uk/science/tools/zmap). Depending on the species and the locus there may be more than 400 datasets available to manual annotators including: gene and pseudogene predictions (including pseudogene predictions from the PseudoPipe[25] and Retrofinder[55] pipelines), cross-species conservation, transcription start and termination sites, regulatory features, mass spectrometry and Ribo-seq data, second (i.e. Illumina) and third (i.e. PacBio) generation RNA-seq data and transcript models and splicing feature predicted from them. Although the output of GENCODE is the final set of annotations, this collection of supporting evidence represents the full breadth of the data currently available within the GENCODE resource.

Functional classification GENCODE genes are assigned a “biotype” associated with one of four broad categories; protein-coding gene, lncRNA gene, small ncRNA gene or pseudogene. Genes derive their biotype from the biotypes assigned to their constituent transcripts and are assigned based an defined series of rules[52]. Briefly, all newly created transcripts and loci are initially assessed to determine their protein-coding potential and the assignment of a non protein-coding biotype is only made when the possibility of coding potential is eliminated. The protein-coding potential of transcripts is determined on the basis of similarity to known protein sequences, the sequences of orthologous and paralogous proteins, the presence of Pfam functional domains[56], clear support of high quality peptides from mass spectrometry (MS) experiments and good evidence of translation from Ribo-seq data. The broad pseudogene biotype definition has multiple subdivisions that describe the mechanism of creation and transcriptional status of the locus. Long non-coding RNA loci are generally more than 200 bases long and require evidence of transcription from EST, cDNA or RNA-seq datasets. They lack any features associated with protein-coding potential described above. Given our current inability to infer the functional potential or mode of action for most lncRNA loci, biotypes are assigned on the basis of genomic position relative to protein-coding loci. For example, antisense transcripts overlap the genomic span of a protein-coding locus on the opposite strand, and lncRNA transcripts are intergenic to protein-coding loci. The lncRNA annotation produced by GENCODE represents a core dataset underpinning the RNA Central lncRNA dataset[57].

We use a controlled vocabulary of attributes to describe important features of transcript and gene annotation that are not captured in other fields. For example, a transcript supported by transcriptional evidence not derived from the same organism is tagged with the attribute ‘non-organism supported’, while a transcript that contains a non-canonical splice site that has been checked and retained in the gene set because it is supported by cross-species conservation, is tagged with the attribute ’non-

canonical conserved'. All attributes may be queried to facilitate the filtering of their associated transcripts and loci.

Updates to annotation: missing loci One of the key objectives of GENCODE to represent all gene loci in human and mouse. Transcriptomics and proteomics data are the predominant data types used to identify previously unannotated loci. Transcribed loci are identified by RNA-seq based gene annotation from Ensembl, ENCODE collaborators and good quality public methods such as PLAR[58], and targeted for manual annotation. Given questions over quality of RNA-seq based transcripts[59], we initially look for intersection of ≥1 intron between RNA-seq based transcripts and splicing ESTs, cDNAs or third generation RNA-seq transcripts. Where this is absent overlap between introns from ≥2 independently created RNA-seq based transcripts is sufficient to support annotation. We will continue to investigate transcription identified in RNA-seq datasets from previously inaccessible tissues and development stages, public third generation transcriptional evidence such as SLRseq[13] and Capture-Seq PacBio data.

We use cross-species conservation information from PhastCons[23] and specific conservation of protein-coding sequence from PhyloCSF[22] to identify novel gene loci. For example, detailed investigation of a refined set of thousands of high scoring PhyloCSF regions generated across the whole human genome yielded more than 100 novel protein-coding loci, many of which had very low expression support in human RNA-seq datasets. We are currently investigating the equivalent dataset in mouse and expect it to be more fruitful given the less complete state of the mouse annotation. Although many of the novel loci have orthologs in both species, we have identified instances where a gene has been lost in one lineage, emphasizing the importance of independent analysis in both species. PhyloCSF frequently highlights unannotated pseudogene loci. More than 200 unitary and unprocessed pseudogenes have been identified suggesting that there remain many unannotated pseudogenes in both human and mouse genomes. Protein-coding loci are also identified by analysis of reprocessed MS data from large-scale public shotgun proteomics datasets. In human this has led to the identification of 16 novel protein-coding loci[42]. Mouse shotgun proteomic datasets are significantly smaller than human but the availability of samples in tissues and developmental stages inaccessible in human may support identification of novel loci. We will continue to use large public proteomics datasets and cross-species conservation information to identify putative novel protein-coding loci and pseudogenes.

Partial annotation and underannotation In GENCODE v24 there are more than 10,000 alternatively spliced (AS) transcripts at annotated protein-coding genes tagged with the "processed transcript" biotype signifying that they cannot be annotated with certainty as protein-coding or nonsense-mediated decay (NMD). A further 33,000 partial protein-coding transcripts are tagged as incomplete (either start or end not found). For non-coding genes, all evidence suggests that the vast majority are currently incomplete[19]. For example, the expected 5' CAGE clusters were found for only 15% of annotated lncRNA genes compared to 55% of annotated protein-coding genes, a difference that persists, albeit reduced, when gene expression is controlled for. Both RACE-seq and CaptureSeq will improve this.

Addressing the annotation of these missing AS transcripts at annotated loci and extending partial AS transcripts to reflect their full length thus remains an important goal of Aim 1. Moreover, adding missing exons and splice junctions is essential in providing the best possible foundation for downstream analysis and interpretation including for applications to population transcriptomics in Aim 2. Even where all exonic sequence is annotated, the accurate extension of all transcripts to full-length is essential to describe their connections. It is also vital to extend all transcripts to full length to allow the proper interpretation of the functional potential of a transcript and promoter. Those transcripts associated with protein-coding loci but lacking a CDS are considered underannotated, in that their structures are correctly described but functional annotation not added. Full-length transcripts combined with knowledge of transcript start sites (TSS) and transcript termination sites (TTS) give all the required information to make a determination of biotype. Specifically, certainty over the TSS allows the translation initiation site (TIS) to be determined and identification of the TTS provides important context

for the position of the stop codon and whether any premature stop codon (PTC) would be likely to trigger NMD.

It is difficult to estimate the precise number of unannotated AS transcripts even within protein-coding loci although it is likely to be large. Recent reannotation of 70 genes on a clinical panel for Early Infantile Epileptic Encephalopathies (EIEE) using PacBio, SLRseq[13] and RNA-seq[60] datasets from brain led to the annotation of 1092 novel AS transcripts, 706 novel exons, 224 novel splice sites in annotated exons and more than 141kb of additional exonic sequence, of which 15.2kb was novel CDS.

It is also essential to extend transcripts of all biotypes to full-length, including transcripts with NMD and retained_intron biotypes as recent studies have shown that transcripts of both these biotypes have been implicated in the post-transcriptional regulation of the genes with which they are associated[61-64] and disruption of their splicing has been associated with disease [65].

While almost all protein-coding loci have at least one full-length transcript, many annotated lncRNA loci do not[19]. Extending transcripts at lncRNA loci allows additional exons to be identified, and allows a more informed determination of transcript and locus biotype, for example confirming that currently annotated lncRNAs are fully intergenic.

Many missing and partial AS transcripts will be detected using third generation transcriptomic data such as Capture LongSeq (see below). However, second generation RNA-seq, with its wider pool of cell-lines, tissues and developmental stages remains useful in identifying and annotating novel splicing features, particularly in combination with PhyloCSF, CAGE/RAMPAGE, and polyAseq data.

One consequence of the addition of a great many more transcript models and the extension of all transcript models to full-length is that the definition of the GENCODE Basic set, which currently contains only full-length transcripts, will need to be redefined. Additionally, we will annotate transcripts or individual exons (where the data may be more reliable) to allow identification and ranking of those most likely to have functional potential. To do so we will integrate multiple datasets: transcriptomic and CAGE/polyAseq data to determine expression level by tissue/cell type/developmental stage; Ribo-seq, targeted and shotgun MS datasets will be used to confirm translation; individual components of the APPRIS pipeline will be used to support protein-coding potential; while cross-species sequence conservation and variation data can be used to identify transcripts that include regions under selection/constraint.

Annotation toolkit The manual annotation toolkit (Figure 2) comprises four main components: ZMap, Otter[66], Annotrack[67] and Seqtools[68]. The ZMap annotation tool is the primary interface annotators create models based on primary evidence alignments and stacked displays to a genomic region. The Otter database stores and tracks manually annotated gene models. Annotrack is a system for recording and prioritizing genes that require manual annotation. Blixem and Dotter from Seqtools provide interactive analyses of sequence alignments to the genome.

ZMap is being extended to support attachment of large data files (BAM, BigBED, BAM, CRAM, VCF and BCF) and will result in a reduction in time required to integrate new data sources. In addition ZMap supports the UCSC Track Hub format so that providers can create data collections alongside valuable metadata used to flag the source and type of data. We also plan extensions to the Blixem software to support additional alignment views including sashimi plots. ZMap will also be expanded to display additional data types such as cis-regulatory and physical interactions in support of Aim 3.

Integrated approach to pseudogene identification and classification

Depending on their formation mechanism, pseudogenes can be referred to as unprocessed (originating through a gene duplication event) or processed (originating through a retrotransposition event). A functional gene may also become a pseudogene by acquiring a disabling mutation, if its function no longer confers a fitness advantage to the organism due to a change in the environment or genetic background. Such pseudogenes are called unitary pseudogenes.

The Yale group has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and HAVANA group, we use a variety of methods to identify pseudogenes[20,25,26].

PseudoPipe Yale's automatic annotation pipeline, is fast and accurate[20] (Figure 4). The pipeline takes as input all known protein sequences in the genome and using a homology search is able to identify disabled copies of functional paralogs (referred to as pseudogene parents). There is a good consensus overlap between the human pseudogene prediction set obtained with PseudoPipe and the set manually curated by the GENCODE annotators. Even more, the PseudoPipe predictions fueled the manual curation of pseudogenes in GENCODE[20].

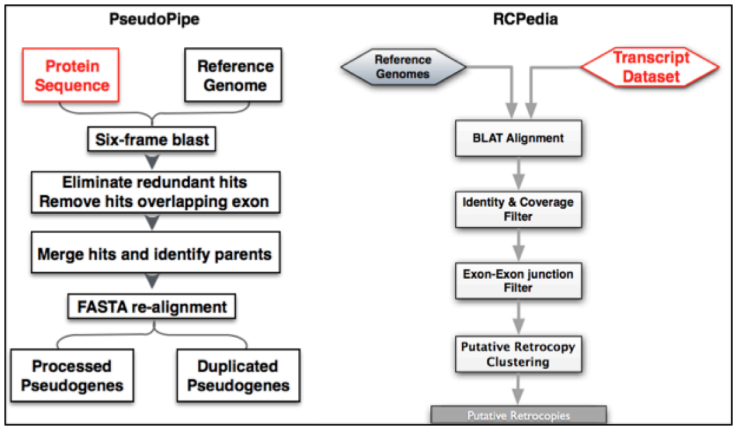


Figure 4: Automatic pseudogene annotation pipelines

RCPedia the newest pseudogene annotation pipeline focuses on the annotation of retrotransposed (processed) pseudogenes[27] (Figure 4). This pipeline takes as input all known protein-coding RNA transcripts and using sequence alignment is able to identify all possible retrocopies of protein-coding genes. Putative retrocopied sites are identified based on exon-exon junction information and direct repeats flanking the event. In the human genome there is an over 85% consensus between processed pseudogenes predicted by RCPedia and those annotated using PseudoPipe.

Retrofinder is the UCSC retrocopy annotation pipeline. Retrocopies can be functional genes that have acquired a promoter, non-functional pseudogenes, or transcribed pseudogenes. Retrofinder finds retrotransposed messenger RNAs (mRNAs) in genomic DNA[55]. Candidate retrocopies overlapping by more than 50% with repeats identified by RepeatMasker[69] and Tandem Repeat Finder[70] are removed. Retrocopies are identified based on a scoring function using a weighted linear combination of features indicative of retrotransposition.

The 3 pipelines will continue to be used to identify pseudogenes in human, mouse and other model organisms[20,26,71] and the structural and functional relationships between the pseudogenes within a gene family described by a pseudogene ontology[72].

Functional characterization of pseudogenes By integrating functional genomics data such as that generated by ENCODE, with available annotation we will obtain a comprehensive map of pseudogenes activity in mouse including transcription signals for some pseudogenes and other biochemical activity such as presence of transcription factor or polymerase II binding sites in the upstream region, active chromatin, etc. Based on previous analysis only 5% of the total number of pseudogenes are broadly expressed in human although, in general, transcribed pseudogenes show higher tissue specificity than protein-coding genes[71], but given the variability of pseudogenes across species, mouse may have different patterns.

Computational methods to evaluate and enhance gene annotation

The Ensembl GeneBuild In parallel to the manual annotation process described above, the Ensembl gene set is created and updated. The Ensembl GeneBuild creates genome-wide annotation quickly and consistently, with thousands of genes annotated in parallel[28]. The GeneBuild contributes gene annotation for regions of the genome that have not benefited from manual curation, gene types that are not manually annotated e.g. small non-coding RNA genes, and provides rapid access to novel transcript isoforms that are identified from new data in the archives.

The Ensembl annotation for human and mouse will be updated when the GRC release a major or minor assembly update or when there are new data sets of major importance. Major assembly updates, such as the update from GRCh37 to GRCh38 result in a change of the chromosome coordinate system, and will trigger a comprehensive update. Minor assembly updates, such as the update from GRCh38.p7 to GRCh38.p8, simply add additional alternate sequence alongside the primary assembly. These minor updates will be annotated quickly and are valuable for providing new genomic sequence that corrects errors identified on the primary chromosomes or novel haplotype sequence not represented on the primary chromosomes.

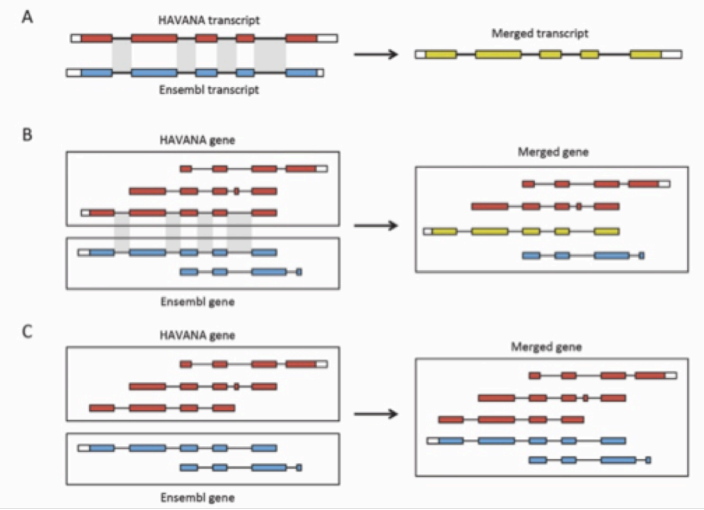


Figure 5: For both Ensembl and HAVANA models, transcripts with overlapping exons are grouped together into genes. A: If the intron-exon boundaries, excluding UTRs, of a transcript from HAVANA completely match those one from Ensembl the result is a merged transcript model based on the HAVANA annotation. B: Exons for a HAVANA gene overlap with those for an Ensembl gene. All transcripts are grouped together in the same merged gene. C: No transcripts with complete matching intron-exon boundaries results in transcripts together into a merged gene but no transcripts are merged.

Ensembl/HAVANA Merge. We will continue to create the final, comprehensive GENCODE gene set by supplementing manual annotation from HAVANA with Ensembl annotation described above. This merge process, described in detail by Harrow et al[18], is performed genome-wide and involves pre-merge quality checks and comparison of all Ensembl transcripts against all overlapping HAVANA transcripts (Figure 5). Where the splicing structure of the Ensembl transcript matches the HAVANA transcript, they are merged and the alignments supporting the Ensembl annotation are combined with the HAVANA data. Novel genes and transcripts contributed by Ensembl are added. The HAVANA biotype takes precedence where data are inconsistent.

PhyloCSF We will seek unannotated protein-coding regions and recent pseudogenes using PhyloCSF, a phylogenetic model based on codon substitution frequencies[22]. In contrast to experimental techniques, PhyloCSF recognizes evidence of functional translation based on its evolutionary signature of selection at the protein-coding level. We achieve this by modeling codon substitution frequencies (CSF), which are highly distinct between true codons in protein-coding exons vs. nucleotide triplets in non-coding regions[22]. PhyloCSF evaluates these codon substitution frequencies in a phylogenetic setting to appropriately account for the relatedness of different mammalian species. Our PhyloCSF pipeline is described as follows.

Annotate putative novel protein-coding regions We will run PhyloCSF on every codon in every frame in the human and mouse genomes, resulting in genome-wide PhyloCSF scores (higher/positive for protein-coding regions, lower/negative for non-coding regions). We will translate these scores into protein-coding intervals using a 4-state Hidden Markov Model (HMM), whose emissions are the PhyloCSF score, and whose hidden states consist of a single protein-coding state (emitting higher PhyloCSF scores) and three non-coding states representing introns, nonconsecutive introns, and intergenic regions. We will calculate the most likely sequence of these hidden states (i.e. the Viterbi path) and report putative novel protein-coding regions those sections of the genomes that do not overlap pseudogenes or previously-annotated coding exons in the same reading frame.

Identification of putative novel protein-coding genes and exons for manual curation In collaboration with HAVANA, we have found features of putative coding exons that are more likely to be biologically meaningful. These will be used to develop a machine-learning framework based on four features: the

average PhyloCSF score per codon; the length of the region; the phylogenetic evidence quantified as the branch length of the aligned species in the region; the difference between the average PhyloCSF score per codon in predicted frame vs. in the antisense frame. We will then train a discriminative machine learning model (a support vector machine, SVM[73]) to discriminate coding regions vs. non-coding regions that do not overlap pseudogenes or antisense exons, and a second SVM to distinguish coding regions vs. antisense regions. We will the group the annotated protein-coding exon predictions into new genes and new exons of existing genes based on their clustering properties on chromosomal segments, and their relationship with existing gene and transcript annotations for GENCODE. We will specifically annotate multiple criteria that distinguish novel coding genes, novel coding exons, and pseudogenes, including: the distance to the nearest annotated coding gene; the branch length of the alignment, which helps distinguish coding genes vs. pseudogenes; the presence of in-frame stop codons far from the exon boundaries, which is more likely in pseudogenes.

Prioritization of novel exons/genes, known exons/genes and pseudogenes Lastly, to help guide the manual annotation process for incorporating new predictions, and also refining existing annotations and potential misannotation, we will calculate the distribution of ranks that already-annotated coding exons would have gotten if not annotated. We will use that distribution to determine a scale factor between known and novel exons, and use them to: (i) set priority thresholds for the manual curation process, (ii) flag potential mistakes in the existing annotations, (iii) estimate the number of novel genes and exons that have not been found after manual examination of first few thousand ranks; (iv) estimate the likely false positive and false negative rate at different rank thresholds.

Isoform analysis The CNIO isoform annotation pipeline (APPRIS; http://appris.bioinfo.cnio.es) uses protein structural and functional features and information from cross-species alignments to annotate alternative splice isoforms[30,74]. APPRIS annotates the likely effects of alternative splicing on protein features, and will select a single CDS as the main (principal) isoform based on these annotations[75]. To date, there data have prompted changes to gene models in more than 300 human genes. The APPRIS principal isoform is generally the isoform with the most conserved protein features and the most evidence of cross-species conservation. APPRIS selects a principal isoform for 73.4% of human genes and 82.1% of mouse genes. For genes in which APPRIS cannot choose a main variant, it selects the main isoform based on CCDS annotations[76] and UCSC TSL. Using information from these two methods APPRIS is able to select a principal isoform for 95.5% of human genes and 96.4% of mouse genes. APPRIS principal isoforms coincide overwhelmingly with the main protein isoform detected in proteomics experiments[77].

Coding gene analysis pipeline The CNIO has developed a methodology for the detection of protein-coding genes with atypical characteristics based on annotations from the APPRIS, Ensembl, GENCODE and UniProt[78] databases. Nineteen features that correlated with lack of protein-level expression, including poor conservation, recent origin, poor supporting evidence and contradictory annotations were used to flag 2,001 coding genes from GENCODE v12 as potentially not coding[38]. Manual annotators have since revisited these genes and 1,026 have been reclassified as either non-coding or pseudogene. The CNIO carried out a similar analysis on coding genes added between GENCODE v12 and v19, leading to the reclassification of almost 500 automatically predicted coding genes. In the most recent analysis (GENCODE v23) a further 2,050 protein-coding genes were labeled as unusual. The pipeline has also been applied to the mouse annotation and the initial analysis identified 4,841 mouse protein-coding genes that were potentially not coding. The CNIO will automate the identification of these unusual coding genes and run the pipeline for each new release of GENCODE.

Transcript support level The Transcript Support Level (TSL) calculation is based on orthogonal sources of information to those used for creating the transcript. This includes comparison drawn from primary data sources, such as GenBank, gene-ortholog comparisons and evolutionary assessment of gene features using conservation patterns. Manual annotations are produced over time, looking at snapshots of evolving primary evidence. By doing a comprehensive, consistent evaluation given the

latest evidence, we will flag and help prioritize problematic transcripts and genes to revisit in the manual annotation process. The orthogonal evidence evaluations we will continue to develop and produce are provided to the manual annotators and the community as TSL scores for each transcript and serve as a metric for users of the GENCODE data set to easily understand the support for a given transcript. We have recently extended this approach to incorporate RNA-seq evidence, which provides a metric for the support of exons in a transcript.

Experimental annotation and discovery

High throughput, complete annotation of novel non-coding RNA transcripts The combination of manual, computational and experimental approaches is a key feature of GENCODE and has been important to its adoption as the reference gene annotation of the human genome. During the initial phases of GENCODE, experimental methods were exclusively aimed to the validation of manually annotated protein coding transcripts. More specifically, splice junctions labelled as putative by the HAVANA team were systematically tested using multiplexed RT-PCR. In the current phase of GENCODE, experimental methods that had been traditionally limited to validation of annotated transcripts have been complemented with discovery methods aimed to recover the full-length structure of genes/transcripts likely to be partially annotated. Thus, in a pilot study (see Progress Report), RACE-seq was employed to target 400 lncRNA loci, leading to the discovery of about 2,600 previously unannotated transcripts.

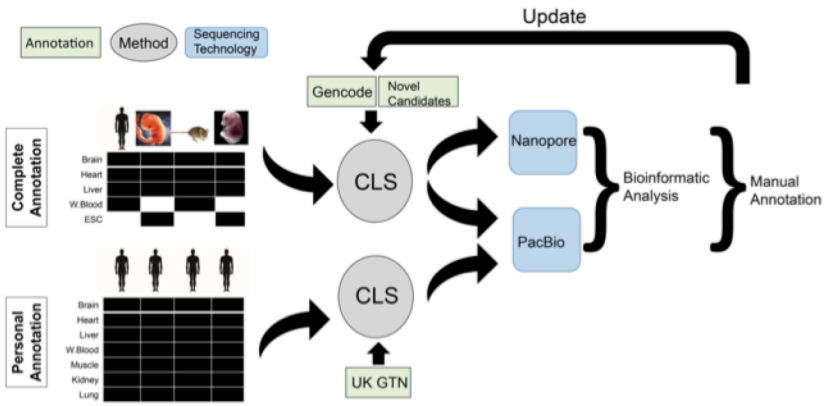


Figure 6: Overview of the RNA Capture LongSeq (CLS) components including complete annotation in matched human / mouse tissues at embryonic and adult time points for Aim 1 and personal annotation of 4 human individuals for Aim 2.

For the next phase of GENCODE we will further prioritize the use of experimental methods for transcript discovery rather than validation. The large and increasing number of RNA-seq data sets available obtained from multiple individuals, tissues and conditions, makes RT-PCR validation increasing dispensable. We found for instance, that most GENCODE putative junctions validated by RT-PCR are also supported by short RNA-seq reads, from the GTEx and ENCODE projects. Building on the results of the current phase, we plan to base the experimental effort of GENCODE on long read sequencing of targeted loci with the aim of inferring the full-length structure of the transcripts encoded in the human and mouse genomes. Towards that end we plan to fully scale the CLS methodology that we have pioneered in the current phase of GENCODE and that combines capture of targeted loci with long read sequencing (see Progress Report).

RNA Capture-LongSeq of GENCODE annotated and unannotated transcripts RNA capture sequencing is a recently-developed method that enables the targeted sequencing of rare transcripts in a complex RNA sample[31]. Starting with a target set of candidate transcribed regions, either previously-annotated or completely novel, researchers can deeply sample their transcripts at a depth that would be impossible with unbiased sequencing. Capture sequencing has huge potential in streamlining the annotation process. The few published methods to date have captured fragmented cDNA libraries, and used Illumina short read sequencing as a readout[79,80]. The consequent reliance on de novo transcript assembly methods means that the quality of the resulting “novel” transcript structures is unclear. We have solved this problem by creating a protocol for capturing full-length cDNAs and using third generation long-read Pacific Biosciences sequencing. Our bioinformatics pipeline will filter, demultiplex and map these data, as well as estimate quality metrics for each experiment. Results from this pilot

constitute the most extensive and reliable set of transcript sequences from targeted loci obtained so far, a unique resource towards high quality annotation.

For the next phase of GENCODE, we will perform CLS with the following specific objectives: 1) exhaustively annotate of all protein-coding and non-protein-coding RNA transcript structures, in human and mouse, at adult and foetal time points, and in all transcript length ranges 200nt and above, in four complex organs; 2) exhaustively annotate of personal/private gene structures within disease-related genes in seven organs from four human individuals (see **Annotation of individual and population data** section). Our plans are shown schematically (Figure 6) and described as follows.

Target annotation definition and capture library design We will target GENCODE annotated loci, with emphasis on lncRNAs, as well genome regions with evidence of transcription and/or protein-coding capacity. Specifically, evidence of transcription will be based on short RNA-seq data from GTEx, ENCODE and other projects, while PhyloCSF regions will be used to define protein-coding potential.

CLS capture and sequencing Custom capture libraries will be designed for the annotations above, and used to capture in RNA samples from brain, heart and liver in human and mouse, as well as in white blood in human, and embryonic stem cells in mouse. Full length, captured sequences will be sequenced by PacBio and Nanopore technologies. For the case of PacBio, this process will be carried out on a yearly basis, using updated and refreshed annotations based on external input and the capture results themselves. Depending on performance of the Nanopore sequencing, we may decide to divert resources from PacBio to Nanopore in later years of the project. For PacBio, we plan 25 SMRT cells per individual sample, equivalent to approximately 950,000 reads based on our previous experience of on average 38,000 reads per SMRT. For libraries targeting 200,000 candidate loci, we estimate this would equates to an average of 5 reads per locus. Given that a significant fraction of loci in our candidate set will not be transcribed in any given sample, and given that eight distinct cell samples will be probed per feature, we believe this approve adequate will provide adequate read depth.

Bioinformatics analysis Reads will be filtered and aligned to extract TSS and TTS information, compare splice junctions to Illumina short reads, and merge putative transcript models. The resulting models will be to HAVANA for incorporation into GENCODE. The transcript models thus generated will be sorted into categories of confidence. High confidence structures, with identified 5' and 3' ends, and all splice junctions supported by Illumina short reads, will require minimal annotation effort. Novel structural features with lower confidence will be flagged and annotated with more care. This experimental effort represents an intermediate step in the GENCODE annotation process, and the two parts together will create a cycle of annotation extension and refinement that can be carried out multiple times. We believe that the CLS approach will greatly facilitate the work of the annotators and enhance the GENCODE annotation in a very cost effective and comprehensive way.

Proteomics Proteogenomics, in which proteomics data from mass spectrometry is used to interrogate genomic sequence data for genome annotation, is a growing field[81-83]. GENCODE will apply quantitative shotgun tandem mass spectrometry to characterize the proteome using established protocol, workflow and analysis methods. To attain deep sequencing and accurate quantitation we will apply approaches for sample fractionation, together with multiple enzymatic digestions for peptide generation and repeat mass spectrometry analysis such that we will quantify >10000 proteins per proteome. We will use synthetic peptides as reference standards to align and compare between sample sets as well as to target specific gene features.

The complementary proteomics expertise and pipelines from the CNIO and the WTSI, together provide experimental confirmation by generating dataset internally as well as dedicated capacity for analyzing data sets external to the consortium. All GENCODE proteomics data generation will be on the same biological samples as the CLS study described above. Protein samples will be processed according to established shotgun proteomics protocols and analyzed by high-resolution tandem mass spectrometry. These analyses feedback to improve annotation in Aim 1and will be valuable for components of Aim 2.

Although comparisons between GENCODE and known coding databases such as UniProt and RefSeq suggest that relatively few coding genes are still missing from the human reference set, this approach remains a profitable strategy: 61 missing coding genes have been added to the reference set from the CNIO proteomics analysis that uses established sequence databases and a further 16 coding regions were added by WTSI using a compendium of genomics and prediction databases[42]. More generally, while recent large-scale experiments have identified peptides for more than two thirds of genes[84,85], reliable proteomics data identifies only a fraction of annotated alternative isoforms, which suggests that most protein-coding genes have a single main protein isoform[77], although a systematic search for alternative isoforms has not yet been conducted.

The WTSI have used OpenMS as a platform on which to build pipelines for the analysis of large-scale proteomics datasets[42] and will deploy their pipeline to process public data focusing on the mouse tissues in the first instance. The pipeline can also be used for personal annotation by uploading associated DNA or RNA sequencing files as reference database (see below) and includes a priority annotation score to identify peptides that are more likely to lead to novel annotation[42]. The output results can be formatted in common genomics file formats (GTF, GFF, BAM, BED) that are directly useful to the manual annotators. These peptide mappings, which can additionally include peptide abundance, modification, and uniqueness information, can be then loaded into Ensembl or UCSC or used within a proteomics Track Hub.

The CNIO proteomics analysis has been run for the GENCODE v3C, v7, v12, v20 and mouse M2 releases and will continue to be run on a regular basis in the future. In each case, spectra from multiple different experiments and databases are analyzed with stringent filters to improve the reliability of the identifications. For example, **t**he GENCODE v20 analysis[85] employed eight datasets (including the spectra from the recent Nature papers[39,40]) and to identify 277,244 peptides that mapped to 12,716 coding genes (64%). The mouse M2 analysis used three datasets and identified 12,000 genes[85].

Coordination with related data resources

RefSeq We coordinate with RefSeq, HGNC and MGI via the CCDS collaboration (LOS Pruitt). This forum holds monthly conference calls to discuss release cycles and difficult annotation cases.

UniProt The UniProt archive hosts the current protein sequences records for human and mouse and is updated on a monthly basis. GENCODE and UniProt have a close collaboration including an active mailing list where we share information regarding missing proteins, protein variants, and dubious protein records with no support. In this way, GENCODE and UniProt are working to converge on an agreed proteome for human and mouse, with results to be stored in a database hosted at EMBL-EBI.

Gene Expression Atlas GENCODE and EMBL-EBI Expression Atlas team have a history of collaboration (LOS Brazma). The Gene Expression Atlas imports GENCODE annotation and processes transcriptome data from large published datasets, the results of which are provided back to GENCODE and can be used directly to provide support for GENCODE's annotation.

Sequence Ontology We are currently working with the Sequence Ontology (SO) consortium[86] to identify the best SO terms relating to our broad gene biotypes. Having achieved this, we will extend our integration with SO to our more detailed locus and transcript level biotypes and then to attributes, using appropriate existing SO terms where possible but modifying current or creating new SO terms as necessary. Integration with SO will enable computational reasoning across GENCODE annotations and make it easier to integrate them with other information described with compatible ontologies.

Creating the GENCODE resource

The creation, advancement and maintenance of the GENCODE resource requires both adherence to and optimization of defined processes that ensure the genome annotation created going forward remains at or above the high standards that we already achieve. We must also be attuned to the new technologies and opportunities that arise as the field of genomics evolves. The GENCODE consortium is extremely well equipped for this task. Over the past 13 years from before the commercialization of next generation sequencing and through the enormous growth in genomics over the last decade, the GENCODE annotation has moved continually forward becoming more complete and simply better with time. This section will describe our plans to continue this trajectory and ensure that GENCODE in 2020 will be significantly more valuable for research and clinical applications in genomics than it is today.

We will start with a description of our quality control (QC) process that extend across all activities in GENCODE and support all four Aims and then describe how we maintain infrastructural and data stability. We will spend the majority of this section describing our plans for improving GENCODE, which include continuing our world leading manual annotation pipeline with additional input from newly generated experimental data (Aims 1 and 2). We will also describe two pilot projects that will help define the most effective way to support future GENCODE annotation: a graph genome representation as part of Aim 2 and a effort to connect genes to regulatory regions that is the whole of Aim 3. We will end this section with plans to ensure that GENCODE scales to meet the project requirements.

Quality control

Manual annotation QC To assess possible structural misannotation (i.e. inclusion of unsupported introns or exons) HAVANA confirm the structure of features from aligned transcriptomic datasets. Manual confirmation is particularly important when supporting sequence evidence is limited or does not align perfectly to the genome. While cDNAs and ESTs were traditionally used to identify introns with little or no support, we are now using RNA-seq data from multiple tissues produced by the GTEx project. We will tag transcripts containing these introns and remove them from the gene set. As the project progresses, will use the increasing amounts of third generation transcriptomic data to confirm complete transcript structures as the scope and depth of such datasets permits.

The quality of functional annotation (i.e. the biotype that is assigned to transcripts and loci) is assessed in three ways. Biotypes of all transcripts are compared to other transcripts at the same locus to identify aberrant combinations. To identify unannotated protein-coding transcripts, PhyloCSF regions and proteomics data intersecting with lncRNA annotation are checked as part of CCDS discussions or following literature review. To confirm true protein-coding functionality of annotated transcripts we will utilize coverage by shotgun and targeted proteomics and ribosome profiling data. Quantitative proteomics will also give insight into the likely functional significance of loci.

Validating protein-coding genes The CNIO will continue to validate coding genes by searching against databases of known coding sequences (UniProt, RefSeq) to which we will add small numbers of likely coding variants. This protocol generates small numbers of novel coding genes with a high hit rate and employs manual inspection to separate spectra that identify novel coding regions from false positive matches. While the protein-coding set of GENCODE human genes is close to complete, the isoform annotation is not. A comparison of the UniProt and GENCODE reference sets showed that only half the genes had the same main isoform. For the analysis of human genes and isoforms CNIO will employ a pipeline that reanalyses spectra from eight different large-scale proteomics using two different search engines and rigorous quality controls that we have already published[77,85]. Peptide-spectrum matches that identify novel coding isoforms (such as isoforms present in databases other than GENCODE, alternative isoforms not previously recognized in proteomics experiments or isoforms predicted from PacBio transcripts) will be validated manually.

Proteomics QC The CNIO has shown that great care must be taken to avoid false positives when using data from large-scale proteomics experiments[38,77,87], although there is currently no reliable strategy to estimate false positive rates in large-scale proteomics experiments[88] for technical reasons

such as the narrowness of the mass precursor windows in the newest high-resolution mass spectrometers[89,90], the difficulty of identifying post-translational modifications[90] and the multiplication of errors when smaller experiments are combined into one[91]. These problems are especially critical when the experimental evidence is used to verify variant translation. The CNIO's partner, the Spanish National Center for Cardiovascular Research – CNIC, is investigating the feasibility of improving methods for calculating false positive identifications in large-scale proteomics experiments. In the meantime, we feel there is no substitute for the manual inspection of all spectra that identify previously unidentified genes and variants. The CNIO and CNIC will perform this manual verification, in order to guarantee the reliability of peptide detection in proteomics experiments.

Final GENCODE gene set QC Before final release, Ensembl compares the GENCODE gene set to other sets (UniProt and cDNA alignments, and imported RefSeq data) to check for missing genes or transcripts. GENCODE is also compared to the most recent CCDS release to ensure that it is a proper superset of the CCDS models. The alignments of cDNAs in the INSDC are updated ever 2.5 months for each Ensembl release and if any annotation in these external datasets is missing and requires manual annotation, it will be stored in AnnoTrack to ensure the HAVANA annotators inspect these loci.

Maintaining GENCODE's infrastructure

GENCODE stability applies at multiple levels. At the most fundamental level we ensure our computational infrastructure is well maintained to support our efforts, that adequate processes are in place to ensure the highest possible quality and that annotation is made freely available in easy to consume high value formats. Continued stability is key to all of GENCODE's overall goals.

Data release We will continue our current release schedule of 4-5 releases per year. Each release will represent a merge of frozen data sets from the GENCODE consortium members. All models will be cross-referenced and assigned stable IDs according to Ensembl's rule set[92] before release. Briefly, stable ID mapping involves identifying overlapping models, transferring identifiers and incrementing ID versions if changes have occurred in the resulting sequences or splicing models. If ambiguity in a transfer target appears new identifiers are created and the ambiguous one retired. These events are tracked by the Ensembl infrastructure. Annotation metadata, including TSL, APPRIS, and GENCODE Basic assignment, is also frozen with each release and made available with the gene sets. Periodically releases will be flagged as "reference" indicating their use in large consortia or marking the last annotation release on an assembly. For example, GENCODE v7 was used for the 2012 ENCODE analysis[93] and GENCODE v19 was the last release native to the GRCh37 human assembly.

Data access GENCODE makes its annotation available through as many data providers as possible. We will to continue these mechanisms of distribution to maximise the annotation's value. Annotation is available through the GENCODE Genes website (genocodegenes.org), Ensembl and UCSC Genome browsers and all of their tools. In addition the primary distribution sites (GENCODE Genes and Ensembl) also provide archival access to each data freeze in the form of flat-files on our FTP servers and archived Ensembl websites. GENCODE annotation will also be made available over GA4GH APIs as these develop. Most of the data formats used by GENCODE are amenable to representing the newer annotations we will make available and when this is not the case additional metadata is made available as tab-separated files. All data freeze files are also versioned. Further information is provided below in the Management, Dissemination and Training section.

Cross-referenced resources GENCODE annotation is cross-referenced to major bioinformatics resources including HGNC, UniProt and RefSeq with every annotation merge as part of the Ensembl release cycle using the Ensembl cross references system. This system can link annotation based on sequence alignment, coordinate overlap and direct assertions; other manual annotation groups such as HGNC and UniProtKB extensively use the latter strategy. Annotation is also linked to GO terms via GOA and by transferring terms from homologues genes as predicted by the Ensembl Compara GeneTrees pipeline[94]. We also cross-reference to disease phenotype data including OMIM, Orphanet and IMPC via transitive mappings to nomenclature committees and then to the GENCODE annotation.

Software All components of the main annotation toolkit (ZMap, Otter, Annotrack and Blixem) are held in source code control repositories from which all development takes place using standard software development methodology for controlled introduction, release or (if necessary) rollback of new features and bug fixes. These will be migrated into the popular source management system GitHub making external contributions to the projects easier. All software is made available under open source licenses. Each release of the annotation tools is extensively tested on Unix/Linux and OSX. Annotation is stored in the Otter database with access controlled via a Single Sign On and will migrate to OAuth2 enabling authentication via ORCID. An audit trail is kept for all database edits allowing roll back in case of error.

Annotation consistency To ensure consistency across the different manual annotators, all new annotators go through a process of training, feedback and mentoring in their first year. Extensive guidelines are available to ensure consistent SOPs and annotation procedures. Periodically we perform consistency checks where a single region is annotated by all annotators and compared for consistency and accuracy with results fed back into the annotation SOPs.

Improving the GENCODE annotation

We will take two major approaches to improve the GENCODE annotation. These will be applied both to the human and the mouse annotation, but from slightly different perspectives. The first major approach (corresponding to Aim 1) will complete human and mouse annotation to the extent technically and operationally possible (see Milestones) with a focus on extending existing human partial models to full length, expanding the improvement of human lncRNA annotation and completing the initial full pass of mouse. The second major approach (corresponding to Aim 2) will be to incorporate individual genome representation and population data represented by available human variation data at the sequence and transcriptomic level and by the 16 mouse strain genomes produced by the Mouse Genomes Project. Two planned pilot projects, one in Aim 2 and one in Aim 3, will be undertaken with the goal of improving the process of annotation and to expand the overall utility of GENCODE.

Aim 1: Toward completing GENCODE's reference gene annotation

Supplementing manual annotation with validated automatic models The scale of data available from second and third generation transcriptomic datasets (RNA-seq, CaptureSeq, SLR-seq, PacBIO) necessitates an alternative strategy to annotation. We will automatically create "pre-GENCODE" models and prioritize them for manual evaluation. All pre-GENCODE transcripts will be excluded from the GENCODE gene set until an annotator evaluates them. Second and third generation based transcripts will be intersected with spliced RNA-seq reads, sequence conservation, complete protein domains, APPRIS, and the UniProt canonical isoform. Where available CAGE and polyAseq data will be used to identify the ends of the novel transcript. Transcripts at protein-coding loci will have a biotype assigned and CDS added based on the existing CDS annotation at the locus where appropriate

Annotation of novel features Extending annotation to full-length models provides the opportunity to annotate additional features at the level of the transcript. Upstream open reading frames (uORFs) are particularly relevant as they regulate the passage of the ribosome along the mRNA and play a role in controlling translation initiation[95-97]. Variation affecting uORFs can have adverse effects on translational regulation[98,99], and creating high quality annotation for them is related to our pilot project annotating regulatory features. uORFs are not comprehensively represented in any other gene set and the quality of their annotation is directly affected by the annotation of TSS and 5' UTR. Annotating uORFs in conjunction with such features adds considerable benefit over methods lacking the same contextual information. Integration of uORF annotation with other datasets such as CAGE and RNA-seq data will allow us to potentially capture the consequences of alternative TSS usage on the gene regulation. By integrating Ribo-seq and MS data we will build on the initial annotation of uORFs to provide information relevant to their functionality such as translation initiation efficiency and encoding of stable proteins.

Finishing the Mouse Pseudogene Annotation We will complete the mouse reference genome pseudogene annotation as part of Aim 1 and have plans to develop customized pseudogene annotations for the available mouse strains within Aim 2. At present, prior to the refinement of the

pseudogene annotation anticipated from more accurate protein-coding gene annotation, PseudoPipe identifies 18,627 putative pseudogenes on GRCm38, RCPedia finds 9,755 processed pseudogenes and Retrofinder predicts 18,467 retrocopies. The tri-way consensus between the pipelines is approximately 80% for processed pseudogenes. We will evaluate the annotation accuracy of our pipelines and refine the pseudogene identification and characterization process using the manually annotated pseudogenes as a gold standard and comparing them with the automatic predictions.

Annotating loss-of-function events We will build on our experience in identifying and analyzing loss of function events in human[46], to develop a reliable annotation framework for unitary and polymorphic pseudogenes and LOF variants in mouse. We will also develop a specialized tool to annotate the impact of LOF variants on gene function.

We will annotate unitary pseudogenes by creating a global inventory of orthologs between the mouse strains using the multi sequence alignment data from UCSC, annotating the syntenic regions, surveying gene disablements. To identify polymorphic pseudogenes, we will extend the Variant Annotation Tool (VAT)[100] to annotate variants that revert disabling stop codons. We will create a universal framework to identify putative LOF variants by combining function based annotation, evolutionary conservation and biological networks data. For this we will integrate resources such as Pfam domains, signal peptide, transmembrane annotations, post-translational modification sites, NMD prediction, and structure-based features (e.g. SCOP domains) and calculate variant position-specific GERP scores and dN/dS values.

We will build a LOF variants characterization module that will include network features to predict disease causing variants by using a proximity parameter that gives the number of disease genes connected to a gene in a protein-protein interaction network and the shortest path to the nearest disease gene. We will also develop a prediction model to classify premature stop causing variants into those that are benign, those that lead to recessive disease and those that lead to dominant disease using the annotation output as predictive features. We will validate our classifier using LOF from Mendelian Diseases, Cancer samples and healthy control datasets such as 1000 Genomes and ExAC.

Annotating pseudogene activity We will leverage our experience in pseudogene transcription analysis, using RNA-seq data to calculate a RPKM value for each pseudogene as an indicator of transcriptional activity in mouse. We will highlight tissue and strain specific transcribed pseudogenes. We will also integrate evolutionary and regulatory data to further characterize pseudogene activity. We will significantly improve on our annotation of human pseudogene activity to understand the regulatory potential of human transcribed pseudogenes by leveraging manual annotators experience and focusing on the transcriptomics (ENCODE, BrainSpan, TCGA), epigenomics (ENCODE, Roadmap Epigenomics) and cis-regulatory interaction data (GTEx, PsychENCODE).

Aim 2: Annotation of individual and population data

Current human genome annotations are based on the reference genome, which does not provide a full representation for the large genomic diversity of the human population. Personal annotation will allow us to account for differences due to individual variation in genes and other genomic elements. Further, it has been demonstrated that using the diploid genome with an individual's variants improves both mappability of the reads[101] and downstream analysis results[1].

We will develop an individual genome GENCODE annotation resource containing a number of tools and utilities to identify GENCODE-annotated features characteristic to an individual, such as a distinctive set of functional genes or the structures of variant-affected transcripts. We will detect and annotate novel sequence in the individual genomes and will experimentally validate the consequences of this variation.

As the international community is providing more and more genomic data, in the form of haplotypes and strain genomes, there is a general push to define a broader reference genome that incorporates all known variants. We already have significant experience annotating alternate sequences in non-linear genomes: the GRC, who maintain the human and mouse reference genomes, has been releasing both

haplotypes and fix patches for several years now and we annotate all of them (Figure 7). An advanced representation of genome assemblies and variation, known as a graph genome, will require us to adjust our methods accordingly. Pilot project 1 will explore this question further.

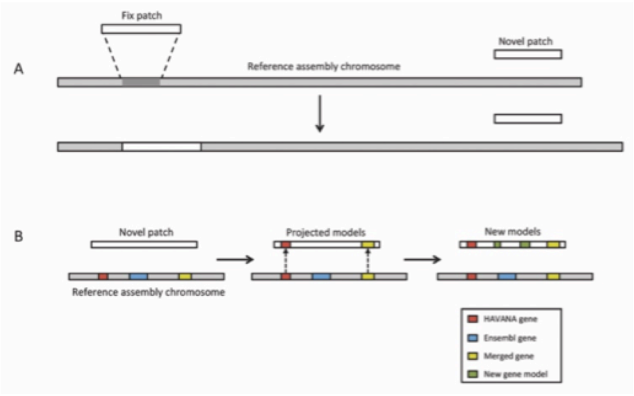


Figure 7: Patch annotation in GENCODE. A: GRC provide two types of patches: fix patches are integrated at the GRC's next major version of the assembly and novel patches will remain as alternative sequence. B: When annotating a novel patch we copy gene models from the primary assembly onto the patch. In this example, the red yellow genes are copied to the patch. The blue gene is not copied because the underlying genomic DNA is too different to enable the projection process. After projection, a patch will be annotated fully using the Ensembl annotation pipeline. In this case two new gene models (green) have been annotated on the novel patch.

Protein-coding gene annotation in the mouse strains As part of the Mouse Genomes Project, UCSC has developed gene and transcript sets for sixteen mouse strains. We have improved three interrelated and interdependent problems of genome assembly, genome alignment and genome annotation, iterating on each repeatedly and progressively to create a gene set for *Mus musculus* with several hundred new genes (see Figure 8 for an example of a substantial new gene), thousands of new isoforms and tens of thousands of novel gene haplotypes. This is helping to improve the existing mouse GENCODE reference set: identifying polymorphic pseudogenes, finding unannotated loci in the C57/BL6 reference genome and identifying reference assembly errors. It is also providing new insights about gene variation present across *Mus* species and, in conjunction with RNA-seq data, is allowing us to assess differential expression between strains using more accurate transcript sets. We have created a new, general purpose Clade Genomics Toolkit, which provides a range of tools that together provide a pipeline for simultaneous and consistent comparative annotation of many genomes, leveraging existing annotations and RNA-seq data.

HAVANA has also manually annotated regions of specific interest on genome sequences unique to individual mouse strains[102]. Where lineage specific regions of the genome are identified, an automatic gene prediction pipeline is run and those regions with potentially interesting genic features are targeted for manual annotation. The target regions are passed through the standard analysis pipelines to ensure a comparable annotation to the reference genome and where strain-specific transcriptomic evidence is available it can also be used to aid specific annotation.

Annotation of individual human genome sequence As high quality personal genome sequence becomes available, either publicly or via collaboration, we will apply our pipelines to regions distinct to the human reference genome sequence, allowing us to identify and capture all novel loci.

Currently, alternative sequences to the reference genome are passed through our standard analysis pipelines and manual annotated using the same guidelines to ensure the annotation is equivalent. Manual annotation has proven essential for difficult regions such as the leukocyte receptor complex (LRC) haplotypes, where there is a combination of tandem gene duplication and pseudogenization events. The repetitive nature of the DNA underlying these biologically complex regions pose particular challenges for alignment algorithms; for this reason the automated pipelines are prone to misannotate transcript structures, join distinct loci together and misannotate pseudogenes as protein-coding.

For small variants, the Ensembl, Yale and HAVANA groups have considerable experience in identifying variants and building specialised annotation pipelines such as the Ensembl Variant Effect Predictor (VEP)[103], VAT and ZMap. In particular HAVANA have developed experience in annotating LOF variation[2,47], however, the constraints of annotating on the reference genome made capturing and storing insights gained from manual annotation problematic. Recent updates to the ZMap annotation software now enable viewing variation and annotation on non-reference genome sequence, allowing

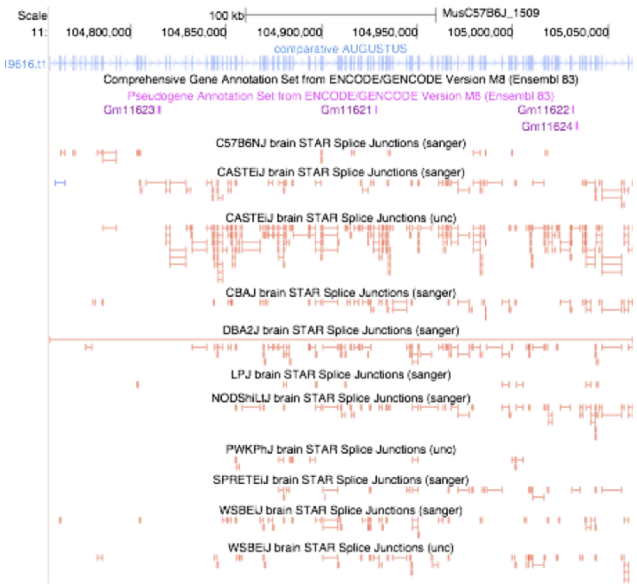


Figure 8: Comparative Augustus identified a 138 exon transcript in a locus not previously annotated in mouse. This transcript has varying splice junction support, with the most coming from *Mus castaneus*. The function of this transcript is being investigated.

us to represent the functional effect of the variant in its correct context within the transcript. This is significant as a variant consequence pipeline might indicate a nonsense codon as having a significant functional impact, whereas annotation of the same variant in the context of a CDS and transcript structure might modify that prediction if, for example, the LOF variant was close to the 5' or 3' end of the CDS. We are developing the ZMap software to make this annotation process more straightforward and are investigating alternative ways to save and distribute this information.

We have developed approaches and tools[101] to integrate personal variation data into the reference genome and produce an individual diploid genome, which can be annotated by mapping GENCODE annotations onto it. We have a large experience with constructing personal genomes, splice-junction libraries and personalized annotations and using them in functional genomic analyses[104-107]. A key aspect of personalized annotation is correctly adjusting the annotation for loss-of-function events, polymorphic and unitary pseudogenes. We have explored in detail their implications for the reference annotation[47] and observed that, in general, LOF events and polymorphic pseudogenes are just different versions of the same event, mostly depending on the major allele frequency. We characterized putative LOF events in individuals using the 1000 Genomes Phase 3 data[108]. Some LOFs may impact only one individual, resulting in the inactivation of an essential gene and leading to disease, while other LOFs can become fixed in the population as nonfunctional relics through pseudogenization. We also surveyed the impact of LOFs on personal annotation[47] and found that LOF variants that introduce premature stop codons resulting in a gene truncation in the reference genome can lead to an incorrect annotation of the gene. This highlights the importance of correct LOF identification for accurate annotation[47]. To this end, we developed a pipeline to identify unitary pseudogenes in human[46] and explored the functional constraints faced by different species and the timescale of functional gene loss[46]. These results along with fully annotated pseudogene sets are deposited at gencodegenes.org.

We will use the newly constructed personal annotations to identify LOF and pseudogenization events by comparison with the reference genome. We will use the personal protein coding annotation as input in our pseudogene annotation pipelines to create a comprehensive personal pseudogene complement. We will assess the annotated personal SNPs for allele specific expression using the data from AlleleDB[107], a repository of genomic annotation of cis-regulatory single nucleotide variants associated with allele-specific binding and expression. Next, by integrating Mendelian disease and cancer data we will use our variant annotation tool and the proposed LOF analysis pipeline to filter the LOF and pseudogenization variants and characterize them with respect to their disease driver potential.

Experimental validation of individual variation We will use variation and transcriptomic data from the same individuals to capture TSS, alternative splicing and TTS and construct representative transcripts associated with specific variants. In this way we can investigate variants tagged as eQTLs, sQTLs, and LOF variants and describe them in their proper transcript context. Furthermore we will work with WTSI to integrate ENTEx proteomics data from the same samples, to compare the impact of variation on proteins and transcripts with particular reference to protein and transcript abundance.

We plan to carry out a pilot CLS to investigate to what degree the human gene and transcript set varies between individuals (see section **High throughput, complete annotation of novel non-coding RNA transcripts)**.Little information exists about variation in gene and transcript number among individuals, and about the number of transcript isoforms private to an individual or individuals—possibly to escape LOF mutations. We will target specifically a small subset of genes of medical relevance, the 541 protein coding genes in the UKGTN data set. We will design a smaller library to capture these genes in RNA samples from seven organs from four individuals that have been deeply-characterised as part of the ENTEx project. At the sequencing depth we propose, we expect to analyse approximately 100 PacBio reads per gene, in each individual and organ. The analysis of this data will provide the first estimation of the level of variation in the gene and transcript set among individuals. This has practical consequences, since if this variation is not negligible it could impact gene expression estimates from RNA-seq data.

Another workflow within our pipeline focuses on personal proteomics, whereby we compare samples from multiple individuals to identify differences in gene and transcript expression, determine allele specific expression, differences in alternate splicing of genes, and to identify sequence variation. We will use multiplex TMT labeled ENTEx tissues samples to enable direct comparison of peptide abundance within a spectrum and highlight cases where a peptide is not present in an individual.

Pilot project 1: Graph genomes representation

This pilot project will explore the value for genome annotation and GENCODE in the use of a graph-based representation of the reference genome. Graphs potentially provide a universal structure for genomics with our early applications focused on human and mouse. They would be an elegant alternative to the current process of adding large alternative loci to the genome, which cannot easily scale to include a significant fraction of human variation because it would take the equivalent of hundreds of complete human genomes to do so. In a graph representation, the variation can be added in a fine-grained manner to create a structure that can represent multiple sequence "paths" simultaneously (Figure 9). A number of groups in the context of the Global Alliance for Genomics and Health (GA4GH) are now collaborating to construct complete reference graph genomes, annotated with rich haplotype information, and we expect them to be available in late 2016. This pilot will build on our existing work to develop a tool chain for graph genomes. We are working actively with Erik Garrison, the lead developer of vg[109] a sequencing read mapper, genetic variant caller, to create an efficient path indexing scheme that can be used to store thousands of haplotypes of complete genomes using a positional Burrows-Wheeler Transform (PBWT) encoded against a graph within a small amount of memory[110]. We believe this index will form the basis for storing and referencing isoforms compactly against a population graph.

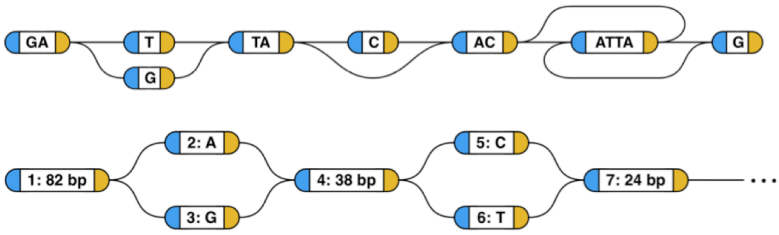


Figure 9: Example graph genomes: Each segment (node) holds some number of bases. A join (edge) can connect, at each of its ends, to a base on either the left (5', blue) or the right (3', yellow) side of the base. When reading through a thread to form a DNA sequence, you leave each base on the opposite side from which you entered, and reverse complement it if you enter on the 3' side and leave on the 5' side. The graph at the top shows these capabilities: one thread spells out (reading from left to right, along the nodes drawn in the middle) the sequence "GATTACACATTAG". Straying from this path, there are three variants available: a substitution of "G" for "T", a deletion of a "C", and an inversion of "ATTA". If all of these detours are taken, the sequence produced is "GAGTAACTAATG". All 8 possible threads from the leading G to the trailing G are allowed. The graph at the bottom is the beginning of the genome graph for BRCA2 derived from the 1000 Genomes phase 3 data, with long sequences elided. Only the first few nodes of the graph are shown. (Adapted from Novak et al, A Community Evaluation of Reference Genome Graphs, submitted)

Mapping evidence to the graph To pilot the approach, starting with a small number of genes, we envisage storing the isoforms collected across many individuals (see Aim 2) efficiently against a

reference graph genome. Establishing means to store other sources of GENCODE evidence such as CAGE tags for TSS against the graph will use similar mechanisms. Mapping the evidence onto a graph can be done quickly[110] and has the key advantage of removing reference bias as it has been demonstrated that sequencing reads that overlap variants to the reference are statistically prone to be left unmapped, biasing allele specific quantification and reducing sensitivity to rare gene products[101].

Analyzing and annotating the graph Having projected the experimental evidence onto the graph, we will want to provide analysis and visualization tools to explore and synthesize the data. Some variants affect transcription, splicing or translation in significant ways, although most are functionally neutral. It will therefore be necessary to map the data across haplotypes and compare the results. In some cases, the evidence will be mapped across alleles, sometimes not. Determining a precise and reliable method for doing so will be imperative for efficient and robust graph-based annotation. As an output, each gene annotation will itself be a sub-graph that covers a subset of haplotypes. At the moment, the GENCODE annotation assigns a new locus ID for every occurrence of a gene on an alternative reference sequence. This cannot scale up to a graph and its astronomical number of possible haplotypes. Conversely, explicitly listing all possible mappings of a single gene, across all known variants of that region would be impractical. We will therefore require automated methods to define when an annotation and its identifier can implicitly be projected across alleles or not. In particular, we will take advantage of the knowledge GENCODE accumulated over the years to define valid gene sequences.

Pilot project 1 reporting and plans We define success in this pilot project initially as mapping of a set of GENCODE transcripts within the graph. Secondary success will be marked by efficiently reading annotation from the graph for display or analysis purposes. We have no plans to convert GENCODE's Annotation Toolkit to work directly on a graph data structure, but if early parts of this pilot are successful, we may consider software to transform annotation in a graph representation to a linear representation in support of Aim 2.

We plan for roughly 1 FTE effort with the majority at UCSC and the remaining at EMBL-EBI over the first two years of GENCODE for this pilot project. Progress will be discussed by the GENCODE investigators and also within the GA4GH as appropriate and formally presented each year to the SAB for comments (see Milestones). After year 2, we will ask the SAB for their recommendation to continue, expand or end this pilot project.

Aim3 and Pilot Project 2: Connecting regulatory regions to regulated genes

This pilot project will determine the scope and parameters of Aim 3. It is based on the concept that the regions of the genome that regulate genes should be considered as part of the genes they regulate. Therefore, where possible, these regions should be directly connected to and annotated with the genes themselves. We believe that a comprehensive solution to this problem in all tissues and conditions is some distance into the future; however, we also believe that the state of our knowledge about links between regulatory regions and genes in 2016 is similar to the state of our knowledge about genes in the early 2000s, when GENCODE was established, and thus this pilot is warranted and necessary.

For a gene to function effectively, it has to produce the right molecular product in the right cell at the right time. The regulatory regions that modulate its expression are therefore key components of the proper function of a gene (Figure 10). GENCODE's gene annotations provide little

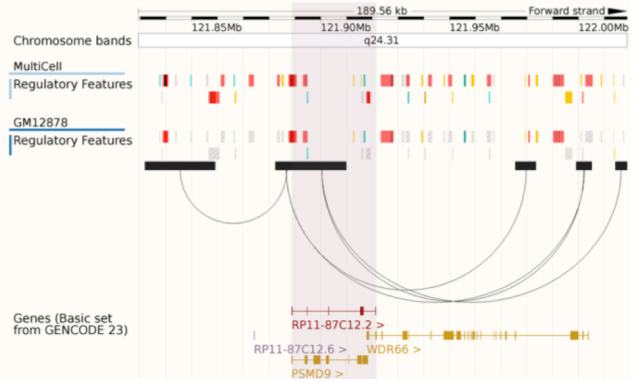


Figure 10: A prototype gene annotation. RP11-87C12.2 is currently represented as a set of exons and UTRs along the genome, however it is surrounded by cell type dependent regulatory elements. In addition, the ties between these regulatory elements and promoters (represented as arches) are also tissue dependent.

indication of the dynamic function of genes although alterations of these regulatory mechanisms have been shown to play significant roles in health, phenotype and evolution[111-113]. Genome regulation depends on many factors including cell type, developmental stage and environmental conditions, and it is estimated that a large fraction of the genome is directly involved in modulating gene expression[114]. Although understanding is incomplete, we are able to create useful computational maps of genome regulation[115-117] based, in part, on molecular details of how regulatory regions are associated with genes[118]. Connections between regulatory variation and gene expression and the ability to measure them are increasingly clear[119-121].

Currently, detecting regulatory elements in a reasonably comprehensive manner requires collecting many datasets across many conditions. Large consortia, which pool and coordinate resources, including ENCODE[93,122], FANTOM5[123], Epigenomics Roadmap[124] and Blueprint[125] have successfully collected these data and done an initial analysis. Having identified possible regulatory regions, various strategies have been adopted to attach target genes to these regulatory regions including correlation of dynamic signal[123,126], genetic association[127,128] and physical proximity[129-132]. Despite their potential, these results have led to many separate maps of gene regulation rather than few integrated ones and no careful approach has considered how to evaluate and integrate the variety of evidence that would enable reliable annotation of the connection between genome regulatory regions to the genes they regulate. This pilot will do so in a deliberate and limited way. If successful, we will incorporate our methods more widely into the GENCODE annotation.

Collecting relevant regulatory datasets We will collect regulatory regions, define the attributes of these elements and attempt to annotate their target genes depending on tissue. The initial pilot will a) leverage the cis-regulatory datasets already collected and uniformly stored by Ensembl[133] and in the ENCODE Encyclopedia[117]; b) analyze them with appropriate machine learning methods; then c) manually review selected regions and compare them to reported regulatory variants[134] to better understand the limitations of the automatic pipelines and feedback improvements into the automatic annotation process.

The Ensembl Regulatory Build[116] already incorporates ChIP-Seq datasets segmented across the genome and annotated with transcription factor binding site motifs. It will expand via separate funding from the Ensembl core grant to DNA methylation[135], enhancer RNA (eRNA)[123] and other relevant data types in 2016 and 2017. GTEx eQTLS (expanding this to all available eQTL datasets in late 2016) are also available within Ensembl. We will further access the ENCODE Encyclopedia and FANTOM5 data and collect physical proximity measurements, including Hi-C, Chia-PET or Promoter Capture Hi-C[121].

Integration of experimental evidence We will link these regulatory annotations to GENCODE gene and transcript annotations using three lines of evidence: functional links, based on activity correlation; genetic links based on eQTL information; and physical links based on 3D conformation. For functional links, we and others have previously developed correlation-based links across many cell types[115,136] to find potential targets of regulatory regions, and we showed that the predicted links are supported by eQTL information[115]. We have recently extended this work in the context of a probabilistic model that links enhancers and genes into modules of similar activity patterns using Latent Dirichlet Allocation (LDA), which results in less noisy predictions than simple correlation. Briefly, we use the matrix R of enhancer activity signals and matrix T of gene expression signals cell types, and assume there are K_1 enhancer modules and K_2 gene modules. The cell type specific activity of each enhancer and gene module is expressed probabilistically as $\mathrm{p}(\mathrm{E}_\mathrm{k}|\mathrm{t})$ and $\mathrm{p}(\mathrm{G}_\mathrm{l}|\mathrm{t})$ respectively for cell type t. We then calculate the probability of linking each enhancer module to each gene module, using a diffusion model, so that modules correlated with many other modules are inherently less probable to be linked to any of them. In preliminary work, we applied this probabilistic model on 56 tissues (Figure 11a), resulting in 290,561 statistically-significant interactions (FDR<0.01), involving 21% of enhancers and 88% of genes at a median distance of 50kb (up to 1Mb was allowed). The resulting links showed strong agreement with eQTLs across lymphoblastoid, blood, brain, fat, skin and liver studies[137-140] (22%-42% eQTLs

overlap with our activity links), Hi-C datasets in six cell types[141] (28%-35% overlap), and ChIA-PET datasets (26.5% of links overlap) (Figure 11b)).

The interaction data will finally be integrated across evidence types using a Bayesian model that takes into account all the available data: sequence, epigenome, genetics, Hi-C-based, ChIA-PET, etc. We could combine these datasets using a number of machine learning approaches that use cross validation as well as experimental validation through targeted mutagenesis using technologies as CRISPR[134]. We are concurrently collaborating with Oliver Stegle whose group is developing such methods (See Letter of Support). This will produce an automatic gene assignment for all enhancers depending on tissue and cell type.

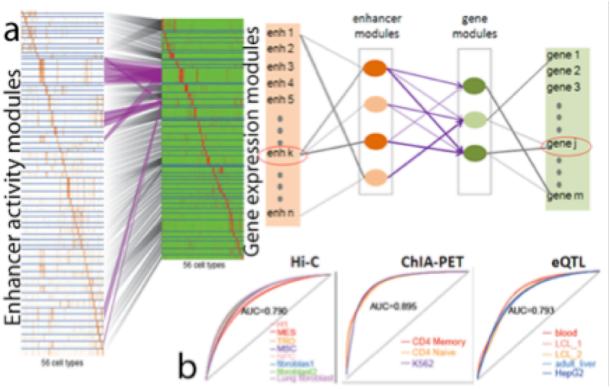


Figure 11: Linking regulatory variants to their target genes. a. Top. Module-based linking of distal enhancer regions (orange heatmap) to their target genes (green heatmap) based on clustering of both enhancers and genes into modules (left) and then linking modules to each other based on correlated activity (purple links). Bottom. The resulting activity-based links are strong predictors of chromatin conformation capture by Hi-C (left) and ChIA-PET (middle) in matched cell types (colors), and of eQTLs (right), establishing their biological relevance.

Manual curation of results and feedback process GENCODE has demonstrated the importance of manual annotation for creating gene annotation based on a diverse collection of evidence. Just as genes that were annotated genome-wide using systematic prediction algorithms were shown to have blind spots at non-typical loci, detailed examination of a subset of the current computationally-based regulatory and cis-regulatory annotations will reveal, we are certain, critical unexpected biases and edge-cases leading to errors. Feedback from manual annotation will improve existing computational processes and create a trusted "gold standard" data set connecting genes to regulatory regions that has the potential to encourage considerable innovation in the bioinformatics community.

Manual annotators have various tools available to detect possible annotation errors. The planned developments for ZMap and the other components of the annotation toolkit (see below) to support new file formats will enable the display of diverse additional data types such as cis-regulatory and physical interactions. We will pilot the integration of these data types with those used currently to annotate regulatory features, annotate the connection of regulatory feature to genes and annotate transcripts and genes associated with regulatory features such as elncRNAs, bidirectional lncRNAs, alternative 5' UTRs originating from alternative promoter and enhancer sequences.

Pilot project 2 reporting and plans We define success in this pilot project as successful computational integration of the functional, genetic and physical data sets linking regulatory regions to genes and the ability of select manual annotators to evaluate the collection of available evidence at key loci. Our effort will be collaborative and complementary (not competitive) to the ongoing work to create the ENCODE Encyclopedia: several of us (Kellis, Gerstein, Flicek, Zerbino) are long-term members of the ENCODE consortium.

We plan for between 1 and 1.5 FTE effort split roughly as 20% at MIT and 80% at EMBL-EBI (and as 70% computational and 30% manual annotation) over the first two years of GENCODE for this pilot project. Progress will be evaluated by the GENCODE investigators during our regular phone calls and formally presented each year to the SAB for comments (see Milestones). After year 2, we will ask the SAB for their recommendation to continue, expand or end the pilot project.

GENCODE scalability

Wide-scale adoption of automatic methods of annotation is essential to increasing the throughput of manual annotation. Without this the comprehensive annotation described in Aim 1, population annotation of Aim 2 and regulatory annotation of Aim 3 will be impossible to achieve. We envisage solutions where annotators are directed towards locus needing attention rather than methods relying on

de novo manual model creation. Automatic methods will annotate new models for manual verification and supplement existing models with additional information.

Our current baseline manual annotation rate is approximately 23,000 loci per year. Our first year annotation efforts will focus extensively on mouse with a rebalancing to focus strongly on human in years two to four. We do not expect the number of unannotated transcript loci to be the rate-limiting figure and so increases beyond this 23,000 loci figure will depend on the methods detailed below to scale up the annotation process.

Driving manual annotation from automatic methods Volumes from second and third generation transcriptomic methods must be manageable by a manual annotator to achieve Aim 1's goal of comprehensive annotation. To assist, we will automatically create models from these data sets and supplement them with additional annotation. These will be included into the final GENCODE gene sets only after manual review. For example, Ensembl already posses methods for predicting RNA-seq derived models and for the case of SLR-seq, CaptureSeq and PacBio data, these results can be plotted to the genome using modified versions of the existing cDNA alignment infrastructure. In all cases these models will flow into the automatic annotation sets, be flagged for inspection, corrected/rejected/accepted and be given a functional assignment. Manual annotators will verify splicing and functional assignment and (for pilot 2) the assignment of regulatory annotation of a model.

In addition, automatic QC will be run on models previously annotated and used to flag additional suspect models such as detecting split-genes from comparative genomics data. Key to this will be ensuring that any QC system does not repeatedly raise issues a manual annotator has deemed acceptable and does correctly identify when an issue should be re-raised. We will extend the existing Ensembl QC systems to provide this functionality.

These events will flow into Annotrack and will drive the opening of regions in the ZMap tool. ZMap's quick region open will be enhanced to respond to the type of issue raised. This allows the tool to display those data sets essential for the QC process required such as RNA-seq model verification with the selected data sets driven by manual annotator experience. This will have the effect of using Annotrack as a central repository of issues raised by GENCODE collaborators and a hit list of loci requiring annotation.

Efficient and directed mouse strain annotation The small divergence time frame[142] among the 16 mouse strains facilitates the dependable transfer of annotation across genomes. The current volume and expected increase of strains available mean that population based annotation as describe in Aim 2 requires supplementing manual methods with of automatic ones.

Possible transfer targets will be informed by alignments of the strains back to the reference mouse assembly. We envisage two complementary methods tuned to the type of annotation being transferred. In the case of non-pseudogene loci, we elect to use the Clade Genomics Toolkit in conjunction with primary evidence such as RNA-seq. Annotation, which has been found to be significantly different from the reference mouse strain, will be flagged via Annotrack for manual review. When working with pseudogene annotation we will use UCSC's LiftOver tool to perform the annotation transfer. In addition, we will extend our pseudogene annotation pipelines to produce an accurate map of pseudogenes in mouse strains.

We also will need to extend our pseudogene detection pipelines to enable novel detection in the mouse strains. We will use the conserved protein-coding genes between each strain and the reference genome as input for identifying pseudogenes. The extended PseudoPipe workflow is summarized in the following steps: 1) identify and extract consensus proteins from Ensembl; 2) mark the coordinates of protein-coding genes; 3) six frame blast homology search to match the consensus peptides to the strain sequence; 4) refine results and eliminate redundant hits; 5) merge hits and identify parents; 6) align parents and pseudogenes and check for disablements (e.g. premature stop codons); 7) assign pseudogene biotype.

RCPedia will also be adapted to integrate gold standard transcript annotation, such as GENCODE mouse annotation and strain annotation. The extended RCPedia pipeline is summarized as follow: 1) Merge multiple annotations using an hierarchical prioritization; 2) align transcripts sequences to the target genome and extract alignments; 3) prioritize intronless alignments; 4) remove alignments parental introns and remove putative genomic duplications; 5) rank parental transcripts; 7) calculate properties of the putative pseudogene, such as target site duplication sequence, identity and polyA length.

Improvements to the manual annotation toolkit ZMap represents the primary analysis platform used by the manual annotators. It is a fast application written in C++ capable of displaying hundreds of tracks of related data and coordinating data set loads across multiple remote resources. A number of features in ZMap require the accompanying Otter tool including annotation remapping between assemblies and transcript model storage. To improve annotation workflow, we will extend ZMap to support “on the fly” remapping of annotation when remapping alignments are available between two genome assemblies allowing manual annotators to interactively load any data they require.

In addition we will support the UCSC developed Track Hub format to access more primary annotation data sets in the manual annotation process. This will accompany developments to support compressed file formats (CRAM, BigBED, BigWig) so that ZMap will be capable of handling any source data set from our annotation pipelines. As new formats are created, including those from GA4GH, we will continue to extend ZMap’s support for external annotation. In addition we will re-code portions of ZMap’s display and data fetching code into the modern Qt GUI toolkit to take advantage of desktop advances such as OpenGL for display and faster rendering of dense data sets. We will transfer these enhancements to the accompanying tools including Blixem to ensure a consistent interface between tools and to support new data views.

Finally we aim to help manual annotators by automating a number of processes performed in ZMap. We will augment the current gene model building process to aggregate isoform predictions and to automatically add supporting evidence metadata. This is essential to ensuring that the correct supporting evidence is included with the model and to reducing potential misannotation of loci.

Incorporating community annotation

Supporting external annotation efforts

Although we have described plans to provide both reference and individual genome annotation, demand for the manual annotation of transcripts in specific situations across individuals, strains and species dramatically outstrips our ability to provide such services using the mechanisms described here. Thus, we will offer a two-tier system combined with workshops and training described below for those wanting to use the world-leading GENCODE infrastructure for the manual annotation of genome sequences.

General Submission System We will support the submission of external annotation back to a new Transcript Archive at EMBL-EBI, developed separately to this application. This archive will make use of existing Otter technology to allow the archive to trace annotation calls back to individual submitters or consortia and as a way for automated processes to retrieve annotation. This linking will be accomplished via ORCID or another suitable authentication/authorization system. Annotation will be QC’d upon submission and then, should it pass, integrated into annotation builds as merge annotation. ZMap can already be used standalone and will have basic QC integrated to support allow for high-quality annotation to be generated for submit to archives. Annotators can choose to use ZMap or to supply submissions from their own software.

“Trusted” Submission System Otter uses a “Single Sign On” system to give access to its editing system to trusted users. These users can directly edit gene models whether working at the EMBL-EBI or remotely. To support this, we will develop Otter servers for cloud-based platforms.

Organizational structure and staff responsibilities

The GENCODE consortium arose from groups that started working together during the initial sequencing and annotation of the human genome and has evolved to encompass major informatics, annotation and genome browser groups. The GENCODE wiki page provides an uncommonly useful view of this history: https://en.wikipedia.org/wiki/GENCODE. The current application represents continuity even as GENCODE moves its lead institute from the Wellcome Trust Sanger Institute (WTSI) to the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) concurrent with the 2017 move of the Human and Vertebrate Analysis and Annotation (HAVANA) group. The current move from represents a gradual change that is also reflective of the project's history: Ensembl has been a part of GENCODE since the beginning and was a joint project of WTSI and EMBL-EBI until 2014 when Ensembl consolidated EMBL-EBI. This renewal application will also see a change in the GENCODE lead PI to Paul Flicek (EMBL-EBI), who has led Ensembl since 2007 and been active in genome annotation for 15 years.

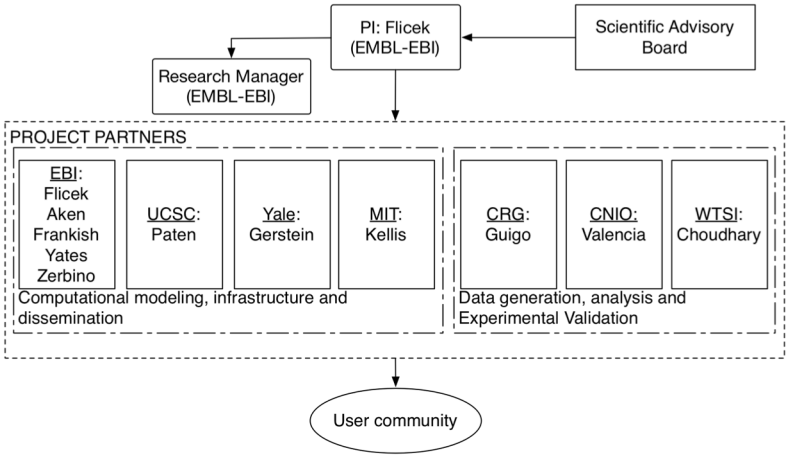


Figure 12: The GENCODE management structure with principal investigator names for each of the partner institutions.

At EMBL-EBI, GENCODE will be part of the Genes, Genomes and Variation (GGV) cluster and the various project components will be managed by four key personnel: Adam Frankish, Bronwen Aken, Andrew Yates and Daniel Zerbino. These managers already work closely together to deliver several GGV resources including Ensembl. They are responsible for different activities within GENCODE and are therefore well-placed to ensure that the aims are met. Specifically, Frankish will lead the manual annotation activities within GGV and report to Flicek. Note that Frankish is currently an employee of WTSI and will not become an EMBL-EBI employee until 1 April 2017 (see HAVANA transition plan below). Software pipelines supporting HAVANA will be maintained and developed by Zerbino synergistic with his current responsibilities for Ensembl core software and overall GGV genome analysis pipelines. Zerbino will also be involved in the regulatory pilot. Data access, distribution and training will be led by Yates who is also responsible for the Ensembl web site, Ensembl release process and GGV outreach and training. Aken will lead the Ensembl GeneBuild and QC activities. A half time research manager will assist Flicek in the administrative and reporting aspects of the grant. The PIs at each of the performance sites will be responsible for project management and reporting for their site (Figure 12).

Among the other partners, Mark Gerstein (Yale) will be primarily responsible for pseudogene annotation; Benedict Paten (UCSC) for engagement with the UCSC Genome Browser group and the graph genome pilot project; Manolis Kellis (MIT) for comparative genomics algorithms and be involved in the regulatory pilot; Roderic Guigo (CRG) for transcript validation and analysis; Michael Tress (CNIO) and Jyoti Choudhary (WTSI) for proteomics data generation and analyses. We have also engaged Tim Hubbard, Professor of Bioinformatics and head of the department of Medical and Molecular Genetics and King's College London and former GENCODE PI and head of informatics at WTSI as a special advisor to the project for the sake of continuity and for his experience in genome annotation and clinical applications of genomics.

We will continue with the current methods that we have successfully used in GENCODE for monitoring progress and ensuring that partners are up to date over the past four years:

- A research support assistant to coordinate formal progress reports, calls, and the annual meeting
- A dedicated “closed” mailing list to communicate progress
- Bi-weekly teleconferences between the consortium members to monitor actions and discuss issues and to provide progress updates
- A password-protected internal wiki site to maintain internal progress documentation
- An external public website highlighting major updates (gencodegenes.org)
- An annual GENCODE consortium meeting to discuss progress and report to the SAB
- Annual formal progress reports to NHGRI

Conflict resolution and transition planning All of the investigators have significant experience working together in a variety of projects over the last 13 years and no major conflicts are anticipated. Should any arise, we will aim to resolve any differences of opinion regarding the direction, process or strategy of GENCODE informally; if this is not possible, issues will be escalated first to the group of investigators and the lead PI (Flicek), then our SAB and the NHGRI. As PI transitions were required twice in the current GENCODE funding due to staff departures at WTSI, we have an established model in the unlikely event this is needed. Should Flicek no longer be able to carry out the responsibilities of PI, a transition plan would be developed by the EMBL-EBI team and presented to the other investigators and then to NHGRI.

HAVANA transition plan The intent to transition the HAVANA project from WTSI to EMBL-EBI was announced in early 2014 and the planning to ensure a smooth transition both scientifically and administratively has been actively underway since then. Flicek, Aken, Yates and Zerbino were all directly involved with the transition of Ensembl and have used this experience in the planning. The date of final transition is set for 1 April 2017 coincident with the start of the proposed funding from this application and will follow an orderly process informed by previous resource transitions from WTSI to EMBL-EBI. Specifically, we will incorporate HAVANA into the Genes, Genomes and Variation cluster of resources (http://www.ebi.ac.uk/services/dna-rna), which is led by Paul Flicek and with the scientific responsibilities divided as described above. As the technical transition proceeds, we will move computational and software infrastructure from WTSI to EMBL-EBI through 2016 and early 2017 and provide HAVANA staff with guest logins to for testing and use before they physically move between institutes. The staff and physical transition will occur in 2017 although Adam Frankish has already been offered an EMBL contract with start date of 1 April 2017 underwritten by EMBL core funds to facilitate necessary his necessary administrative tasks. Because WTSI and EMBL-EBI are separated by less than 100 feet the physical transition will incur a disruption of no more than 1 day.

Scientific Advisory Board

GENCODE will receive advice from a planned six member scientific advisory board (SAB) covering essentially all aspects of the project. The SAB will include three members of the SAB from the current iteration of the project for continuity and three new members. The fourth member of the current SAB, Steve Brown, MRC Harwell, was asked to step down due to his status as co-PI with Flicek on another grant. The role of the SAB will be to provide advice on progress, priorities, new technologies, operational processes of the consortium and serve as representatives of our user community. The SAB will also assist in evaluating the progress of the GENCODE pilot projects. They will also provide advice on any other improvements within their areas of expertise including operating in a cost-effective manner. The returning members of the SAB are Tom Gingeras, Cold Spring Harbor Laboratory; Ross Hardison, Penn State University; and John Rinn, Harvard University. In accordance with guidance, new SAB members have not been asked, but are expected to represent proteomics, mouse informatics and population transcriptomics.

The SAB will meet annually over the course of two days with the project PIs and may be called upon at other times for specific advice. The SAB meeting agenda will be set in discussions between the GENCODE PIs and the chair of the SAB. The format and process of the SAB meetings will follow a

similar pattern to other EMBL-EBI project SABs including Ensembl and the NHGRI-EBI GWAS Catalog. In both of these cases, the SAB report is responded to formally via conference call six months after the SAB meeting (or sooner if required) and further updates are provided as part of the following SAB meeting.

Aim 4: Access and dissemination

It is paramount to the project's impact that GENCODE be made available to as many researchers as possible. This includes both making the annotation easily available and consumable and making all GENCODE developed software available for offsite use. In this section, we describe current methods for GENCODE data access as well as future plans intended to make GENCODE more accessible and, together with the Training section below, easier to use.

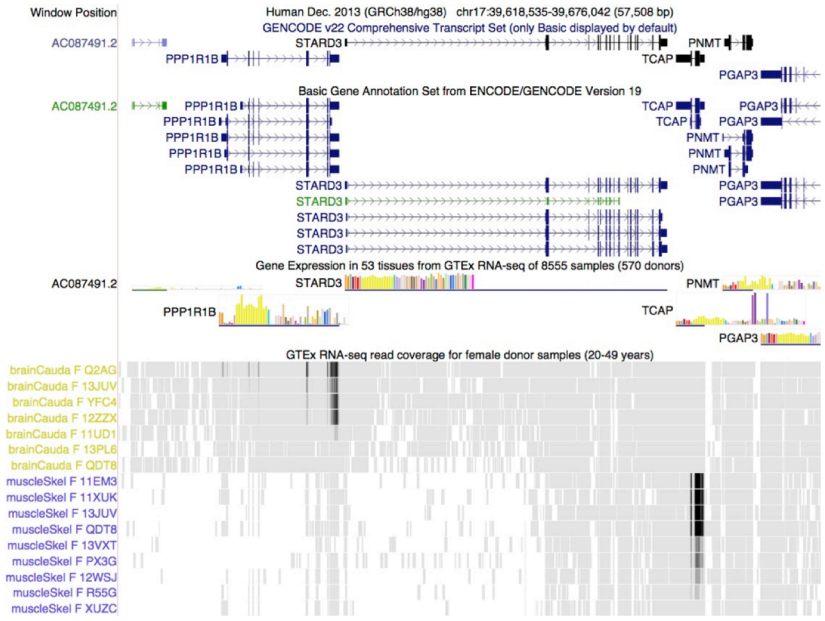


Figure 13: Gencode and GTEx in the UCSC Browser. View of a 56 Kbp region of human chromosome 17 where GENCODE annotates one non-coding and 5 protein-coding genes. Two genes in the region display tissue-specific gene expression as evidenced by GTEx RNA-seq including TCAP (titin cap protein) in muscle tissue and PP1R1B (a therapeutic target for neurologic disorders) in brain basal ganglia but not muscle. In this UCSC Genome Browser view, the main UCSC genes track (based on GENCODE v22, and colored by evidence strength) is configured to show a single isoform, while the earlier GENCODE v19 (used as the basis of the GTEx analysis shown here) shows all isoforms in the basic annotation set.

Genome browser access

Access to the GENCODE annotation is primarily though the Ensembl and UCSC Genome Browsers, two of the most widely used resources for genome science. Both are funded separately for the majority of their activities and through this grant only for specific additional details directly related to GENCODE.

As both Ensembl and UCSC use GENCODE as their default human annotation, GENCODE is deeply imbedded into the tools and interfaces that biologists and bioinformaticians use everyday. The Ensembl and UCSC genome browsers each serve approximately 150,000 active individual users per month and a combined total of well more than 1 million unique users each year. Many researchers use both browsers as part of their workflow. Together we believe that UCSC and Ensembl reach essentially all researchers in vertebrate genomics and are used regularly by the overwhelming majority of all researchers, clinicians and even interested members of the public working in genomics. This grant also supports an interface between the UCSC Browser group and Ensembl, helping to ensure data consistency between the two browsers.

Integration with other datasets and tools As the primary human gene set in the UCSC and Ensembl browsers, the GENCODE track is displayed by default when either browser is first visited. GENCODE annotations now serve as the linkage from a locus in the genome to a number of external resources, including OMIM, GTEx, RefSeq, and UniProt. The UCSC GENCODE group recently computed GTEx expression quantifications of GENCODE genes and isoforms as individual tissue expression profiles for the GENCODE sets (Figure 13) and Ensembl will add GTEx data later in 2016. In the proposed project period we will update these quantifications with the growing GTEx dataset and recompute the quantifications for each updated GENCODE release and then provide them to the community.

GENCODE is also incorporated into the specific and distinctive tools of the browsers including Ensembl's BioMart, Perl and REST APIs, TreeFam orthology and paralogy annotation and the Variant Effect Predictor (VEP) and well as UCSC's Table Browser, Gene Sorter and Variant Annotation Integrator.

Ensembl is responsible for the data merge to create the comprehensive GENCODE annotation set from HAVANA manual annotation and Ensembl GeneBuild automatic annotation. Thus GENCODE annotation is fully integrated at every step. The UCSC GENCODE group directly handles the ingestion of GENCODE data into the UCSC Browser, which frequently involves updating the browser source base to support GENCODE requirements.

To support users who have not migrated to the new human genome assembly, the UCSC and the HAVANA groups developed a methodology for mapping GENCODE from GRCh38 to GRCh37. This data set is distributed both via UCSC and the genecodegenes.org web site. We will continue to support and enhance this approach and apply it to the next and subsequent versions of the mouse genome assembly.

New interfaces for genomic annotation display and access

Interactive web interfaces To fully utilize the annotation of multiple mouse strain genomes as well as the annotation of a graph-based genome representation, a way of visually highlighting the differences and similarities in annotation is required. Ensembl has a number of static views to view synteny data and viewing smaller regions of differences and will develop these into new dynamic interfaces capable of quickly switching between the various strains (paths) and anchor regions. The goal is to extend the Genoverse scrollable genome browser, which was built for and has been a part of Ensembl since 2012 (www.genoverse.org), to make possible navigation of multiple regions simultaneously (Figure 14).

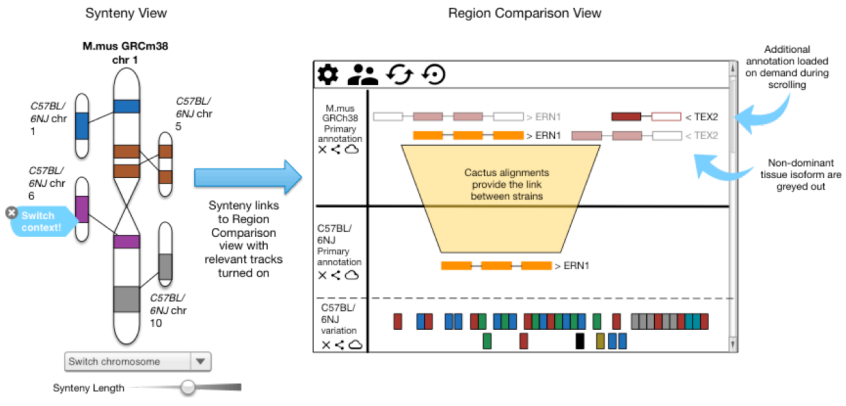


Figure 14: A view of mouse strain supported views moving from high-level synteny views to a configurable client side genome browser. Non-dominant tissue specific isoforms are greyed out.

In addition to the comparison view, we will add the ability in Genoverse to focus on dominant isoforms while greying out hiding others from the display. This will require the ability to select a panel of tissues calculate comparisons among these within the Genoverse web code via metadata attached to the isoforms indicating their status and read by the client web application. Other Ensembl tools will use the same metadata regarding tissue-specificity and isoform dominance. For example, the VEP will be update to prioritize variants according to tissue, as will our sequence searches. Finally our supporting evidence interfaces will be modified to keep pace with the new sources of evidence being generated from this proposal.

Programmatic data distribution We seek to provide the GENCODE annotation through as many sustainable and modern distribution methods as practical. This will require the development of new publicly accessible APIs in addition to continued support for our more traditional pre-generated flat file freezes of the datasets. For example, to support interactive queries we will distribute GENCODE annotation via Global Alliance for Genomics and Health (GA4GH) compatible APIs integrated into the suite of APIs that are funded separately and will be developed at EMBL-EBI. We will also enhance these APIs where appropriate to distribute additional metadata, such as links between genes and

regulatory elements and tissue specific isoforms, as required. GENCODE annotation data freezes will be accompanied by unique identifiers based on annotation checksums generated separately as part of the Transforming Genetic Medicine Initiative (http://www.thetgmi.org). These checksums are intended to become the global unambiguous identifier for these sets. Change sets will also be released identifying new, retired and modified models with every GENCODE release.

We will maintain the current dedicate GENCODE portal (http://www.gencodegenes.org) for dedicated data download and specific project news. Annotation will continue to be made available over FTP and HTTP using common bioinformatics formats including GTF, GFF3, BED and BigBED. In addition we will continue to provide metadata as tab-delimited datasets and as structured JSON. New data will be promoted using the UCSC developed Track Hub system and made available through the EMBL-EBI hosted Track Hub Registry (http://www.trackhubregistry.org).

Software release

The primary output of GENCODE is genome annotation and not software. All annotation will be released without restriction in accordance with EMBL-EBI's terms of use (http://www.ebi.ac.uk/about/terms-of-use). All software developed by Ensembl grant will be released publicly and generally via GitHub including via the existing Ensembl GitHub (https://github.com/Ensembl). Support for the use of the software will be though our existing our RT services linked from the GENCODE portal. The portal will be expanded to provide in-depth documentation and details of the processes used within the GENCODE consortium. We also plan to incorporate information on our software into face-to-face workshops and meetings. All GENCODE software produced by EMBL-EBI is open source and this will continue with the distribution of all project software under the Apache 2.0 license.

Training and outreach

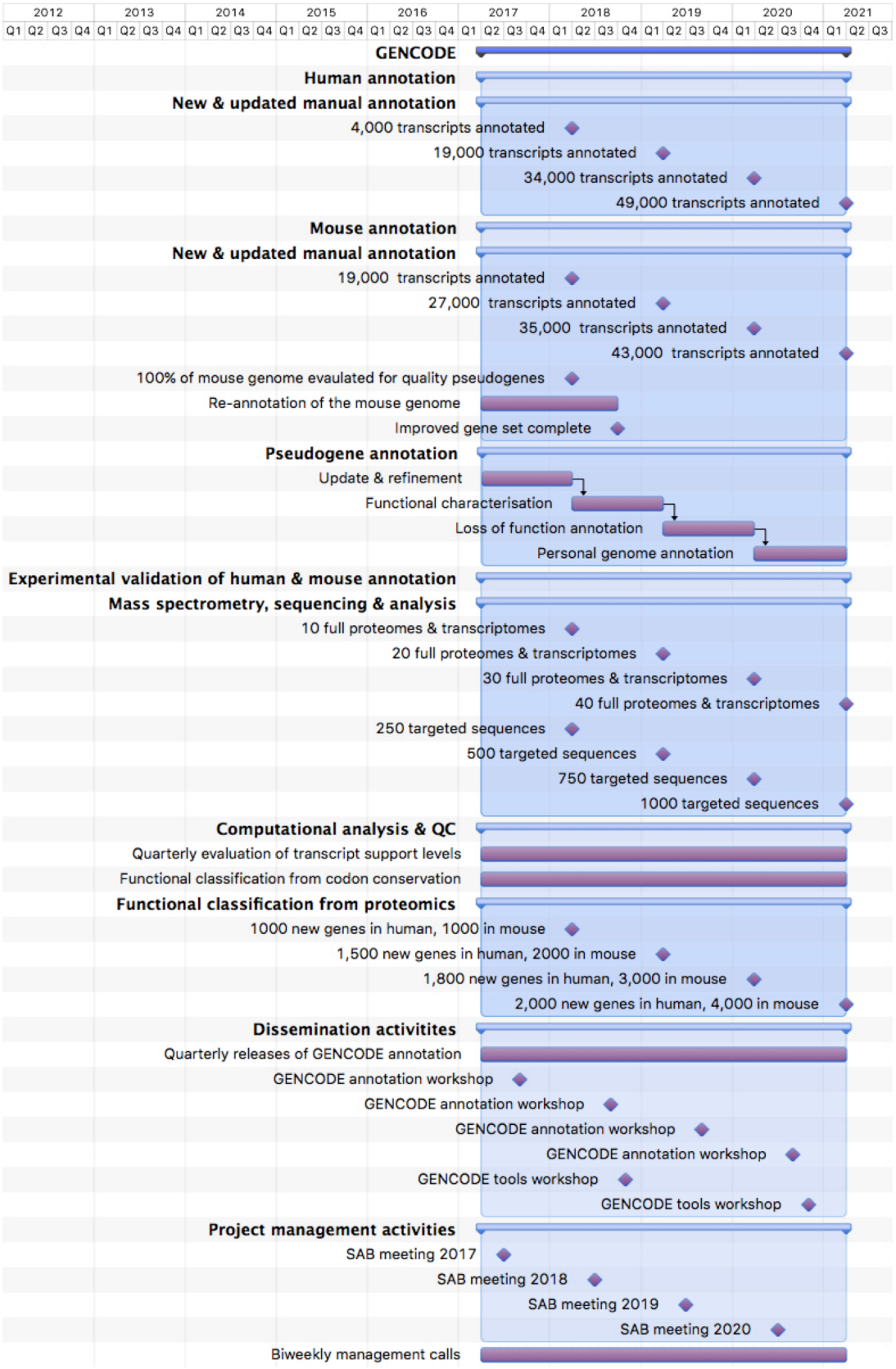
GENCODE will leverage the established Ensembl and EMBL-EBI active worldwide outreach program primarily funded by the core Wellcome Trust Ensembl grant. Ensembl hosts over 100 workshops a year across the US, Europe and Asia. Due to the highly integrated nature of the GENCODE annotation its teaching is integral to all Ensembl workshops. These workshops include details of the GENCODE annotation as well as tools for working with GENCODE datasets for downstream analysis. The UCSC Genome Browser operates a similar worldwide training program including details of the GENCODE annotation and how it can be used within the wider ecosystem of UCSC-developed tools.

We will establish a program presenting one-day workshop per year on GENCODE annotation. These will target researchers in all communities utilizing gene annotation. Our training will cover the various methods of access, including their appropriateness, as well as information on evidence behind annotation calls and utilizing GENCODE annotation in downstream applications such as expression quantification. Training would include promoting the new data types provided by GENCODE and how to best use new annotation such as tissue specificity and population differences between transcripts and the possible impacts on analysis. The location of this workshop would alternate between the US (hosted by UCSC, Yale or MIT) and Europe (hosted by EMBL-EBI) each year.

We will host a one-day workshop in year one and year three on how to use GENCODE annotation tools to annotate genomes and how to submit annotation back to archives. This will be hosted in year one at EMBL-EBI where facilities exist to deliver a workshop to 30 people and may be at EMBL-EBI or other location in year three. We would also supplement this by offering a shorter workshop at the Biocurators conference consisting of a 2-hour program to briefly describe the tools and practices employed when performing manual annotation. This will serve partly as an introduction to the in-depth workshop.

In addition user support will be made available via RT hosted at EMBL-EBI. Consortium members will have access to this

GENCODE milestones



References

1. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M.K., Malhotra, A., Stütz, A.M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M.J.P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H.Y.K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J.M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R.A., Marth, G., Mason, C.E., Menelaou, A., Muzny, D.M., Nelson, B.J., Noor, A., Parrish, N.F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E.E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A.A., Untergasser, A., Walker, J.A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M.A., McCarroll, S.A., Mills, R.E., Gerstein, M.B., Bashir, A., Stegle, O., Devine, S.E., Lee, C., Eichler, E.E. and Korbel, J.O. An integrated map of structural variation in 2,504 human genomes. (2015) *Nature,* **526**, 75-81. PMCID: PMC4617611.

2. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A., Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.-M., Hunt, T., Barnes, I.H.A., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B. and Tyler-Smith, C. A systematic survey of loss-of-function variants in human protein-coding genes. (2012) *Science,* **335**, 823-828. PMCID: PMC3299548.

3. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y. and Pritchard, J.K. RNA splicing is a primary link between genetic variation and disease. (2016) *Science,* **352**, 600-604.

4. Lowe, W.L. and Reddy, T.E. Genomic approaches for understanding the genetics of complex disease. (2015) *Genome Res,* **25**, 1432-1441. PMCID: PMC4579328.

5. Brown, S.D.M. and Moore, M.W. Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. (2012) *Dis Model Mech,* **5**, 289-292. PMCID: PMC3339821.

6. Haendel, M.A., Vasilevsky, N., Brush, M., Hochheiser, H.S., Jacobsen, J., Oellrich, A., Mungall, C.J., Washington, N., Köhler, S., Lewis, S.E., Robinson, P.N. and Smedley, D. Disease insights through cross-species phenotype comparisons. (2015) *Mamm Genome,* **26**, 548-555. PMCID: PMC4602072.

7. Marx, V. The DNA of a nation. (2015) *Nature,* **524**, 503-505.

8. Ashley, E.A. The precision medicine initiative: a new national effort. (2015) *JAMA,* **313**, 2119-2120.

9. GTEx Consortium Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. (2015) *Science,* **348**, 648-660. PMCID: PMC4547484.

10. McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B. and Donnelly, P. Choice of transcripts and software has a large effect on variant annotation. (2014) *Genome Med,* **6**, 26. PMCID: PMC4062061.

11. Zhao, S. and Zhang, B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. (2015) *BMC Genomics,* **16**, 97. PMCID: PMC4339237.

12. Frankish, A., Uszczynska, B., Ritchie, G.R., Gonzalez, J.M., Pervouchine, D., Petryszak, R., Mudge, J.M., Fonseca, N., Brazma, A., Guigo, R. and Harrow, J. Comparison of GENCODE and

RefSeq gene annotation and the impact of reference geneset on variant effect prediction. (2015) *BMC Genomics,* **16**, S2. PMCID: PMC4502323.

13. Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C.D., Rasmussen, M. and Snyder, M.P. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. (2015) *Nat Biotechnol,* **33**, 736-742. PMCID: PMC4832928.

14. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. (2009) *Science,* **324**, 218-223. PMCID: PMC2746483.

15. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. (2003) *Proc Natl Acad Sci U S A,* **100**, 15776-15781. PMCID: PMC307644.

16. Batut, P. and Gingeras, T.R. RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. (2013) *Curr Protoc Mol Biol,* **104**, Unit 25B.11. PMCID: PMC4372803.

17. Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M. and Babak, T. A quantitative atlas of polyadenylation in five mammals. (2012) *Genome Res,* **22**, 1173-1183. PMCID: PMC3371698.

18. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. and Hubbard, T.J. GENCODE: the reference human genome annotation for The ENCODE Project. (2012) *Genome Res,* **22**, 1760-1774. PMCID: PMC3431492.

19. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J. and Guigó, R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. (2012) *Genome Res,* **22**, 1775-1789. PMCID: PMC3431493.

20. Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X.J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T.J., Harrow, J. and Gerstein, M.B. The GENCODE pseudogene resource. (2012) *Genome Biol,* **13**, R51. PMCID: PMC3491395.

21. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S.J., Cunningham, F., Aken, B.L., Zerbino, D.R. and Flicek, P. Ensembl 2016. (2016) *Nucleic Acids Res,* **44**, D710-D716. PMCID: PMC4702834.

22. Lin, M.F., Jungreis, I. and Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. (2011) *Bioinformatics,* **27**, i275-i282. PMCID: PMC3117341.

23. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W.

and Haussler, D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. (2005) *Genome Res,* **15**, 1034-1050. PMCID: PMC1182216.

24. Severin, J., Lizio, M., Harshbarger, J., Kawaji, H., Daub, C.O., Hayashizaki, Y., Bertin, N. and Forrest, A.R.R. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. (2014) *Nat Biotechnol,* **32**, 217-219.

25. Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M. and Gerstein, M. PseudoPipe: an automated pseudogene identification pipeline. (2006) *Bioinformatics,* **22**, 1437-1439. PMCID: Open access at doi://10.1093/bioinformatics/btl116.

26. Zheng, D. and Gerstein, M.B. A computational approach for identifying pseudogenes in the ENCODE regions. (2006) *Genome Biol,* **7 Suppl 1**, S13.1-S1310. PMCID: PMC1810550.

27. Navarro, F.C.P. and Galante, P.A.F. RCPedia: a database of retrocopied genes. (2013) *Bioinformatics,* **29**, 1235-1237. PMCID: PMC3634192.

28. Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garcia Giron, C., Hourlier, T., Howe, K., Kahari, A. and Kokocinski, F. The Ensembl Gene Annotation System. (2016) *Database (Oxford)*, In press. PMCID: PMC Journal - In Process.

29. Collins, J.E., White, S., Searle, S.M.J. and Stemple, D.L. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. (2012) *Genome Res,* **22**, 2067-2078. PMCID: PMC3460200.

30. Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A. and Tress, M.L. APPRIS: annotation of principal and alternative splice isoforms. (2013) *Nucleic Acids Res,* **41**, D110-D117. PMCID: PMC3531113.

31. Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E. and Mattick, J.S. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. (2014) *Nat Protoc,* **9**, 989-1009.

32. Gritsenko, A.A., Hulsman, M., Reinders, M.J.T. and de Ridder, D. Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. (2015) *PLoS Comput Biol,* **11**, e1004336. PMCID: PMC4537299.

33. Dana, A. and Tuller, T. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. (2012) *PLoS Comput Biol,* **8**, e1002755. PMCID: PMC3486846.

34. Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., Andersson, R., Mungall, C.J., Meehan, T.F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y.A., Plessy, C., Vitezic, M., Severin, J., Semple, C.A., Ishizu, Y., Young, R.S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G.M., Arakawa, T., Archer, J.A.C., Arner, P., Babina, M., Rennie, S., Balwierz, P.J., Beckhouse, A.G., Pradhan-Bhatt, S., Blake, J.A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A.M., Califano, A., Cannistraci, C.V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H.C., Dalla, E., Davis, C.A., Detmar, M., Diehl, A.D., Dohi, T., Drabløs, F., Edge, A.S.B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M.C., Faulkner, G.J., Favorov, A.V., Fisher, M.E., Frith, M.C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J.-I., Geijtenbeek, T.B., Gibson, A.P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K.J., Ho Sui, S.J., Hofmann, O.M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B.R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A.S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y.I., Kawashima, T., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Klinken, S.P., Knox, A.J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A.T., Laros, J.F.J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-Sim, A., Manabe, R.-I., Mar, J.C.,

Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D.A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C.L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Noazaki, T., Ogishima, S., Ohkura, N., Ohimiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J.G.D., Rackham, O.J.L., Ramilowski, J.A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M.B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E.A., Schulze-Tanzil, G.G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J.W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R.K., 't Hoen, P.A.C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyodo, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L.M., Verado, R., Vijayan, D., Vorontsov, I.E., Wasserman, W.W., Watanabe, S., Wells, C.A., Winteringham, L.N., Wolvetang, E., Wood, E.J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S.E., Zhang, P.G., Zhao, X., Zucchelli, S., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D.A., Carninci, P. and Hayashizaki, Y. A promoter-level mammalian expression atlas. (2014) *Nature,* **507**, 462-470. PMCID: PMC4529748.

35. Tilgner, H., Raha, D., Habegger, L., Mohiuddin, M., Gerstein, M. and Snyder, M. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. (2013) *G3 (Bethesda),* **3**, 387-397. PMCID: PMC3583448.

36. Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. A single-molecule long-read survey of the human transcriptome. (2013) *Nat Biotechnol,* **31**, 1009-1014. PMCID: PMC4075632.

37. Tilgner, H., Grubert, F., Sharon, D. and Snyder, M.P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. (2014) *Proc Natl Acad Sci U S A,* **111**, 9869-9874. PMCID: PMC4103364.

38. Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M.L. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. (2014) *Hum Mol Genet,* **23**, 5866-5878. PMCID: PMC4204768.

39. Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D.N., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S.K., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H. and Pandey, A. A draft map of the human proteome. (2014) *Nature,* **509**, 575-581. PMCID: PMC4403737.

40. Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. and Kuster, B. Mass-spectrometry-based draft of the human proteome. (2014) *Nature,* **509**, 582-587.

41. Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N. and Aebersold, R. The PeptideAtlas project. (2006) *Nucleic Acids Res,* **34**, D655-D658. PMCID: PMC1347403.

42. Wright, J.C., Mudge, J., Weisser, H., Barzine, M.P., Gonzalez, J.M., Brazma, A., Choudhary, J.S. and Harrow, J. Improving GENCODE Reference Gene Annotation Using High Stringency Proteomics Workflow. (2016) *Nat Commun*, In Press. PMCID: PMC Journal - In Process.

43. Balasubramanian, S., Zheng, D., Liu, Y.-J., Fang, G., Frankish, A., Carriero, N., Robilotto, R., Cayting, P. and Gerstein, M. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. (2009) *Genome Biol,* **10**, R2. PMCID: PMC2687790.

44. Zhang, Z.L., Harrison, P.M. and Gerstein, M. Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes. (2002) *J Mol Biol,* **323**, 811-822.

45. Liu, Y.-J., Zheng, D., Balasubramanian, S., Carriero, N., Khurana, E., Robilotto, R. and Gerstein, M.B. Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity. (2009) *BMC Genomics,* **10**, 480. PMCID: PMC2770531.

46. Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J. and Gerstein, M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. (2010) *Genome Biol,* **11**, R26. PMCID: PMC2864566.

47. Balasubramanian, S., Habegger, L., Frankish, A., MacArthur, D.G., Harte, R., Tyler-Smith, C., Harrow, J. and Gerstein, M. Gene inactivation and its implications for annotation in the era of personal genomics. (2011) *Genes Dev,* **25**, 1-10. PMCID: PMC3012931.

48. E-MTAB-513 - RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map). http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/.

49. Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.-M., Cao, X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J., Iyer, M.K., Barrette, T., Shanmugam, A., Dhanasekaran, S.M., Palanisamy, N. and Chinnaiyan, A.M. Expressed pseudogenes in the transcriptional landscape of human cancers. (2012) *Cell,* **149**, 1622-1634. PMCID: PMC3597446.

50. Poliseno, L. Pseudogenes: newly discovered players in human cancer. (2012) *Sci Signal,* **5**, re5.

51. Abyzov, A., Iskow, R., Gokcumen, O., Radke, D.W., Balasubramanian, S., Pei, B., Habegger, L., Lee, C. and Gerstein, M. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. (2013) *Genome Res,* **23**, 2042-2052. PMCID: PMC3847774.

52. Havana Annotation Guidelines Version 24: 30 March 2016. *Havana Annotation Guidelines Version 24: 30 March 2016,* ftp://ftp.sanger.ac.uk/pub/project/havana/Guidelines/Guidelines_March_2016.pdf.

53. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. (1990) *J Mol Biol,* **215**, 403-410.

54. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. (1997) *Comput Appl Biosci,* **13**, 477-478.

55. Baertsch, R., Diekhans, M., Kent, W.J., Haussler, D. and Brosius, J. Retrocopy contributions to the evolution of the human genome. (2008) *BMC Genomics,* **9**, 466. PMCID: PMC2584115.

56. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. and Bateman, A. The Pfam protein families database: towards a more sustainable future. (2016) *Nucleic Acids Res,* **44**, D279-D285. PMCID: PMC4702930.

57. RNAcentral Consortium RNAcentral: an international database of ncRNA sequences. (2015) *Nucleic Acids Res,* **43**, D123-D129. PMCID: PMC4384043.

58. Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P. and Ulitsky, I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. (2015) *Cell Rep,* **11**, 1110-1122. PMCID: PMC4576741.

59. Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S.E., Behr, J., Bertone, P., Bohnert, R., Bucher, P., Cloonan, N., Derrien, T., Djebali, S., Du, J., Dudoit, S., Gerstein, M., Gingeras, T.R., Gonzalez, D., Grimmond, S.M., Guigó, R., Habegger, L., Harrow, J., Hubbard, T.J., Iseli, C., Jean, G., Kahles, A., Lagarde, J., Leng, J., Lefebvre, G., Lewis, S., Mortazavi, A., Niermann, P., Rätsch, G., Reymond, A., Ribeca, P., Richard, H., Rougemont, J., Rozowsky, J., Sammeth, M., Sboner, A., Schulz, M.H., Searle, S.M.J., Solorzano, N.D., Solovyev, V., Stanke, M., Stevenson, B.J., Stockinger, H., Valsesia, A., Weese, D., White, S., Wold, B.J., Wu, J., Wu, T.D., Zeller, G., Zerbino, D. and Zhang, M.Q. Assessment of transcript reconstruction methods for RNA-seq. (2013) *Nat Methods,* **10**, 1177-1184. PMCID: PMC3851240.

60. Jaffe, A.E., Shin, J., Collado-Torres, L., Leek, J.T., Tao, R., Li, C., Gao, Y., Jia, Y., Maher, B.J., Hyde, T.M., Kleinman, J.E. and Weinberger, D.R. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. (2015) *Nat Neurosci,* **18**, 154-161. PMCID: PMC4281298.

61. Wollerton, M.C., Gooding, C., Wagner, E.J., Garcia-Blanco, M.A. and Smith, C.W.J. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. (2004) *Mol Cell,* **13**, 91-100.

62. Sureau, A., Gattoni, R., Dooghe, Y., Stévenin, J. and Soret, J. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. (2001) *EMBO J,* **20**, 1785-1796. PMCID: PMC145484.

63. Wong, J.J.-L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., Thoeng, A., Khoo, T.-L., Bailey, C.G., Holst, J. and Rasko, J.E.J. Orchestrated intron retention regulates normal granulocyte differentiation. (2013) *Cell,* **154**, 583-595.

64. Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M. and Blencowe, B.J. Widespread intron retention in mammals functionally tunes transcriptomes. (2014) *Genome Res,* **24**, 1774-1786. PMCID: PMC4216919.

65. Lynch, D.C., Revil, T., Schwartzentruber, J., Bhoj, E.J., Innes, A.M., Lamont, R.E., Lemire, E.G., Chodirker, B.N., Taylor, J.P., Zackai, E.H., McLeod, D.R., Kirk, E.P., Hoover-Fong, J., Fleming, L., Savarirayan, R., Majewski, J., Jerome-Majewska, L.A., Parboosingh, J.S. and Bernier, F.P. Disrupted auto-regulation of the spliceosomal gene SNRPB causes cerebro-costo-mandibular syndrome. (2014) *Nat Commun,* **5**, 4483. PMCID: PMC4109005.

66. Searle, S.M.J., Gilbert, J., Iyer, V. and Clamp, M. The otter annotation system. (2004) *Genome Res,* **14**, 963-970. PMCID: PMC479127.

67. Kokocinski, F., Harrow, J. and Hubbard, T. AnnoTrack--a tracking system for genome annotation. (2010) *BMC Genomics,* **11**, 538. PMCID: PMC3091687.

68. Barson, G. and Griffiths, E. SeqTools: visual tools for manual analysis of sequence alignments. (2016) *BMC Res Notes,* **9**, 39. PMCID: PMC4724122.

69. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. Repbase Update, a database of eukaryotic repetitive elements. (2005) *Cytogenet Genome Res,* **110**, 462-467.

70. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. (1999) *Nucleic Acids Res,* **27**, 573-580. PMCID: PMC148217.

71. Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., Balasubramanian, S., Harte, R., Wang, D., Rutenberg-Schoenberg, M., Clark, W., Diekhans, M., Rozowsky, J., Hubbard, T., Harrow, J. and Gerstein, M.B. Comparative analysis of pseudogenes across three phyla. (2014) *Proc Natl Acad Sci U S A,* **111**, 13361-13366. PMCID: PMC4169933.

72. Holford, M.E., Khurana, E., Cheung, K.-H. and Gerstein, M. Using semantic web rules to reason on an ontology of pseudogenes. (2010) *Bioinformatics,* **26**, i71-i78. PMCID: PMC2881358.

73. Byvatov, E. and Schneider, G. Support vector machine applications in bioinformatics. (2003) *Appl Bioinformatics,* **2**, 67-77.

74. Rodriguez, J.M., Carro, A., Valencia, A. and Tress, M.L. APPRIS WebServer and WebServices. (2015) *Nucleic Acids Res,* **43**, W455-W459. PMCID: PMC4489225.

75. Tress, M.L., Wesselink, J.-J., Frankish, A., López, G., Goldman, N., Löytynoja, A., Massingham, T., Pardi, F., Whelan, S., Harrow, J. and Valencia, A. Determination and validation of principal gene products. (2008) *Bioinformatics,* **24**, 11-17. PMCID: PMC2734078.

76. Farrell, C.M., O'Leary, N.A., Harte, R.A., Loveland, J.E., Wilming, L.G., Wallin, C., Diekhans, M., Barrell, D., Searle, S.M.J., Aken, B., Hiatt, S.M., Frankish, A., Suner, M.-M., Rajput, B., Steward, C.A., Brown, G.R., Bennett, R., Murphy, M., Wu, W., Kay, M.P., Hart, J., Rajan, J., Weber, J., Snow, C., Riddick, L.D., Hunt, T., Webb, D., Thomas, M., Tamez, P., Rangwala, S.H., McGarvey, K.M., Pujar, S., Shkeda, A., Mudge, J.M., Gonzalez, J.M., Gilbert, J.G.R., Trevanion, S.J., Baertsch, R., Harrow, J.L., Hubbard, T., Ostell, J.M., Haussler, D. and Pruitt, K.D. Current status and new features of the Consensus Coding Sequence database. (2014) *Nucleic Acids Res,* **42**, D865-D872. PMCID: PMC3965069.

77. Ezkurdia, I., Rodriguez, J.M., Carrillo-de Santa Pau, E., Vázquez, J., Valencia, A. and Tress, M.L. Most highly expressed protein-coding genes have a single dominant isoform. (2015) *J Proteome Res,* **14**, 1880-1887. PMCID: PMC4768900.

78. Pundir, S., Magrane, M., Martin, M.J., O'Donovan, C. and UniProt Consortium Searching and Navigating UniProt Databases. (2015) *Curr Protoc Bioinformatics,* **50**, 1.27.1-1.2710. PMCID: PMC4522465.

79. Clark, M.B., Mercer, T.R., Bussotti, G., Leonardi, T., Haynes, K.R., Crawford, J., Brunck, M.E., Cao, K.-A.L., Thomas, G.P., Chen, W.Y., Taft, R.J., Nielsen, L.K., Enright, A.J., Mattick, J.S. and Dinger, M.E. Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. (2015) *Nat Methods,* **12**, 339-342.

80. Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S. and Rinn, J.L. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. (2012) *Nat Biotechnol,* **30**, 99-104. PMCID: PMC3710462.

81. Brosch, M., Saunders, G.I., Frankish, A., Collins, M.O., Yu, L., Wright, J., Verstraten, R., Adams, D.J., Harrow, J., Choudhary, J.S. and Hubbard, T. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. (2011) *Genome Res,* **21**, 756-767. PMCID: PMC3083093.

82. Gascoigne, D.K., Cheetham, S.W., Cattenoz, P.B., Clark, M.B., Amaral, P.P., Taft, R.J., Wilhelm, D., Dinger, M.E. and Mattick, J.S. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. (2012) *Bioinformatics,* **28**, 3042-3050.

83. Kumar, D., Mondal, A.K., Kutum, R. and Dash, D. Proteogenomics of rare taxonomic phyla: A prospective treasure trove of protein coding genes. (2016) *Proteomics,* **16**, 226-240.

84. Deutsch, E.W., Sun, Z., Campbell, D., Kusebauch, U., Chu, C.S., Mendoza, L., Shteynberg, D., Omenn, G.S. and Moritz, R.L. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. (2015) *J Proteome Res,* **14**, 3461-3473. PMCID: PMC4755269.

85. Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J.M., del Pozo, A., Vázquez, J., Valencia, A. and Tress, M.L. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. (2015) *PLoS Comput Biol,* **11**, e1004325. PMCID: PMC4465641.

86. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. (2005) *Genome Biol,* **6**, R44. PMCID: PMC1175956.

87. Ezkurdia, I., Calvo, E., Del Pozo, A., Vázquez, J., Valencia, A. and Tress, M.L. The potential clinical impact of the release of two drafts of the human proteome. (2015) *Expert Rev Proteomics,* **12**, 579-593. PMCID: PMC4732427.

88. Savitski, M.M., Wilhelm, M., Hahne, H., Kuster, B. and Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. (2015) *Mol Cell Proteomics,* **14**, 2394-2404. PMCID: PMC4563723.

89. Cooper, B. The problem with peptide presumption and the downfall of target-decoy false discovery rates. (2012) *Anal Chem,* **84**, 9663-9667.

90. Bonzon-Kulichenko, E., Garcia-Marques, F., Trevisan-Herraz, M. and Vázquez, J. Revisiting peptide identification by high-accuracy mass spectrometry: problems associated with the use of narrow mass precursor windows. (2015) *J Proteome Res,* **14**, 700-710.

91. Reiter, L., Claassen, M., Schrimpf, S.P., Jovanovic, M., Schmidt, A., Buhmann, J.M., Hengartner, M.O. and Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. (2009) *Mol Cell Proteomics,* **8**, 2405-2417. PMCID: PMC2773710.

92. Ensembl Stable IDs. http://www.ensembl.org/info/genome/stable_ids/versions.html.

93. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. (2012) *Nature,* **489**, 57-74. PMCID: PMC3439153.

94. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. (2009) *Genome Res,* **19**, 327-335. PMCID: PMC2652215.

95. Wethmar, K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. (2014) *Wiley Interdiscip Rev RNA,* **5**, 765-778.

96. Capell, A., Fellerer, K. and Haass, C. Progranulin transcripts with short and long 5' untranslated regions (UTRs) are differentially expressed via posttranscriptional and translational repression. (2014) *J Biol Chem,* **289**, 25879-25889. PMCID: PMC4162188.

97. Johnstone, T.G., Bazzini, A.A. and Giraldez, A.J. Upstream ORFs are prevalent translational repressors in vertebrates. (2016) *EMBO J,* **35**, 706-723. PMCID: PMC4818764.

98. Barbosa, C., Peixeiro, I. and Romão, L. Gene expression regulation by upstream open reading frames and human disease. (2013) *PLoS Genet,* **9**, e1003529. PMCID: PMC3738444.

99. Calvo, S.E., Pagliarini, D.J. and Mootha, V.K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. (2009) *Proc Natl Acad Sci U S A,* **106**, 7507-7512. PMCID: PMC2669787.

100. Habegger, L., Balasubramanian, S., Chen, D.Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M. and Gerstein, M. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. (2012) *Bioinformatics,* **28**, 2267-2269. PMCID: PMC3426844.

101. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M. and Gerstein, M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. (2011) *Mol Syst Biol,* **7**, 522. PMCID: PMC3208341.

102. Steward, C.A., Gonzalez, J.M., Trevanion, S., Sheppard, D., Kerry, G., Gilbert, J.G.R., Wicker, L.S., Rogers, J. and Harrow, J.L. The non-obese diabetic mouse sequence, annotation and variation resource: an aid for investigating type 1 diabetes. (2013) *Database (Oxford),* **2013**, bat032. PMCID: PMC3668384.

103. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. (2010) *Bioinformatics,* **26**, 2069-2070. PMCID: PMC2916720.

104. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. and Gingeras, T.R. Landscape of transcription in human cells. (2012) *Nature,* **489**, 101-108. PMCID: PMC3684276.

105. Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.-K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A.P., Cayting, P., Charos, A., Chen, D.Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E.C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T.E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K.Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P.J., Myers, R.M., Weissman, S.M. and Snyder, M. Architecture of the human regulatory network derived from ENCODE data. (2012) *Nature,* **489**, 91-100. PMCID: PMC4154057.

106. Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U.S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüs, Z.H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S.M., Macarthur, D.G., Marth, G., Muzny, D., Pers, T.H., Ritchie, G.R.S., Rosenfeld, J.A., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., Dermitzakis, E.T., Yu, H., Rubin, M.A., Tyler-Smith, C. and Gerstein, M. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. (2013) *Science,* **342**, 1235587. PMCID: PMC3947637.

107. Chen, J., Rozowsky, J., Galeev, T.R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L. and Gerstein, M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. (2016) *Nat Commun,* **7**, 11101. PMCID: PMC4837449.

108. The 1000 Genomes Project Consortium A global reference for human genetic variation. (2015) *Nature,* **526**, 68-74. PMCID: PMC4750478.

109. Tools for working with variation graphs. https://github.com/vgteam/vg.

110. Novak, A.M., Garrison, E. and Paten, B. A Graph Extension of the Positional Burrows-Wheeler Transform and its Applications. *bioRxiv,*, http://dx.doi.org/10.1101/051409.

111. Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., James, M., Liu, P., Tichelaar, J.W., Vikis, H.G., You, M. and Mills, I.G. Principles for the post-GWAS functional characterization of cancer risk loci. (2011) *Nat Genet,* **43**, 513-518. PMCID: PMC3325768.

112. McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke, D.B., Schaar, B.T., Wenger, A.M., Bejerano, G. and Kingsley, D.M. Human-specific loss of regulatory DNA and the evolution of human-specific traits. (2011) *Nature,* **471**, 216-219. PMCID: PMC3071156.

113. Levine, M. and Tjian, R. Transcription regulation and animal diversity. (2003) *Nature,* **424**, 147-151.

114. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., Dunham, I., Elnitski, L.L., Farnham, P.J., Feingold, E.A., Gerstein, M., Giddings, M.C., Gilbert, D.M., Gingeras, T.R., Green, E.D., Guigo, R., Hubbard, T., Kent, J., Lieb, J.D., Myers, R.M., Pazin, M.J., Ren, B., Stamatoyannopoulos, J.A., Weng, Z., White, K.P. and Hardison, R.C. Defining functional DNA elements in the human genome. (2014) *Proc Natl Acad Sci U S A,* **111**, 6131-6138. PMCID: PMC4035993.

115. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and Bernstein, B.E. Mapping and analysis of chromatin state dynamics in nine human cell types. (2011) *Nature,* **473**, 43-49. PMCID: PMC3088773.

116. Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T. and Flicek, P.R. The Ensembl Regulatory Build. (2015) *Genome Biol,* **16**, 56. PMCID: PMC4407537.

117. ENCODE Encyclopedia: Genomic Annotations. *ENCODE Encyclopedia: Genomic Annotations,*, https://www.encodeproject.org/data/annotations/.

118. Merkenschlager, M. and Odom, D.T. CTCF and cohesin: linking gene regulatory elements with their targets. (2013) *Cell,* **152**, 1285-1297.

119. Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K. and Gilad, Y. Genomic variation. Impact of regulatory variation from RNA to protein. (2015) *Science,* **347**, 664-667. PMCID: PMC4507520.

120. Albert, F.W. and Kruglyak, L. The role of regulatory variation in complex traits and disease. (2015) *Nat Rev Genet,* **16**, 197-212.

121. Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R. and Higgs, D.R. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. (2014) *Nat Genet,* **46**, 205-212.

122. Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaöz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., Carter, N.P., Vetrie, D. and Dunham, I. The landscape of histone modifications across 1% of the human genome in five human cell lines. (2007) *Genome Res,* **17**, 691-707. PMCID: PMC1891331.

123. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M. and Sandelin, A. An atlas of active enhancers across human cell types and tissues. (2014) *Nature,* **507**, 455-461.

124. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen,

T.S. and Thomson, J.A. The NIH Roadmap Epigenomics Mapping Consortium. (2010) *Nat Biotechnol,* **28**, 1045-1048. PMCID: PMC3607281.

125. Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E.T., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Giehl, C., Graf, T., Grosveld, F., Guigo, R., Gut, I., Helin, K., Jarvius, J., Küppers, R., Lehrach, H., Lengauer, T., Lernmark, A., Leslie, D., Loeffler, M., Macintyre, E., Mai, A., Martens, J.H., Minucci, S., Ouwehand, W.H., Pelicci, P.G., Pendeville, H., Porse, B., Rakyan, V., Reik, W., Schrappe, M., Schübeler, D., Seifert, M., Siebert, R., Simmons, D., Soranzo, N., Spicuglia, S., Stratton, M., Stunnenberg, H.G., Tanay, A., Torrents, D., Valencia, A., Vellenga, E., Vingron, M., Walter, J. and Willcocks, S. BLUEPRINT to decode the epigenetic signature written in blood. (2012) *Nat Biotechnol,* **30**, 224-226.

126. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E. and Stamatoyannopoulos, J.A. The accessible chromatin landscape of the human genome. (2012) *Nature,* **489**, 75-82. PMCID: PMC3721348.

127. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavaré, S., Deloukas, P., Hurles, M.E. and Dermitzakis, E.T. Relative impact of nucleotide and copy number variation on gene expression phenotypes. (2007) *Science,* **315**, 848-853. PMCID: PMC2665772.

128. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C.C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G.M., Abecasis, G.R. and Cookson, W.O.C. A genome-wide association study of global gene expression. (2007) *Nat Genet,* **39**, 1202-1207.

129. Dostie, J. and Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. (2007) *Nat Protoc,* **2**, 988-1002.

130. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G.Y., Huang, P.Y.H., Welboren, W.-J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D.S.A., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K.V., Thomsen, J.S., Lee, Y.K., Karuturi, R.K.M., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E.T., Wei, C.-L., Cheung, E. and Ruan, Y. An oestrogen-receptor-alpha-bound human chromatin interactome. (2009) *Nature,* **462**, 58-64. PMCID: PMC2774924.

131. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. and Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. (2009) *Science,* **326**, 289-293. PMCID: PMC2858594.

132. Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., Dimitrova, E., Dimond, A., Edelman, L.B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., LeProust, E., Osborne, C.S., Mitchell, J.A., Luscombe, N.M. and Fraser, P. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. (2015) *Genome Res,* **25**, 582-597. PMCID: PMC4381529.

133. Zerbino, D.R., Johnson, N., Juetteman, T., Sheppard, D., Wilder, S.P., Lavidas, I., Nuhn, M., Perry, E., Raffaillac-Desfosses, Q., Sobral, D., Keefe, D., Gräf, S., Ahmed, I., Kinsella, R., Pritchard, B., Brent, S., Amode, R., Parker, A., Trevanion, S., Birney, E., Dunham, I. and Flicek, P. Ensembl regulation resources. (2016) *Database (Oxford),* **2016**, bav119. PMCID: PMC4756621.

134. Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., Abdennur, N.A., Liu, J., Svensson, P.-A., Hsu, Y.-H., Drucker, D.J., Mellgren, G., Hui, C.-C., Hauner, H. and Kellis, M. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. (2015) *N Engl J Med,* **373**, 895-907.

135. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B. and Ecker, J.R. Human DNA methylomes at base resolution show widespread epigenomic differences. (2009) *Nature,* **462**, 315-322. PMCID: PMC2857523.

136. Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E. and Stamatoyannopoulos, J.A. Circuitry and dynamics of human transcription factor regulatory networks. (2012) *Cell,* **150**, 1274-1286. PMCID: PMC3679407.

137. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., Urban, A.E., Montgomery, S.B., Levinson, D.F. and Koller, D. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. (2014) *Genome Res,* **24**, 14-24. PMCID: PMC3875855.

138. Xia, K., Shabalin, A.A., Huang, S., Madar, V., Zhou, Y.-H., Wang, W., Zou, F., Sun, W., Sullivan, P.F. and Wright, F.A. seeQTL: a searchable database for human eQTLs. (2012) *Bioinformatics,* **28**, 451-452. PMCID: PMC3268245.

139. Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., Aldred, S.F., Trinklein, N.D., Schuetz, E., Nickerson, D.A., Thummel, K.E., Rieder, M.J., Rettie, A.E., Ratain, M.J., Cox, N.J. and Brown, C.D. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. (2011) *PLoS Genet,* **7**, e1002078. PMCID: PMC3102751.

140. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., Hedman, A.K., Bataille, V., Tzenova Bell, J., Surdulescu, G., Dimas, A.S., Ingle, C., Nestle, F.O., di Meglio, P., Min, J.L., Wilk, A., Hammond, C.J., Hassanali, N., Yang, T.-P., Montgomery, S.B., O'Rahilly, S., Lindgren, C.M., Zondervan, K.T., Soranzo, N., Barroso, I., Durbin, R., Ahmadi, K., Deloukas, P., McCarthy, M.I., Dermitzakis, E.T., Spector, T.D. and The MuTHER Consortium The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. (2011) *PLoS Genet,* **7**, e1002003. PMCID: PMC3033383.

141. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. and Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. (2012) *Cell Res,* **22**, 490-503. PMCID: PMC3292289.

142. Zheng, J., Chen, Y., Deng, F., Huang, R., Petersen, F., Ibrahim, S. and Yu, X. mtDNA sequence, phylogeny and evolution of laboratory mice. (2014) *Mitochondrion,* **17**, 126-131.