Application Table of Contents

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 for 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 as the primary method to increase the quality and completeness of the protein coding annotation. Key well-established technologies for this aim include annotation based on cDNA, EST, RNA-seq and mass spectroscopy data as well as core informatics methods for gene annotation and coding potential. We will investigate the best approaches to incorporate long transcriptome data (Iso-Seq) and other relevant emerging technologies. Non-coding annotation will build on our experience of the last four years during which we have created the most complete and comprehensive non-coding gene sets. We will expand and validate the methods that we use to annotate full-length non-coding genes such as long read RACEseq and other approaches.

Aim 2: Population based genome annotation

The overall goal of this aim is to ensure that any transcript isoform 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/prioritisation of variable transcripts from population transcriptomics datasets such as GTEx. Finally, as graph genome representations mature, GENCODE will pilot methods to present its annotation on a graph representation of the genome that fully incorporates population and/or individual variation as graph methods mature.

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

This aim will seek to bring new data types that directly connect transcripts to relevant regulatory regions and thus annotate a more comprehensive definition of what is a gene. We will proceed as a series of pilot projects within GENCODE focused on data generated to initially measure polymerase recruitment and transcription initiation, epigenomes, cis-regulatory interactions and physical interactions. The most informative of these datasets will be incorporated into the GENCODE annotations using a combination of computational and manual approaches.

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 provisional support for distribution of GENCODE annotation via Global Alliance for Genomics and Health (GA4GH) compatible APIs. We will establish new mechanisms for prioritising genes for manual annotation with community input and seek to establish GENCODE as the standard annotation set leading research and clinical genomics efforts.

Research Strategy (Overall): 6 pages

The overview should explain the rationale for the community resource, describe the resource to be generated, document community support for the proposed resource, and explain the anticipated impact of the resource widely across biomedical research. The overview should also include the project's elements, including the technologies that will be used to produce the resource. Applications proposing to develop new databases, repositories, or other resources should clearly explain why there is a need or why the current resources are not adequate.

Describe of the concepts, methods, technologies, treatments, services, or preventative interventions that drive this field will be changed if the proposed aims are achieved.

Organize the Research Strategy in the specified order and using the instructions provided below, or as stated in the Funding Opportunity Announcement. Start each section with the appropriate section heading - Significance, Innovation, Approach. Cite published experimental details in the Research Strategy section and provide the full reference in Section G.220 - R&R Other Project Information Form, Bibliography and Reference Cited.

1. Significance

- *Explain the importance of the problem or critical barrier to progress in the field that the proposed project addresses.*
- *Describe the scientific premise for the proposed project, including consideration of the strengths and weaknesses of published research or preliminary data crucial to the support of your application.*
- *Explain how the proposed project will improve scientific knowledge, technical capability, and/or clinical practice in one or more broad fields.*

2. Innovation

- *Explain how the application challenges and seeks to shift current research or clinical practice paradigms.*
- *Describe any novel theoretical concepts, approaches or methodologies, instrumentation or interventions to be developed or used, and any advantage over existing methodologies, instrumentation, or interventions.*
- *Explain any refinements, improvements, or new applications of theoretical concepts, approaches or methodologies, instrumentation, or interventions.*

3. Approach

- *Describe the overall strategy, methodology, and analyses to be used to accomplish the specific aims of the project. Describe the experimental design and methods proposed and how they will achieve robust and unbiased results. Unless addressed separately in Item 15 (Resource Sharing Plan), include how the data will be collected, analyzed, and interpreted as well as any resource sharing plans as appropriate.*
- *Discuss potential problems, alternative strategies, and benchmarks for success anticipated to achieve the aims.*
- *If the project is in the early stages of development, describe any strategy to establish feasibility, and address the management of any high risk aspects of the proposed work.*
- *Explain how relevant biological variables, such as sex, are factored into research designs and analyses for studies in vertebrate animals and humans. For example, strong justification from the scientific literature, preliminary data, or other relevant considerations, must be provided for applications proposing to study only one sex.*
- *If your study(s) involves human subjects, the sections on the Inclusion of Women and Minorities and Inclusion of Children can be used to expand your discussion on inclusion and justify the proposed proportions of individuals (such as males and females) in the sample, but it must also be addressed here in the Approach section.*
- *Please refer to NOT-OD-15-102 for further consideration of NIH expectations about sex as a biological variable.*
- *If research on Human Embryonic Stem Cells (hESCs) is proposed but an approved cell line from the NIH hESC Registry cannot be identified, provide a strong justification for why an appropriate cell line cannot be chosen from the Registry at this time.*

If an applicant has multiple Specific Aims, then the applicant may address Significance, Innovation and Approach for each Specific Aim individually, or may address Significance, Innovation and Approach for all of the Specific Aims collectively.

As applicable, also include the following information as part of the Research Strategy, keeping within the three sections listed above: Significance, Innovation, and Approach.

Significance

The mapping of the human genome and the resulting reference human assembly is one of the great scientific products of the 21st century. We are now on the cusp of 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 understaing of a 'normal' genome has emerged and understanding the range of normal is critical to defining what it means to have a genetic disease. Indeed, the variety of the genome has often been surprising. We have discovered that structural and copy number variation is pervasive (cite history and current) and consequential, we have found that everyone's genome contains a significant number of protein truncating or loss of function mutations (cite MacArthur and others) and we are only beginning to understand the spectrum of functional sequence changes that occur in and modify disease causing pathways (cite).

Highly accurate genome annotation is a vital foundation to these studies and a critical companion to the planned large-scale initiatives to sequence humans for research and clinical care. 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 lead eventually lead to an error in interpretation. Many of these interpretation errors will be inconsequential, some will not.

Overall goals and established usage

The objective of the GENCODE consortium is to create this foundational reference genome annotation. Our overall goal is to identify and classify all evidence-based gene features in the human and mouse genomes with high accuracy and release these annotations for the greatest possible benefit for biomedical research and genome interpretation. GENCODE focuses on protein-coding and non-coding loci including alternatively spliced isoforms and pseudogenes. It relies on a series of well-tested manual and computational methods within a high functioning consortium to produce regular annotation releases. We will continue our successful approach to investigate and incorporate new technologies and new annotation types via a series of well chosen pilot projects addressing Iso-Seq data, population-based gene expression, regulatory regions and other topics that complement major funded projects and resources with long-term connections or relevance to the GENCODE consortium.

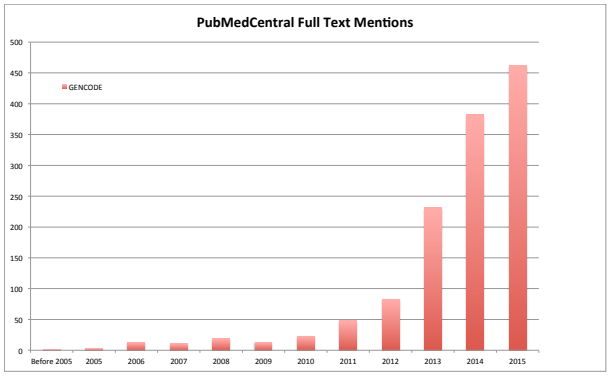


Figure 1: Number of times each year that the text "GENCODE" appears in PubMedCentral. This is a full text search of only the articles that are in PMC, and thus undercounts usage because only a fraction of papers are there. Note before the project, there is apparently only one mention in PMC.

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 at UCSC. UCSC also provides the mouse GENCODE set, but not yet as default (this is planned for the near future). GENCODE is the gene set used for major projects including the Exome Aggregation Consortium (EXAC), GTEx, 1000 Genomes Project and ENCODE. GENCODE engages directly with the Mouse Genome Informatics (MGI) resource at the Jackson Laboratory and with NCBI as part of the Consensus Coding Sequence (CCDS) project and the CCDS gene set is then used in Ensembl, UCSC and other places. The International Mouse Phenotyping Consortium (IMPC) uses the mouse gene set arising from this work.

Although the uses of GENCODE are not always correctly cited or the citations are not to GENCODE: users may cite Ensembl or UCSC, the growth of GENCODE usage has been dramatic over the past four years (Figure 1). Even with this undercounting, Google scholar has more than 2200 citations for the main GENCODE papers.

There have been several comparisons conducted by independent groups of the GENCODE genes to other gene sets for various purposes and these universally recommend the use of GENCODE as the best human annotation. We have also done specific comparisons and published our results. These efforts have helped us understand exactly how the GENCODE annotation is used and were partly responsible for some changes to the resources that we provide including the introduction of GENCODE-Basic as a way to deal with a concern that number of GENCODE transcripts may make RNA-seq analysis more complicated.

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 it encodes remains incomplete. This insufficiency applies to all classes of genic features; protein-coding genes, pseudogenes, long non-coding RNAs and small RNAs. The deficit manifests at mutiple levels; complete absence of annotation, partial annotation, under annotation and mis-annotation. 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 count for recent GENCODE releases, it is likely that the majority of unannotated loci will be 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 extent, almost certainly because it is based on non-full-length or truncated evidence e.g. ESTs. Underannotation, which often co-occurs with partial annotation, happens where a feature is present in the geneset but is lacking the level of functional annotation it is possible to add; for example where a transcript in a protein-coding locus starts or ends with a novel internal exon, no CDS is added given the uncertainty over whether the true start or end of the transcript has been found. Misannotation, where incorrect structural or functional annotation is present can be attributable to error, although GENCODE's extensive QC seeks to reduce this to a minimum, or more likely absence of required orthogonal dataset at the time of annotation; for example at the time of annotation a locus may be annotated as a one biotype but having more evidence may clarify the structure and functional potential allowing the biotype to be updated. All these modes apply to the reference genome sequence and patch and haplotype sequences created by the genome reference consortium (GRC), however, genes and transcripts not present on these sequences eg in alternative haplotypes will not be captured.

The emergence and improvement of 3rd generation sequencing technologies such as PacBio and SLRseq together with the extension of recent techniques based on second generation short-read sequences such as RP, CAGE, RAMPAGE and polyAseq and Mass Spectrometry will individually all us to reduce the degree of error and incompleteness. However, GENCODE's strength has been derived from its ability to integrate multiple different data-types to achieve the best possible annotation of gene and transcript structure and function and going forward, it is by utilizing multiple orthogonal datasets in combination that we will be able to shrink the gaps in annotation.

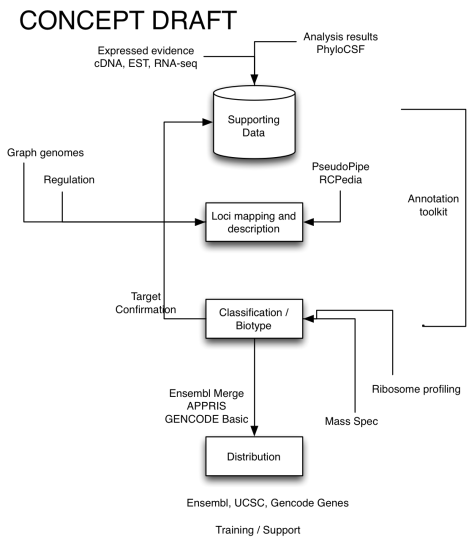


Figure 2: A schematic of the core GENCODE data and analysis flow.

Innovation

Getting all of the details right in genome annotation requires integration of a diverse set of evidence data and the application of clear and consistent processes (Figure 2). This effort is partly manual and time consuming (cite Harrow), but to date no computational or automatic approach is able replace the tasks of an experienced and trained annotator. Within GENCODE our clear experience is that consistent procedures lead to the best possible results and that when it comes to creating highly accurate genome annotation, sometimes conceptually unexciting, well established methods and processes are necessary for a comprehensive solution. 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.

GENCODE has developed since its founding as a reasoned combination of well-established and conservative procedures with targeted investigations ("pilot projects") into the value of new technologies, new data and new sources of evidence. These pilots are a major source of innovation in the project and over the course of this proposal 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 almost ten 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). This enduring collaboration has **four fundamental components: (1) a comprehensive gene annotation pipeline; (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 that 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.

Over the last four years, GENCODE has completed full first pass manual annotation of the human genome, conducted extensive QC on the annotation and investigated many novel data types and data sets. Going forward, the annotation of human genome sequence will follow a similar path of testing new data types and extension of existing data types into new cell-lines, tissues, and developmental stages generated within the GENCODE consortium, by other collaborators and deposited in the public repositories. GENCODE will develop annotation strategies to utilise 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 formalise our guidelines for edge-case resolution while the large volumes of new data with direct relevance to gene annotation will require our continued development of providing direct computational assistance to manual annotation downstream of the alignment and prediction steps.

The GENCODE annotation of the mouse reference genome is less complete than that of the human reference genome. As such part of the continuing annotation effort will be to continue 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. 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; human has much more experimental data, but mouse has access to tissues and developmental datasets that are unavailable to human researchers.

With this application we will continue curating the GENCODE resource for human and mouse, deepening the annotation and its utility to include tissue-specific isoforms and expression. As new data become available, existing annotation will be refined. For example, more accurate transcription start and end positions will be identified using CAGE and 3prime pull down data.

We will also expand the GENCODE resource is deliberate and unique ways in response to community feedback and opportunities presented by new technology. Our targeted areas for these future GENCODE expansions are (1) comprehensive annotation via identification and characterization of

transcripts that are may not be present on the current reference genome or that are polymorphic pseudogenes; (2) leverage graph-based representations of genome structure for incorporating and distributing population or individual specific annotation and (3) identification of genome regulatory regions that are confidently connected to specific genes and transcripts. To these ends, we will use the growing collection of human and mouse genome sequences, transcriptional resources and functional data within the current GENCODE pipelines and in line with GENCODE's overall goal to annotate all evidence-based gene features in the human and mouse genomes with high accuracy.

Comprehensive gene annotation pipeline

The manual gene annotation process, by which experienced human annotators assess and curate gene structures and interpret their functional potential based on experimental evidence and predictions processed through a diverse set of computational pipelines, remains central to the GENCODE project. Manual annotation of protein-coding, long non-coding RNA and pseudogene loci for the GENCODE human and mouse genesets is carried out according to the guidelines of the HAVANA (Human And Vertebrate Analysis and Annotation) group; available at https://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf).

Historically the HAVANA group has produced transcript models largely based on the alignment of transcriptomic (ESTs and mRNAs) and protein sequence data from GenBank and Uniprot. Although new data types have been introduced into the annotation process, the vast majority of transcript models in GENCODE were created using support from these sources of evidence. Data were aligned to the individual BAC clones that make up the reference genome sequence using BLAST (ref), with a subsequent realignment of transcriptomic data by Est2Genome (ref). Transcript and protein data, along with other data useful in its interpretation were viewed in the ZMap annotation interface (http://www.sanger.ac.uk/science/tools/zmap).

Integrated approach to pseudogene identification and classification

The Yale group has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and Sanger Havana group, we have developed a variety of methods to identify pseudogenes \cite{16574694,16925835,22951037}. These including Pseudopipe, which takes as input all known protein sequences in the genome and using an homology search is able to identify disabled copies of functional paralogs (referred to as pseudogene parents). Based on their formation mechanism pseudogenes are classified into 3 different biotypes: processed, unprocessed and ambiguous. A second and newer method, RCPedia, focuses on the annotation of retrotransposed (processed) pseudogenes \cite{23457042}. These pipelines and extended versions of them will be used to finish the annotation of mouse genome including annotation of both the mouse reference and the recently available 18 mouse strain assemblies. The pseudogene collection will be characterized by 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 psuedogene variability across human individuals and, in doing so, help to understand the boundary between protein coding genes and psuedogenes.

Computational methods to evaluate and enhance gene annotation

The Ensembl GeneBuild provides an automated, independent method to identify and annotate all genes including protein coding, non-protein coding, and pseudogenes. Our gene annotation has a reputation for high quality, as judged by community assessments, which is achieved by a well-established core data flow that integrates alignments of expressed protein, cDNA and other biological sequences (Aken et al, 2016). The primary data used to inform gene annotation are: protein sequences from UniProt, full-length mRNA and transcriptome sequences from ENA, and RNAcentral resources for noncoding RNA genes. The Ensembl GeneBuild is merged with the Havana manual annotation to create the full GENCODE geneset and is especially valuable for filling in transcripts that may be expressed in difficult to assess human tissues and for regions of the mouse genome that have not yet been had comprehensive manual annotation.

The Ensembl RNA-seq pipeline (Collins et al 2012) 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 provide tissue-specific expression information.

Additional specific 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 additional information that is added to the transcripts of the final GENCODE set. Specifically, we use the current version of phyloCSF (MIT) to help identify the 1000s or tens of 1000s of novel protein coding exons that are unannotated within existing 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) and UCSC's Transcript Support Level (TSL) methods provide valuable and complementary information about the quality and consistency of the final GENCODE set will continue to be developed.

Experimental validation

Complementary experimental approaches will be used to verify and validate various annotations within the GENCODE project. Specifically, we will use our recently-developed methodology for the targeted annotation of known and novel RNA transcripts by PacBio third-generation sequencing - "Capture Long-Seq" (CLS). This approach enables us to focus on a candidate genomic space for new transcript discovery, whilst providing complete or almost-complete transcript models for each. CLS will be deployed for a series of complex tissues in both adult and embryonic timepoints. In addition, we will use mass spectroscopy for evidence-based annotation of biotype, to validate novel protein coding genes and transcribed pseudogenes as well as to identify alternative isoforms and non-sense mediated targets.

Research Strategy (Resource Project): 12 pages

The central focus of the project should be the generation of a research resource that is broadly useful. This component should describe the resource in more detail than the Overview Component, how it will be produced, and how it will be integrated or coordinated with related resources.

Complex applications for large awards should include all the elements below. Applications that are less complex will likely require fewer pages and may skip elements below that are not relevant.

The application should include preliminary data that support the technological approach, if appropriate.

Renewal applications should include a progress report.

For renewal/revision applications, provide a Progress Report. Provide the beginning and ending dates for the period covered since the last competitive review. Summarize the specific aims of the previous project period and the importance of the findings, and emphasize the progress made toward their achievement. Explain any significant changes to the specific aims and any new directions including changes to the specific aims and any new directions including changes resulting from significant budget reductions. For any studies meeting the NIH definition for clinical research, discuss previous participant enrollment (e.g., recruitment, retention, inclusion of women, minorities, children etc.) as part of the progress report, particularly if relevant to studies proposed in the renewal or revision application. You should not submit a PHS Inclusion Enrollment Report form unless the enrollment is part of new or ongoing studies in the renewal or revision application.

Other guidance may apply to some, but not all, types of resource applications:

Training about the use of the resource is a common feature of NHGRI community resource awards. For the types of resources where this would be useful, such as informatics tools and data resources, the application should include information on how training in use of the resource will be provided.

If appropriate, the application may have an applied research component to improve the methods used to develop the resource. (Note that hypothesis-driven R21- or R01-like research is not considered applied research). This research component may comprise up to 10 percent of the direct costs of the award.

The resources that NHGRI will support must represent work in genomics that is broadly applicable to many diseases and research questions. If appropriate, applications may include plans for obtaining additional support and co-funding for the resource.

Special requirements for informatics community resource projects

These projects include human or model organism databases and other informatics resources that involve curation of data from the literature and integration with other genomic or genetic data. Applications for these resources should also address the following issues:

1. The data types to be included in the resource: The application should present a rationale for including or excluding particular data types (incorporating community priorities), and should provide a plan for adding new data types as they arise. In addition to the main data types that are the focus of the resource, the resource should also include types of evidence, measures of data quality, descriptions of curation methods and associated metadata, and attribution of data sources, for both experimental and computational data.

2. The curation processes to be used: These should be described and justified. Processes addressed should include high-quality manual and computational methods as well as extraction of information from the literature. The application should describe the plans to present the curated data and descriptions of the curation process to users. The application should outline the controlled vocabularies that will be used to describe the data. The application should list the amounts of the various data types to be curated.

3. Any plans to leverage and integrate data from other genomics data resources: The application should explain which resources were chosen and why the data are appropriate for inclusion in the proposed resource. If data from existing resources that provide similar or overlapping information are included, the application should justify their value and how unnecessary duplication will be avoided. Applications should discuss how the resource will clearly attribute data to other resources.

4. Any plans to coordinate with related data resources: This may include playing an active role in securing agreement on controlled vocabularies and common data exchange formats where necessary. Applicants should discuss their track record in coordinating with other resources.

The resource sharing plans should be provided only in the Overall Component.

Appendix: Do not use the Appendix to circumvent page limits. Follow all instructions for the Appendix as described in the SF424 (R&R) Application Guide.

Progress report (~2 pages)

Progress report to be divided into the four components listed above.

SANGER PORTION TO BE COMPLETED

Pseudogene annotation. Our experience with annotating pseudogenes spans more than 15 years. Over time we have annotated and reviewed pseudogenes in a variety of species including prokaryotic organisms \cite{15345048,14583187}, yeast \cite{11866506,12417195}, plants \cite{12083509}, worm \cite{11160906}, fly \cite{12034841,12560500}, and a wide range of vertebrates (e.g. zebrafish, mouse, rat, chimp, and human) \cite{19835609,12052146,12417195,12909341,18065488}. Our involvement in the GENCODE project started over a decade ago and ever since we have generated the complete and comprehensive set of pseudogenes in human and model organisms. Moreover, we elucidated the evolution and activity of the pseudogenes by using variation and functional genomics information.

In detail, in our most recent and updated publication, we identified 14,505 pseudogenes in human, 911 in worm, and 145 in fly \cite{25157146,22951037}. The numbers of pseudogenes are not proportional to the genome sizes or the numbers of coding genes in the genomes, highlighting the species specific evolution of pseudogenes. This specificity is also reflected in pseudogene types, where processed pseudogenes dominate over duplicated ones in human more than in the other species. This indicates a burst of retrotransposition events at the dawn of primate lineage \cite{25157146}. We also conducted systematic analyses of human pseudogenes focusing on large pseudogene families \cite{19123937,12417195,19835609} or particular types of pseudogenes such as unitary \cite{20210993} and polymorphic pseudogenes \cite{21205862}. 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.

Despite the presence of disabling mutations such as premature stop codons or loss of promoters, numerous studies have shown that pseudogenes can be transcribed and even translated \cite{15860774,16680195,15876366,17568002,16683022}. Using the RNA-seq data from Human BodyMap, we investigated the expression pattern of pseudogenes across 16 human tissues. 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. While testis holds the largest number of transcribed pseudogenes, skeletal muscle holds the least \cite{22951037}.

Data types to be included in the resource (~1 page)

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 goal standard. We have been collaborating for almost ten years, and have expertise in: gene and transcript isoform identification, pseudogene evolution, sequence conservation, gene expressions, proteomics and post-translational modification, and gene regulatory elements.

Huge strides have been made by the consortium, including the first-pass annotation of the human genome. However, there is much more to be done. Many of the annotated human genes are necessarily incomplete because the sequence data available at the time of annotation was incomplete. In particular, thousands of transcript isoforms are known to be 5-prime or 3-prime incomplete. New

transcriptome data 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 and there is ongoing research in this area that will be important for GENCODE to incorporate.

Beyond the coding and non-coding genes, GENCODE creates reference pseudogene annotation. Pseudogenes are defined as disabled copies of functional genes. Depending on their formation mechanism they 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. Pseudogenes provide valuable opportunities to study the dynamics and evolution of gene functions.

Pseudogenes have long been considered nonfunctional elements. However, recent studies indicate that pseudogenes can be transcribed, translated and can play key regulatory roles. In particular pseudogenes can regulate the expression of functional protein coding genes by serving as a source of siRNAs, antisense transcripts, microRNA binding sites, or competing mRNAs \cite{22726445,21080588,22990117}. 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 \cite{24026178,22344438,21205862}. Finally, the annotation of pseudogenes is important in the analysis of personal genomes, providing 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. 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. As this reference genome expands, so GENCODE will provide annotation appropriate to these new genomic sequences.

There are similar challenges for mouse, not least that the mouse has not yet achieved the gold standard of a first-pass manual annotation. On the other hand, the mouse reference genome is ahead of the human genome in that the GRC have already replaced the linear genome with a ‘graph’ of 16 mouse strains. Annotating reference gene resources for both mouse and human has many advantages, allowing for scalability as well as early access to pilot new data types in one species that are not yet available for the other.

Curation processes to be used (~6.5 pages)

Comprehensive gene annotation pipeline (~2.5 pages)

The manual gene annotation process, by which experienced human annotators assess and curate gene structures and interpret their functional potential based on experimental evidence and predictions processed through a diverse set of computational pipelines, remains central to the GENCODE project. Manual annotation of protein-coding, long non-coding RNA and pseudogene loci for the GENCODE human and mouse genesets is carried out according to the guidelines of the HAVANA (Human And Vertebrate Analysis and Annotation) group; available at https://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf).

Historically the HAVANA group has produced transcript models largely based on the alignment of transcriptomic (ESTs and mRNAs) and protein sequence data from GenBank and Uniprot. Although new data types have been introduced into the annotation process, the vast majority of transcript models in GENCODE were created using support from these sources of evidence. Data were aligned to the individual BAC clones that make up the reference genome sequence using BLAST (ref), with a

subsequent realignment of transcriptomic data by Est2Genome (ref). Transcript and protein data, along with other data useful in its interpretation were viewed in the ZMap annotation interface (http://www.sanger.ac.uk/science/tools/zmap).

The core GENCODE process (Figure 2) starts with a diverse set of data types that have been either aligned to the reference genome assembly or calculated via one of the several comprehensive annotation pipelines. For example, 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 (Ref) and Retrofinder (Ref) pipelines produced by GENCODE consortium members), cross-species conservation, transcription start and termination sites, regulatory features, Mass Spectrometry and Riboseq data, 2nd (i.e. illumina) and 3rd (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.

Updates to Annotation Our initial usage of short RNA-seq transcriptomic data was limited by its short read length and the quality of transcript models derived from it, which was lower than could be produced using our conventional transcriptomic datasets(Ref). As such, its primary usage was to confirm individual splice features (introns, exons, splice sites) that were constructed using conventional transcriptomic data but were weakly supported for example, uncertain alignments of poor quality EST sequences, non-species evidence, or evidence from a paralogous locus in the same species. Another significant application was using RNA-seq coverage data to identify extensions of 3' UTRs to distal transcription termination sites and polyA features where conventional evidence was patchy but indication connection.

However, more recently was have been using RNA-seq transcriptomic data to support annotation of novel transcripts/loci, almost exclusively lncRNAs. Our pipeline takes models created by multiple analysis pipelines from Ensembl, ENCODE collaborators and lncRNA collaborators (later adding good quality public experiments such as PLAR (REF)). As a consequence the transcript models used are generated using independent methods even where there is an overlap between the RNA-seq datasets they used. Initially the RNA-seq based transcripts were used to identify loci where ≥1 intron from a transcript intersects with an unannotated intron in splicing ESTs and cDNAs (and later 3rd generation RNA-seq). This information was used to create a list of candidate loci for targeted manual annotation. Where there is no overlapping EST or cDNA evidence a separate list of loci where introns from ≥2 independent RNA-seq based transcripts overlap was created to facilitate targeted manual annotation. Going forward we will create a GENCODE transcript based on the most representative RNA-seq-based transcript (i.e. the one that contains the most introns supported in all overlapping RNA-seq-based transcripts). Where available CAGE and polyAseq data are available, they are used to trim the ends of the RNA-seq based transcript. These models are tagged to ensure they do not enter the GENCODE geneset unless they are check by an annotator, modified as necessary to achieve with the best possible annotation at the locus and the tag removed. In this way we let the computational pipeline take the load of checking simple and precise intersections of features like introns and add all relevant attributes and metadata which frees the annotator to spend more time assessing the functional biotype of the locus e.g. whether is it an unannotated protein-coding gene or pseudogene and correcting or improving models where this is appropriate.

We have also implemented a similar pipeline using 3rd generation transcriptomic dataset using CaptureSeq PacBio and SLRSeq dataset. We initially cluster mapped reads within an experiment, then between experiments, then between data types to create a single track. Novel loci, i.e. those that share no overlap with annotated transcripts are targeted for investigation by annotators. Again, we used transcripts derived from clustering to create a GENCODE transcripts tagged to ensure they do not enter the GENCODE geneset without specific authorisation from an annotator. Again the transcripts

may be edited to achieve with the best possible annotation at the locus and the tag removed to allow them to be included in the GENCODE geneset.

On the human genome, annotation is driven by our generation of hierarchical lists of features of interest ie lists of all putative features of interest eg unannotated protein-coding genes, unannotated pseudogenes, unannotated lncRNAs, is generated which are then filtered and prioritised on the basis of their likelihood of representing true positive features ie to provide high specificity. Where initial filtering or prioritisation yields unsatisfactory results, alternative filters/prioritisation is applied to improve the specificity.

On the mouse genome, annotation is prioritised in two ways – while a substantial annotation effort is driven by a similar approach to that in human, we also continue to annotate clone-by-clone, identifying all genic features within the region under investigation. First pass human annotation was completed using a very similar approach giving consistency between the two annotation sets. More recently in mouse we have added an annotation priority track to flag features annotators are required to investigate. This track summarises information in other tracks eg Pseudopipe and Retrofinder pseudogene predictions, protein-coding identified during the Ensembl genebuild but loci lacking manual annotation, putative lncRNAs identified by the RNA-seq and 3rd generation transcriptomic pipelines.

Determining biotype GENCODE genes are assigned a biotype associated with one of three broad categories; protein-coding gene, lncRNA gene or pseudogene. Genes derive their biotype from the biotypes assigned to their constituent transcripts. Not all transcripts within a locus are required to have the same biotype, but all transcripts must have biotypes compatible with their gene biotype. For example, a protein-coding locus must have at least one transcript with the protein-coding biotype but it may have others with the nonsense-mediated decay (NMD) or non-stop decay (NSD) biotypes; it is never permitted to contain a transcript with a biotype belonging to another of the broad biotype categories such as lncRNA or pseudogene.

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 possibility of coding potential is eliminated. Protein-coding potential of transcripts is determined on the basis of similarity to known protein sequences, the sequences or orthologous and paralogous proteins, the presence of Pfam functional domains (ref), clear support of high quality peptides from Mass Spectrometry (MS) experiments and good evidence of translation from Ribosome Profiling (RP/RiboSeq) data. Deep conservation of a CDS can be used to support annotation of a protein-coding transcript, even in the absence of any other support. Once a locus is established as protein-coding the likely susceptibility of AS transcripts to Nonsense-mediated decay (NMD) may be considered. A transcript is annotated as a target for NMD if it fulfils the following criteria; it has a stop codon >50 bases upstream of a splice junction and has either a full-length CDS (start-codon to stop-codon) or, where the start-codon is missing, a CDS that shares its reading-frame with a protein-coding transcript at the same locus.

Where a locus shares homology with UniProt accessions but is deemed unlikely to encode a protein itself because its putative CDS is disrupted by in-frame stop codons, frameshifts, insertions or deletions it is assigned the pseudogene biotype. The pseudogene broad biotype has multiple subdivisions that describe the mechanism of creation and transcriptional status of the locus. Pseudogenes are generated via retrotransposition events while 'unprocessed' pseudogenes are created by duplication. Where ambiguities arise other features are used to resolve the provenance of the pseudogene such as the flanking sequence and genomic location relative to parent loci. Unitary_pseudogenes are identified by the presence of an unambiguous functional ortholog in another species and are checked to confirm both that any disabling mutations are fixed in humans and that they are not due to errors in the reference genome sequence. Where a disabling mutation is not fixed i.e. it is polymorphic or segregating, the locus is annotated as a polymorphic pseudogene i.e. it is a protein-coding gene that happens to possess a loss-of-function variant in the reference genome. Processed, unprocessed and unitary pseudogenes may be annotated as transcribed and translated

where they have locus-specific evidence of transcription or multiple peptide spectrum matches from high quality proteomics studies (eg Brosch et al 2011 and Ezkurdia et al 2012) respectively.

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; similarity to known protein sequences, the sequences of orthologous and paralogous proteins, the presence of Pfam functional domains (ref), clear support of high quality peptides from Mass Spectrometry (MS) experiments and good evidence of translation from Ribosome Profiling (RP/RiboSeq) data. Given our current inability to infer the functional potential or mode of action for all but a handful of lncRNA loci, biotypes are assigned on the basis of genomic position relative to protein-coding loci. For example, 3prime_overlapping_ncRNA transcripts have strong evidence for independent TSS overlapping the 3'UTR of a protein-coding locus on the same strand, antisense transcripts overlap the genomic span of a protein-coding locus on the opposite strand, lincRNA transcripts are intergenic to protein coding loci and bidirectional lncRNA have a TSS within the promoter region of a protein-coding gene on the other strand. The lncRNA annotation produced by GENCODE represent a core dataset underpinning the RNA Central lncRNA dataset (Ref).

We have also included a number of additional datasets that provide specific information required to produce accurate annotation of transcript functional biotype.

CAGE data produced by the ENCODE(REF) and Fantom(REF) Consortia and RAMPAGE data also produced by ENCODE define the position of the transcription start site of a transcript. In protein-coding loci this data is essential in identifying whether annotated AS transcripts are full-length which has implications over the assignment of a CDS. Simply put if we know the TSS has been identified, a translation initiation site can be added with confidence. For lncRNA loci, knowledge of the TSS gives confidence that the locus has been annotated to its full extent and the biotype assigned is correct.

polyAseq data (REF) defines the polyA site i.e. the end of a transcript. As with CAGE data knowledge of the position of the end of the transcript end is useful in defining biotype, particularly for transcripts with a putative premature termination codon, where a decision is required to annotate with the NMD or coding biotype.

Ribosome profiling data (Riboseq) (REF) define the position of ribosomes bound to mRNA. Depending on the preparation technique they indicate either the position of the translation initiation site (TiS) or the CDS. Information from RP data is useful, though not definitive in interpreting whether a putative CDS is translated and which potential TiS is most actively used. As such it is useful in annotating AS in protein-coding loci and determining whether a locus should be assigned a protein-coding or lncRNA biotype.

Mass spectrometry (MS) peptides represent the sequences of detected proteins and, despite potential problems with miscalling and mismapping of sequences, they can provide definitive support for the translation of a CDS. As with RP data they are useful in confirming the translation of AS transcripts in protein coding loci and, in conjunction with orthogonal data to determine whether a locus should be given a protein-coding, pseudogene or lncRNA biotype.

PhyloCSF identifies potential novel coding sequence based on evolutionary signatures. Generally PhyloCSF is run against transcripts based on RNA-seq data in order to identify putative protein-coding genes. In order to identify conserved but unannotated protein-coding loci, even where they lack strong evidence of transcription, we have run PhyloCSF against the whole human and mouse genomes producing more than 600,000 regions with putative protein-coding potential in each. Following extensive filtering to remove lower quality regions and regions intersecting with annotated CDS the number of target regions was reduced to ~60,000 in each genome. We adopted a number of additional filtering strategies to specifically identify unannotated protein-coding loci (as opposed to novel protein-coding exons), looking for high scoring regions that were found in clusters, overlapping annotated lncRNA loci, in intergenic regions, overlapping ab initio gene predictions, overlapping unannotated loci identified using RNA-seq data and 3rd gen transciptomic data. As a result of this analysis we identified

~100 novel protein-coding loci and ~200 novel pseudogenes (which still carry the signature of protein-coding conservation) in addition to ~100 novel protein-coding exons in annotated coding loci. These were generally 5' and 3' extensions to genes as we have excluded intronic PhyloCSF regions from the analysis thus far.

Status and Attributes To provide more detail to the description of transcripts and loci, a status of known, novel or putative is assigned based on the presence in other sequence databases e.g. UniProt and RefSeq and the nature of the evidence that supports the annotation of the transcript. We use controlled vocabulary terms or attributes to describe important features of transcript and gene annotation that are not captured in other fields. For example a transcript build of the basis of support from transcriptional evidence not derived from the same organism eg using a mouse cDNA to support a transcript in human 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 geneset because it is supported by cross-species conservation is tagged with the attribute 'non-canonical conserved'. All attributes may be queried by users to facilitate the filtering of their associated transcripts and loci.

Missing loci It is clearly one of the key objectives of GENCODE to represent all gene loci. The means by which we identify missing loci is determined by their functional and transcriptional properties. For example, the identification of unannotated protein-coding genes will be different to finding unannotated pseudogenes and lncRNAs. The main data types we will use to identifying unannotated loci are transcriptomic data, and proteomics data. Where a locus is transcribed, i.e. protein-coding genes and lncRNAs but also transcribed pseudogenes, it can clearly be identified on the basis of its transcription. We will continue to investigate transcription identified in RNA-seq datasets from previously inaccessible tissues and development stages, public 3rd generation transcriptional evidence such as SLRseq and PacBio and on-target and off-target Captureseq PacBio reads generated within the GENCODE consortium. Transcript models produced from RNA-seq data of present problems to annotators because of their short length, but integrating them with longer reads and CAGE and polyAseq data will allow better resolution of loci.

Recent analysis of reprocessed MS data from large-scale public shotgun proteomics datasets in human has led to the identification of ~20 novel protein-coding loci (REF). We will continue to use large public proteomics datasets to identify putative novel protein-coding loci. While the intended targets of our pipeline were novel protein-coding genes, we determined that many of the loci identified were likely to be pseudogenes, flagged as encoding proteins in error. However, false positive identification of novel protein-coding loci did lead to the identification of unannotated pseudogene loci. Mouse shotgun proteomic datasets are significantly smaller than human and as such less likely to reveal novel protein-coding loci, however, the future availability of samples in tissues and developmental stages inaccessible in human may require the exercise being repeated in mouse.

We have previously used cross-species conservation data to identify novel gene loci. Using both a measure general conservation such as PhastCons (REF) and specific conservation of protein-coding sequence PhyloCSF (REF) to flag unannotated regions of high conservation. For example detailed investigation of a refined set of thousands of high scoring PhyloCSF regions generated across of the whole human genome yielded more than 100 novel protein-coding, many of which had very low expression support in human RNA-seq datasets. However, lower scoring phyloCSF regions led to the identification of novel protein-coding gene loci with an increasingly reduced efficiency and as such it seems likely that we are close to identifying all human protein-coding loci with a strong signature of conservation. The same exercise has begun, but not been completed in mouse and as such it is very likely that unannotated protein-coding loci will be captured within this set. Although many of the novel loci identified in human and mouse on the basis of their conservation have orthologs in both species, we have identified multiple loci where a gene has been lost in one lineage, emphasising the importance of independent analysis in both species. As with MS data, PhyloCSF frequently highlights unannotated pseudogene loci. The identification of novel unitary pseudogenes reflects correct genome alignment and highlights the sensitivity of PhyloCSF in identifying residual protein-coding signature in non-functional genes while the identification of unprocessed pseudogenes lacking orthologs is an

unintended benefit resulting from artefacts in the genome alignments underpinning PhyloCSF. The same analysis that led to the identification of ~100 novel protein coding genes also found more than 200 novel pseudogenes and continues to do so with acceptable efficiency suggesting that there remain many unannotated pseudogenes in both human and mouse genomes

As novel protein-coding loci are added to the GENCODE geneset it is essential to revisit the possible presence of unannotated paralogs and pseudogenes, particularly for loci that lack any previously annotated a paralogs. This will be achieved by iterative reanalysis of the genome using Pseudopipe (Yale). We will continue to closely monitor both the research literature and other reference sequence databases that include curated content, for example UniProt and RefSeq (and MGI for mouse) to test any genes that are unannotated in GENCODE and determine whether there is sufficient support for their annotated according to GENCODE guidelines

Regardless of the data that led to their identification, all novel loci will be assessed to ascertain their functional biotype. For example, pseudogenes can be highlighted by discovery pipelines for protein-coding gene and vice versa. LncRNAs are occasionally found by protein-coding pipelines, however, protein-coding genes and transcribed pseudogene loci may be found by pipelines agnostic to the presence of a CDS or specifically intended to target lncRNAs. All novel loci will be investigated to determine their protein-coding potential based on homology to known protein sequences, the presence of Pfam functional domains, clear support of high quality peptides from Mass Spectrometry (MS) experiments, good evidence of translation from Ribosome Profiling (RP/RiboSeq) data and conservation of the CDS.

Partial annotation and underannotation Adding annotation of missing AS transcripts at annotated loci and extending partial AS transcripts to reflect their full length remains an important goal of the GENCODE projects. Clearly adding missing features such as exons, and splice junctions is essential in providing the best possible foundation for downstream analysis using the GENCODE geneset, for example in the interpretation of variation data where, similar to an unannotated locus, unannotated genic sequence will lead to a difference in interpretation and functional significance of the variant (REF). Even where all exonic sequence is annotated, the accurate extension of all transcripts to full-length is essential to describe the connectivity between these features, which impacts the functional annotation of the transcript and the predicted effect of any variation affecting it. For example, a variant might introduce premature stop codon into the sequence of a transcript, without knowledge of the position of the nonsense codon within the transcript it would be possible to misestimate its severity. Similarly, knowledge of all the AS transcripts at a locus is important for example to identify transcripts that do not include the exon effected by the variant (REF). To add further value to the annotation it is important to extend all transcripts to full length as frequently a proper interpretation of the functional potential of a transcript is not possible where a transcript remains truncated. For example, where an EST supported AS transcript shows a frameshift, it is likely that the CDS will end at a premature termination codon (PTC) that will trigger the NMD pathway. However, unless the PTC and downstream splice junction are identified we cannot be certain of this, as other splicing possibilities could lead to a full-length CDS and preventing recruitment of the NMD pathway. Currently such transcripts are annotated with an agnostic processed transcript biotype; having the full-length transcript in such an instance would make the resolution of the correct biotype straightforward. It is also necessary to extend transcripts of all biotypes to full-length. For protein-coding genes this would require all transcripts with biotypes NMD and retained_intron be extended. Both these biotypes had historically been regarded as reflecting, or at least enriched for missplicing events and poor RNA preparation. However, increasingly over recent years, transcripts of both these biotypes have been implicated in the post-transcriptional regulation of the genes with which they are associated (REFs) and disruption of their splicing has been associated with disease (REF).

While almost all protein-coding loci have at least one full-length transcript, many annotated lncRNA loci do not appear to have even that (REF). In additional to simply better representing lncRNA loci and capturing the genome sequence associated with exons of lncRNA transcripts, extending transcripts at lncRNA loci to TSS and TTS aids their annotation in two significant ways. Firstly, having full-length

transcripts will enable us to gather together associated but currently non-overlapping partial transcripts to provide a more accurate estimate of the number of lncRNA loci and secondly knowing the full-length of a locus will allow a more informed determination of transcript and locus biotype, for example confirming that currently annotated lincRNAs are fully (ie TSS to TTS) intergenic, which will allow properties of the position based biotypes we currently use to be determined.

In GENCODE v24 there are more than 10,000 processed transcripts annotated at protein-coding genes and a further 33,000 partial protein-coding transcripts tagged as either start or end not found. Those transcripts associated with protein-coding loci but lacking a CDS are considered underannotated, in that their structures are correctly described to the extent of their homology with the supporting evidence but due to uncertainty over any splicing events not covered by the supporting evidence and the position of their termini. It is difficult to estimate the precise number of unannotated AS transcripts even within protein-coding loci, however, it is likely to be large. A recent effort by HAVANA to reannotate 70 (seventy) genes on a clinical panel for Early infantile epileptic encephalopathies (EIEE) using PacBio, SLRseq (REF) and RNA-seq (REF) datasets from brain (i.e. the appropriate tissue given our knowledge of the expression pattern of the genes and the disorders in which they are implicated) 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 represented novel CDS.

Missing and partial AS transcripts will primarily be detected and extended using 3rd generation transcriptomic data, SLRseq and particularly PacBio CaptureSeq based on the transcripts we have already annotated (Rory/Roderic). However, other data types are essential to ensure that coverage of unannoted exonic sequence is complete and accurate functional annotation of new transcripts possible. While the high scoring regions idenitified during genome wide PhyloCSF analysis may not reveal a great number of missing protein-coding loci, there are many thousands of intronic, 5' and 3' proximal regions that are likely to represent a large number of conserved unannotated coding exons. Furthermore, in the absence of the preferable 3rd generation transcriptomic data, 2nd generation RNA-seq, with it 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. Full-length transcripts combined with knowledge of TSS and TTS give all the required information to make a determination of biotype, specifically certainty over the TSS allows the translation initiation site to be determined and identification of the TTS provides important context for the position of the stop codon and whether any PTC would be likely to trigger NMD.

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 GENCODE Basic set, which contains only full-length transcripts, will inevitably become redundant as the numbers of transcripts it contains increases and the difference between the Basic and Comprehensive sets shrinks. In order to make the GENCODE geneset filterable for users we will annotate transcripts [could also annotate individual exons where the data is likely to 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 (quantity of transcript) and rate of inclusion in transcripts of splicing features by tissue/cell type/developmental stage, RP, targeted and shotgun MS datasets will be used to confirm translation, score of individual components of the APPRIS pipeline currently used to indicate principal isoform, cross-species sequence conservation and variation 'load' or tolerance relative to rest of gene [cf ExAC] can be used to identify transcripts that include regions under selection/constraint.

Annotation of novel features While our current role in the annotation of protein-coding genes is to ascertain and capture the best possible transcript structures and CDS, extending annotation to full-length provides the opportunity to annotate additional features at the level of the transcript. Particularly relevant are features such as upstream open reading frames (uORFs or upORFs) and RNA secondary structures that regulate the passage of the ribsome along the mRNA (Refs) and consequently play a

role in controlling translation initiation (Refs). Variation affecting these features can have adverse effects on translational regulation (Ref) and as such production of high quality annotation for them is in line with the goals of the project. These functionally significant features are not represented in any other geneset and the quality of their annotation is directly affected by the annotation of other features in GENCODE eg TSS and TiS. As such annotating them in conjunction with such features, as GENCODE are able to do, adds considerable benefit over e.g. third party annotation that lacks understanding of the rest of the geneset. Integration of uORF and secondary structures annotation with other datasets such as CAGE and RNA-seq data will allow capture of consequences of alternative TSS useage on the regulatory repertoire of genes.

Integrated approach to pseudogene identification and classification (~1.5 pages)[[FN shrink + harmonize, ask about the tense, mostly cut LOF]]

The Yale group has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and Sanger Havana group, we have developed a variety of methods to identify pseudogenes \cite{16574694,16925835,22951037}.

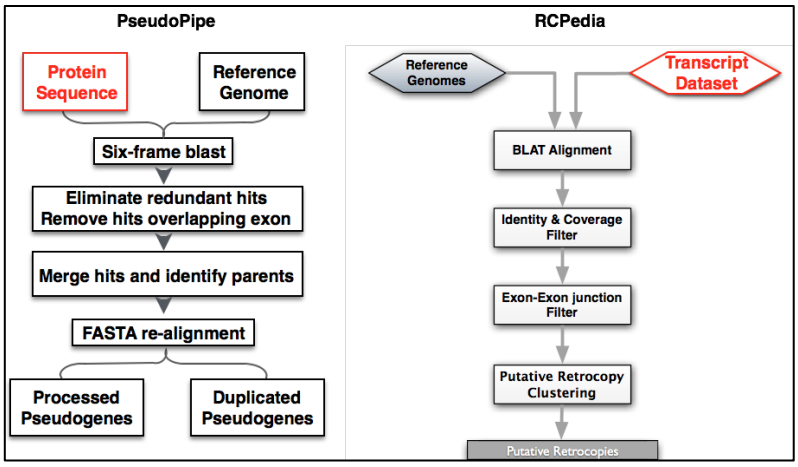


Figure 3: Automatic pseudogene annotation pipelines.

Pseudopipe, Yale's in house automatic annotation pipeline, is fast and accurate \cite{22951037} (**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). Based on their formation mechanism pseudogenes are classified into 3 different biotypes: processed, unprocessed and ambiguous. There is a good consensus overlap between the human pseudogene prediction set obtained with Pseudopipe and the set manually curated by the GENCODE annotators \cite{22951037}. Even more, the Pseudopipe predictions fueled the manual curation of pseudogenes in GENCODE \cite{22951037}.

RCPedia, the newest pseudogene annotation pipeline focuses on the annotation of retrotransposed (processed) pseudogenes \cite{23457042} (**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 \cite{18842134}. Candidate retrocopies overlapping by more than 50% with repeats identified by RepeatMasker \cite{16093699,Smit} and Tandem Repeat Finder \cite{9862982} are removed. Retrocopies are identified based on a score function using a weighted linear combination of features indicative of retrotransposition. [[Maybe cut]]

We use the 3 pipelines to identify pseudogenes in human, mouse, worm, fly, and other model organisms \cite{16925835,22951037,25157146}. We identify pseudogenes with related genomic and epigenomic data and make it available in our online databases \cite{17099229,18957444,22951037,25157146}. Moreover, using data from the 1000 Genomes Project in addition to the pseudogene annotation resulting from our pipelines, we investigate the impact of

variation on the pseudogene population in the human genome. [[ref to 24026178, 26432246 & http://papers.gersteinlab.org/papers/unitary]]In particular, we also describe retrotransposition of mRNAs (creating processed pseudogenes) as a novel class of gene copy number polymorphisms that creates variability across human populations \cite{24026178}. We also evaluated the impact of SVs across 2,504 genomes on pseudogenes \cite{26432246}.

To record the structural and functional relationship between the pseudogenes within a gene family, we developed a **pseudogene ontology** \cite{20529940}. The pseudogene ontology is used in the generation of the GENCODE genomes annotation resource and is available, alongside many other tools for pseudogene analysis at the online pseudogene repository, **pseudogene.org** \cite{17099229}.

Functional characterization We integrate ENCODE functional genomics data to obtain a comprehensive map of pseudogenes activity in human and other model organisms. Using this strategy we are able to find transcription signals for some pseudogenes and describe a large range in their biochemical activity (e.g. presence of transcription factor or polymerase II binding sites in the upstream region, active chromatin, etc). We found 1441, 143, and 23 transcribed pseudogenes in human, worm, and fly, respectively. We also identified 878 transcribed pseudogenes in mouse and 31 in zebrafish. These numbers represent a fairly uniform fraction (~15%) of the total pseudogene complement in each organism reflecting the similarity across phyla observed in their transcriptomes \cite{25157146}.

To consolidate the transcription evidence of pseudogenes in model organism and human we evaluate the expression patter of parent genes and pseudogenes. Parent genes of broadly expressed pseudogenes tend to be broadly expressed as well, but the reciprocal statement is not valid. Specifically, only 5.1%, 0.69%, and 4.6% of the total number of pseudogenes are broadly expressed in human, worm, and fly, respectively. However, in general, transcribed pseudogenes show higher tissue specificity than protein-coding genes \cite{25157146}.

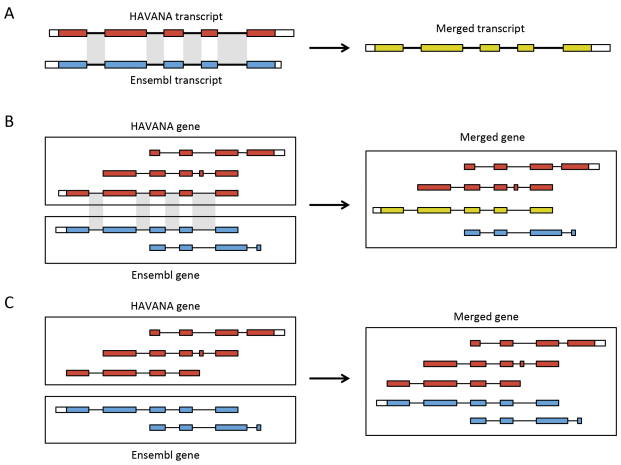
Computational methods to evaluate and enhance gene annotation (~1.5 pages)

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. The GeneBuild process automates the decision-making steps followed by manual curators, as much as they can be, using the underlying same alignment data. This automated annotation provides gene annotation for regions of the genome that have not benefited from manual curation, gene types that are not manually annotated eg. small noncoding RNA genes, and provides rapid access to novel transcript isoforms that are identified from new data in the archives.

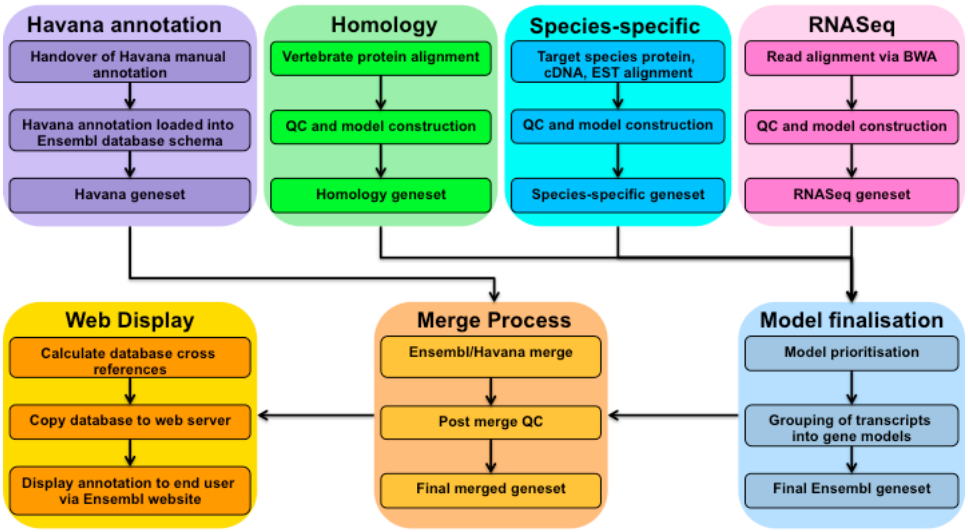
The GeneBuild is constantly improved by the lessons learned and the experience of the manual annotation team. For example, all protein-coding loci and pseudogenes that were only identified by the Ensembl genebuild pipeline were specifically checked which resulted in updating hundreds of loci to alternative biotypes. Moreover, to ensure accuracy of the annotation, regions of the genome that are biologically complex (i.e. immunoglobulins, major histocompatibility complex) or where input data are inconsistent are annotated only with manual curation.

The Ensembl annotation for human and mouse is updated when the Genome Reference Consortium release a major or minor assembly update. Major assembly updates, such as the update from GRCh37 to GRCh38 result in a change of the chromosome coordinate system, trigger a comprehensive update in which all data previously aligned to the genome are re-mapped to ensure the best-in-genome alignments. This is a considerable undertaking taking several months to complete. Minor assembly updates, such as the update from GRCh38.p7 to GRCh38.p2, simple add additional alternate sequence alongside the primary assembly. These minor updates are 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.

Ensembl Merge. The last step in the creation of the full GENCODE gene set is the merging of the Ensembl and HAVANA gene sets, which combines the Ensembl annotation with the manually curated HAVANA set to produce the 'merged' GENCODE gene set. The aim of this process is to create the most comprehensive gene set possible, by including the entire annotation from HAVANA and supplementing it with the Ensembl annotation. The Ensembl models fill the gaps where there are no HAVANA models, and they provide additional transcript isoforms using new sequence data that have not already been annotated. The merge process is run for the entire genome, including all assembly patches. The full merge process has been described in details (Harrow et al. (2012)).



We have developed a system of sanity checks against which the Ensembl and HAVANA annotation is run before the merge starts. These check for both data integrity (eg. that a gene has been annotated with one or more transcripts), and biological integrity (eg. that a 'lincRNA' gene does not have a protein coding transcript). Annotation identified by this system as inconsistent or unexpected in some way is passed back to HAVANA for further inspection.



The merge method involves comparison of all Ensembl transcripts against all overlapping HAVANA transcripts. 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 over the Ensembl biotype, except in regions of the genome that have errors and are under review by the GRC and tagged as such.

The Ensembl Healthcheck system ensures that the final GENCODE gene set meets the specified data consistency, and presence of additional gene-related data such as cross references, before public release.

GENCODE Basic Once the final GENCODE gene set has been produced, the GENCODE Basic gene set is also produced for specific purposes. This process identifies a subset of representative transcripts for each gene by prioritizing full-length protein coding transcripts over partial or non-protein coding transcripts within the same gene, thus highlight those transcripts that most useful in the majority of applications. The GENCODE Basic set is updated every time that the GENCODE genes are updated.

Add phyloCSF section here

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 (Rodriguez 2013 PMID: 23161672, Rodriguez 2015 PMID: 25990727).

APPRIS has two main roles, to catalogue and annotate the likely effects of alternative splicing on protein features, and to select a single CDS as the main (principal) isoform based on these annotations (Tress 2008 PMID: 18006548).

APPRIS is currently composed of six methods. These methods are used to annotate coding variants with structure, function, localisation and conservation information. The six methods are: SPADE, which uses PFAMSCAN (Finn 2016 PMID: 26673716) to estimate the likely effects of splicing events on protein functional domains; *firestar* (Lopez 2011 PMID: 21672959), a method predicts functionally important amino acid residues; MATADOR-3D, a method that estimates the likely effect of splicing events on protein structure; CRASH, which uses SignalP (Petersen 2011 PMID: 21959131) and TargetP (Emanuelsson 2007 PMID: 17446895) to predict the presence and absence of reliable signal sequences; THUMP, which uses three trans-membrane helix prediction methods, MemSat (Buchan 2010 PMID: 20507913), Phobius (Kall 2007 PMID: 17483518) and PRODIV (Virklund 2004 PMID:15215532), to predict the effects of splicing events on trans-membrane helices; and CORSAIR, which determines the conservation of each variant across vertebrate species.

APPRIS uses these annotations to select one CDS as the principal variant for each coding gene and to flag genes and transcripts that have unusual conservation or that code for proteins with altered structure, function or localisation. These annotations have an important quality control role and are at the core of the various CNIO quality control pipelines. Predictions for protein structural and functional features have been used to investigate interesting cases, to update gene models and to annotate new coding variants.

The APPRIS principal isoform is generally the isoform with the most conserved protein features and the most evidence of cross-species conservation. Experimental evidence clearly shows that these principal isoforms are the main protein isoforms in the cell and not just a technical distinction. APPRIS principal isoforms coincide overwhelmingly with the main protein isoform detected in proteomics experiments [Ezkurdia 2015 PMID: 25732134] and with the variant that manual annotators agree on [Farrell 2014 PMID: 24217909], and even agree with the transcript with most reliable RNA-seq evidence [in house results]. In addition principal isoforms are under significantly greater selective pressure than alternative isoforms [Liu and Lin 2015 PMID: 25820936].

The APPRIS database [Rodriguez 2013 PMID: 23161672] houses annotations for seven Ensembl [ref] species (human, mouse, rat, pig, zebra-fish, fruit-fly and *C. elegans*), and for the RefSeq [ref] human gene set. The recently developed APPRIS WebServer and WebServices [Rodriguez 2015 PMID: 25990727] allow users to check Ensembl annotations for nine other species, dog, cat, cow, opossum, chicken, zebra-finch, lizard, *xenopus* and fugu, and to interrogate the APPRIS database in an automatic fashion.

APPRIS is stable and is implemented as part of the GENCODE/Ensembl human genome annotation [ref], and can be visualized in the UCSC Genome Browser [ref] as public track hub.

The core APPRIS methods select a principal isoform for 73.4% of human genes and 82.1% of mouse genes. For those genes in which the core methods are not able to choose a main variant, we select the main isoform based on CCDS annotations and Transcript Support Level since we have found that these two methods also provide information that supports a main isoform. Using information from these two methods in addition to the core methods APPRIS is able to make a reliable prediction for the main isoform for 95.5% of human genes and 96.4% of mouse genes.

Currently 45,223 (48.2%) of the transcripts in GENCODE v24 annotated by APPRIS would generate protein isoforms with either fewer Pfam functional domains [Finn 2016 PMID: 26673716] or with damaged Pfam domains with respect to the constitutional variant for the same gene. 33,467 isoforms (35.6% of all isoforms) would have lost or damaged structural domains based on alignments with known 3D structures, and 19,458 isoforms (20.7% of all isoforms) would lose functionally important residues. In total 48,571 (51.7%) of the translated isoforms would lose either functional or structural domains, or functional residues relative to the constitutive isoform.

For mouse 18,360 of the transcripts annotated in GENCODE vM9 (32.1%) would generate isoforms with either lost or damaged Pfam functional domains, 12,989 transcripts would produce isoforms with lost or damaged 3D structure (22.6%) and 7,430 transcripts code for isoforms that have lost important functional residues (12.9%).

We believe we can use the data from APPRIS to not only select the main protein isoform for each gene, but also predict which alternative isoforms are the most likely to have important functional roles.

Coding gene analysis pipeline (CNIO) 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 [Pundir 2015 PMID: 26088053] databases. These genes may be incorrectly labeled as coding genes. Analysis of the annotations uncovered 19 different features that correlated with lack of protein-level expression. These 19 features (including poor conservation, recent origin, poor supporting evidence and annotations contradicting coding status) were used to flag 2,001 coding genes from the GENCODE v12 human reference set as potentially not coding [Ezkurdia 2012 PMID: 24939910]. The HAVANA manual annotators have revisited these genes and 1,026 have since been reclassified as either non-coding or pseudogene.

The CNIO carried out a similar analysis on new protein coding genes added between GENCODE v12 and GENCODE v19, leading to the reclassification of almost 500 automatically generated protein coding genes. The most recent analysis was with the GENCODE v23 annotation, where a further 2,050 protein coding genes were labeled as unusual. Almost half of these genes are either read-through genes, or are immunoglobulin or T-cell receptor genes.

The pipeline has also now been applied to the mouse annotation and the initial analysis identified 4,841 mouse protein-coding genes that are potentially not coding. In the case of the mouse and GENCODE v23 annotations, the HAVANA manual annotators are in the process of analyzing the likely coding potential of these unusual coding genes.

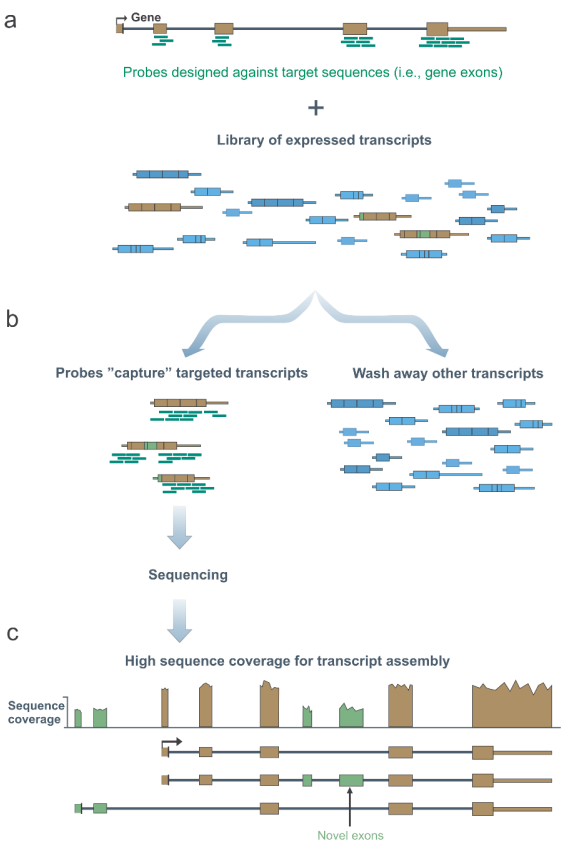
The CNIO plans to automatize the process of identification of these unusual coding genes to allow the pipeline to be run for each new release of GENCODE. This should be most useful for the mouse annotation (and other species), since the annotation of coding genes for the human reference set is close to completion.

Transcript support level (UCSC) In the current iteration of the GENCODE project UCSC creates computational quality control analysis of the generated gene sets drawing on orthogonal sources of information. This includes comparison drawn from primary data sources, such as GenBank, gene-ortholog comparisons and evolutionary assessment of gene features using conservation patterns. The result is that each CCDS release undergoes rigorous conservation, ortholog, and pseudogene evaluation. To motivate this analysis, UCSC works in conjunction with the manual annotation group to provide ad-hoc analyses and to automate and integrate previously manual approaches. In this way we have enhanced the productivity of manual annotators by providing valuable additional lines of evidence. In the proposed project we will continue to develop and produce methods for orthogonal evaluation of the GENCODE annotations using primary evidence. 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 create are provided to the community as Transcript Support Level (TSL) scores for each transcript. TSL scores 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. While RNA-Seq does not generally provide full-length transcript evidence, it provides much better consistency and provenance than mining GenBank.

Validation of Annotation Results (~1.5 pages)

High Throughput Complete Annotation of Novel Noncoding RNA Transcripts The aim of this project is to comprehensively annotate the entire un-annotated transcriptome of a series of complex human tissues in both adult and embryonic timepoints. To achieve this we will leverage our recently-developed methodology for the targeted annotation of known and novel RNA transcripts by PacBio third-generation sequencing - “Capture Long-Seq” (CLS). This approach enables us to focus on a candidate genomic space for new transcript discovery, whilst providing complete or almost-complete transcript models for each. This represents a huge advantage over both (1) manual annotation approaches (in terms of cost and throughput), and (2) previous short-read RNA capture sequencing (in terms of producing full-length transcript models).



The present study will be far more ambitious in scope over previous projects. In the latter, we mainly focused on completing existing GENCODE lncRNA gene annotations, to which we added ~250% more validated splice junctions and Y% full lengths transcripts. In other words, we increased the median length of lncRNA annotations from X bp to Y bp. In the new study, we propose to apply CLS to annotation of novel transcript structures.

Although we are well aware that the mammalian genome contains a wealth of novel genes, both protein-coding and not, it is likely that our annotations of these are highly incomplete. The latest GENCODE lncRNA annotation (v24) contains ~16,000 lncRNA genes. However, a growing number of studies, themselves likely to be incomplete, point to lncRNA genesets in the region of 50,000 - 100,000 (REF Hangauer, NONCODE, Managadze). In addition, there is intriguing evidence for many thousands of novel protein coding genomic regions that remain unannotated, based on multiple genome alignments (REF). Most of these predictions are based on de novo short read assembly and bioinformatic inference, meaning that they are likely to suffer from false positive and false negative predictions. Nevertheless, amongst the 100,000+ candidate regions are likely to reside a substantial core of genuine, unannotated genes.

The aim of the present project will be to screen a large set of candidate transcript structures, including the above mentioned sets, in addition to a series of de novo collections generated in house. The power of CLS will be leveraged to directly identify full-length transcript structures amongst these, at high-throughput and low cost.

The project will have the defined end point of completing the annotation of full-length transcript structures, both coding and non-coding, in a defined series of complex adult and fetal human tissues.

Target Selection For both human and mouse, we will obtain target annotations for human from phyloCSF regions, Stringtie models, NONCODE and orthology (Washietl, PLAR). Stringtie de novo transcript models will be generated in house on the following RNA-seq datasets. For adult tissues, multiple Gtex samples will be merged to give samples of high depth. Similar analyses will be carried out for mouse using ENCODE and other public data. Human embryo data will be sourced from ENCODE and will include cerebellum, cardiac myocyte, diencephalon, frontal cortex, H1-hESC, heart, liver, neural progenitor, occipital lobe, parietal lobe and spinal code. Human adult data will be sourced from GTEx and will include brain, heart, liver and white blood cells.

Samples The project will focus on human adult and embryonic samples from brain, heart and liver. ESCs and adult while blood cells will also be used.

Methodology We will apply the CLS methodology that we recently developed. Briefly, RNA samples are quality tested and reverse transcribed into full length cDNA. These are used to create Illumina barcoded libraries, which are then pooled, captured and sequenced by PacBio.

Proteomics Several recent publications have made dramatic claims about the numbers of coding and non-coding variants that can be verified by large-scale proteomics experiments [Wilhelm 2014 PMID: 24870543, Kim 2014 PMID: 24870542, Ly 2014 PMID: 24596151, Hao 2015 PMID: 26146086]. These studies are all highly flawed either because of the misuse of target-decoy strategies when estimating false discovery rates (FDR), or because of technical errors (or both).

These CNIO has shown that great care must be taken when using data from these large-scale proteomics experiments [Ezkurdia 2014 PMID: 24939910, Ezkurdia 2015 PMID: 25014353, Ezkurdia 2015 PMID: 26496066]. Overestimating the numbers of coding and non-coding variants identified in the experiments wastes time and resources and propagates false positive identifications in databases. This noise will obscure real biological insights, has been used to justify unjustifiable scientific hypotheses [Hao 2015 PMID: 26146086] and in the long term will undermine confidence in large-scale proteomics data [White 2011 PMID: 21325204]. There is currently no reliable strategy to estimate estimate false positive rates in large-scale proteomics experiments [Savitski 2015 PMID: 25987413]; in part because of problems with the narrowness of the mass precursor windows in the most recent high-resolution mass spectrometers [Cooper 2012 PMID: 23106481**,** Bonzon-Kulichenko 2015 PMID: 25494653]; in part because post-translational modifications, which are much more widespread and add much more complexity to the human proteome than once thought, are rarely identified properly [Bonzon-Kulichenko 2015 PMID: 25494653]; and in part because these errors are exaggerated when multiple smaller experiments are combined to make a single large-scale experiment [Reiter 2009 PMID: 19608599]. These problems are common to all proteomics studies, but are especially critical when the experimental evidence is used to verify variant translation. Ideally the community should come up with standards to deal with the growing level of false positive identifications in large-scale experiments. The CNIO's partner, the CNIC are 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.

We will apply high accuracy quantitative shotgun tandem mass spectrometry to characterize the proteome to an unprecedented level of detail. Previously, we have used proteogenomics to identify novel CDS within the non-coding space in GENCODE via a global strategy. More recently, we have performed a targeted investigation of CDS that were initially suggested by phyloCSF data found outside GENCODE annotations, i.e. based on sequence conservation. Moving forward, our overall goal is to use our proteogenomics workflow to target existing or prospective GENCODE CDS annotations that have equivocal supporting evidence based on conservation or transcriptomic libraries. Mass spectrometry peptides can thus provide the requisite orthologonal evidence for these annotations to be made or retained with confidence. Our usage of both discovery and targeted mass spectrometry approaches will allow us to analyse selected gene features on a locus by locus basis. In particular, we will be able to study the coding potential of alternative splicing within protein-coding genes, the relationship between alternative transcription start sites (TSS) and translational initiation sites (TIS), the translation of peptides within 5' UTR regions, and the coding potential of loci currently annotated as pseudogenes. Importantly, mass spectrometry also has the potential to answer long-standing questions about the existence of lineage-specific functionality within the transcriptome, i.e. to investigate translation within potential human or mouse coding regions that are not supported by conservation. Furthermore, we will be working from the same samples as used for the generation of RNA seq

libraries, and the creation of 'multi-layered' expression profiles will allow us to systematically disentangle the factors that determine the relationship between the levels of mRNA and protein in selected human and mouse tissues. Recently, the community-standard approach of using RNA and not protein to measure gene output has been called into question, and our fully integrated approach will therefore provide important insights in this regard.

We will use 3 approaches: (1) proteogenomic analysis of public shotgun proteomics datasets; (2) capture deep quantitative shotgun proteomics using high-resolution shotgun sequencing of key tissues; (3) targeted proteomics to validate expression of selected gene features.

Analysis of shotgun proteomics datasets We have developed an open source pipeline in OpenMS for proteogenomics data analysis. The pipeline is flexible, modular and built to be extendable by the community. We have applied this pipeline to analyse substantial public human datasets for refining genome annotation, this has resulted in the discovery of several new genes (Wright et al). This pipeline will be used to process data for public datasets focusing on the mouse tissues in the first instance. The pipeline can also be used for personal annotation by up loading the associated DNA or RNA sequencing files as reference database. We have also devised a priority annotation score to distinguish peptides that are more likely to lead to novel annotation. For GENCODE we will focus our analysis to the same biological samples as the RNA-seq study.

Use high-resolution shotgun sequencing to refine genome annotation GENCODEv24 contains 79,930 coding transcripts and in the majority of cases this functional annotation remains putative. All results will feed back into the main annotation pipeline and be subject to manual analysis.

Proteogenomics Pipeline Overview We have developed a sophisticated and robust workflow for the analysis of proteogenomic experimental data. So far this work flow has been applied to the analysis of 52 million spectra that were made public as part of the draft human proteomes published in Nature [REFs]. The pipeline is now being applied to other human tissue datasets including the recent Human Proteome Project datasets [REF], and other species such a mouse where we hope to apply the analysis to various strains of mice commonly used in biological research using publically available tissue data, such as the mouse tissue map [REF], and in house collected data from a wide range of mouse tissues.

Bespoke Sequence Database

Construction of a bespoke search database that encompasses the current known proteome of the sample species as well as additional novel protein sequences is essential to proteogenomic analyses. To do this we obtain protein-coding sequences from GENCODE and UniProt, and combine them with translated Pseudogenes, lncRNAs, 5'UTR sequences, predicted genes (ie AUGUSTUS [REF]), PhyloCSF regions [REF], RNA-seq transcripts, and additional selected ORFs from a six frame translation of the genome. For personal proteomics experiments any sample or individual specific sequencing information is also added to the standard proteogenomic database. To this database we add a set of contaminate proteins from the cRAP database [REF] before using an in house tool (DecoyPYrat) to create accurate hybrid reversed and shuffled decoy databases, with low redundancy to the target database. All isobaric amino acids are converted to a single code (ie isoleucine and leucine).

Pre-search Protein Clustering

To improve protein inference and remove ambiguity in protein identifications, the sequences in the database are clustered as described by Nesvizhskii et al. [REF]. Proteins are theoretically digested into their peptides and clustered together based on their proteotypic peptides. Proteins are merged together if the peptides in one protein are a superset of another. The peptides that are unique to each protein cluster are then reported and used at a later point for assigning identified peptides to proteins.

Spectral Conversion

The conversion of raw mass spectra from in-house and publically available experiments into the standardised mzML format [REF] is achieved using the msconvert tool, part of the ProteoWizard suit.

This allows us to convert raw spectra from all commonly used mass spectrometry instruments. At this point we also centroid the spectra, remove any empty spectra, and merge fractionated samples into single mzML files. These mzML files form the basic input into the OpenMS workflow.

Peptide Identification

Our proteogenomics pipeline has been constructed around the open source OpenMS workflow platform [REF]. This platform allows fast high throughput analysis of proteomics experiments making use of the Sangers high performance computing clusters HPC. Many of the components we use already existed in OpenMS, however to fully realise our pipeline we have contributed new analysis nodes and further developed others to give us the functionality we require. The mzML spectra are searched in parallel against the bespoke proteogenomic sequence database using multiple search algorithms. The two main search programmes we use are Mascot [REF] and MSGF+ [REF]. The results of each of these two algorithms are individually post-processed using MSGF+ Percolator [REF] and our in-house MascotPercolator [REFs]. The results of the separate Percolator steps are then merged taking the worst score (Posterior Error Probability, PEP) for each peptide spectrum match across the two searches. This has the effect of only retaining PSMs that are significant in both sets of results. The workflow then has a series of stringent filtering steps to obtain a list of significant peptide identifications. These filtering criteria encompass a minimum FDR, PEP and a minimum peptide length.

Spectral Clustering

As part of our proteogenomics pipeline we have developed and in-house spectral clustering tool, MSSMIV, which examines mass differences between identified and unidentified precursor masses and then scores the similarity between their corresponding MS/MS fragment spectra taking into account the peptide mass difference. This helps identify spectra originating from peptides with modifications and amino acid substitutions which we would not have been able to correctly assign in the sequence database search.

Protein Inference

Peptides mapping to contaminate proteins are removed at this stage. The remaining distinct peptides are used for protein inference using the pre-compiled list of protein clusters and FIDO [REF] which reports protein level probabilities. Only proteotypic non-shared peptides are used for protein inference to avoid ambiguity in identifications. Proteins can additionally be scored using highest score unique peptide mapping and a protein level false discovery rate (FDR) reported. The identified proteins and their mapped peptides are then divided into a set derived from known protein-coding genes and a set that represent novel proteins either from incorrectly annotated genes or unannotated genome loci. The known coding gene set is further divided into those with a single expressed transcript and those with evidence of alternative transcripts.

Peptide Mapping

Our in-house peptide to genome mapping tool finds the specific genomic co-ordinates for all identified peptides, and highlights those which validate exon / intron boundaries. The output from this mapping tool can be formatted as one of the common genomics file formats (GTF, GFF, BAM, BED), the peptide mappings which can additionally include peptide abundance, modification and uniqueness information can be then loaded into a genome browser such as Ensembl [REF], UCSC [REF] or Biodalliance [REF]. The BED format is further converted into a bigBED format and used to create a proteomics track-hub [REF]. We are currently working on Gtf. Format for integration to IGV.

Novel Peptide Analysis

Novel peptides mapping to proteins not currently annotated as CDS in the genome are separated and further filtered to a higher stringency. Overly long or short peptides and heavily mis-cleaved peptides are disregarded. The number and propensity of modifications are also examined, and any modifications that appear over represented amongst the novel peptides are removed. Each novel peptide is then search against all known CDS proteins allowing up to two amino acid variants in the sequence,

matching peptides are again removed from the novel analysis. The remaining novels are ranked based on a priority annotation score (PAS) [REF]. The ranked list is then passed to manual genome annotators for inspection. Additionally any evidence for genes expressing multiple alternative transcripts is extracted and filtered with the same criteria as novel identifications.

Priority Annotation Scoring

This score uses peptide features targeting criteria that appears to distinguish peptides leading to annotation from those do not. The aim of this score is to rank novel peptides identifications, highlighting the protein and peptide mappings most likely to lead to new annotation. The peptide priority annotation score is based on various peptide features that are not normally considered in PSM scoring including the posterior error probability, the number of PSMs observed, the number of replicate identifications, the delta scores between top and second rank assignments, peptide length, and the amino acid sequence complexity. The summed score of all unique peptides identifying a novel protein is then used to rank putative novel protein identifications, prior to manual annotation.

PTM Identification and Localisation

Peptide modifications biological and artefactual are regularly included in the Mascot and MSGF+ searches. Error tolerant search can be conducted to establish the most common modifications occurring in a sample. Further modified spectra can also be discoverer using the spectral clustering tool MSSMIV. For experiments target towards identification of biologically important post translational modifications such as phosphorylation, we have developed part of our workflow to include site localisation tools such as Lucifor [REF] and TurboSLOMO [REF].

Protein Quantification

Our pipeline implements multiple different quantification tools which can be applied depending on the sample. We have developed a TMT [REF] specific workflow in OpenMS for quantification of our labelled samples. For unlabelled experiments we have two different methods, the first is intensity based quantification of proteins using OpenMS tools for feature finding and resolving missing values, the second in an in-house developed method for genes and transcripts which treats identified PSMs like RNA-seq reads and generates a PPKM (PSMs Per Kilobase of transcript per Million) which is similar to the RPKM (Reads Per Kilobase per Million) abundance value reported in RNA-seq experiments.

Pipeline Outputs

The workflow for most of these steps described above is integrated into OpenMS with outputs being obtained from particular steps. The main proteomics outputs from the workflow are mzML formatted XML files containing all the processed spectra, and mzTab files which are tab delimited files containing experimental and analysis meta-data, inferred proteins, quantified peptides, and a full list of all identified PSMs. These are comprehensive files with strict controlled vocabulary containing all information and statistics acquired from the identification process. The mzML and mzTab files along with the FASTA sequence database searched are uploaded to the PRIDE and ProteomeXchange repositories through which they are made publically available. The quantified genes and transcripts reported in the experiment and further submitted to ExpresionAtlas which allows browsing and comparison of the quantified genes / transcripts across the multiple tissues and experimental samples. Also as mentioned previously the genome mapped peptides in bigBed format are used to create a public track hub which can be used to browse the search results in a genomic context. Novel identifications ranked by their PAS, are written to a tab delimited table which is passed to the genome annotation groups to allow further investigation along with other orthogonal evidence for the addition or modification of a new protein-coding gene.

Regular analysis of GENCODE annotation sets The CNIO proteomics analysis has now been run for the GENCODE v3C, v7, v12, v20 and mouse M2 releases. For the GENCODE v12 reference set we analyzed human spectra from seven different experiments and databases and included a series of stringent filters to improve the reliability of the identifications. We identified peptides for 11,840 genes [Ezkurdia 2014 PMID: 24939910] and found a strong relationship between proteomics identification and

conservation. Most of the genes detected were highly conserved. Indeed we found peptides for more than 96% of those genes that evolved before bilateria. The opposite relation was also true; primate-specific genes, genes without any protein-like features and genes with poor cross-species conservation had almost no peptides. This discovery was the inspiration for our coding gene analysis pipeline.

The GENCODE v20 analysis [Abascal 2016 PMID: 26061177] employed eight data sets (including the spectra from the recent Nature papers [Wilhelm 2014 PMID: 24870543, Kim 2014 PMID: 24870542]) and further stringent filters. The study identified 277,244 peptides that mapped to 12,716 coding genes (64%). The mouse M2 analysis used three data sets and identified 12,000 genes [Abascal 2016 PMID: 26061177].

When integrated into the evidence tracks by the annotation team detected peptides can be used to confirm novel isoforms or can help reclassify nonsense-mediated decay targets. We have also used the peptides detected for each GENCODE release to make suggestions for refinements of gene model structure and to annotate new transcripts .The reliable validation of the translation of protein-coding isoforms will continue with each new GENCODE human release and will be extended to mouse.

Plans to leverage and integrate data from other genomics resources (~ 1 page)

From UCSC (Aim 1 Text)

As transcript isoform sequencing data becomes available during the proposed grant period we will extend our RNA-Seq TSL process to integrate information from a broader array of samples, using the growing body of freely available RNA-Seq data, such as GTEx and ENCODE. This large collection of diverse experimental data will be used to more comprehensively establish or refute the expression of apparently rarely expressed transcripts. To make this analysis cost effective, we will use UCSC's high-throughput, low cost, cloud-based genomic analysis platform, which we estimate can be used to analyse expression for less than $0.5/sample. As a proof of principle, we recently used this platform to analyse 20K samples using a single, consistent pipeline with two different methods for estimating isoform expression [Link to BioArxiv preprint]. As long-read technologies, such as PacBio and Oxford Nanopore, mature and become widely available we will adapt our pipelines to incorporate them. This will further improve our ability to characterise the expression of rare isoforms over the course of the project period. UCSC also has several key tools for annotating pseudogenes. In the proposed project, coordinating with the Yale group, we propose to continue running the PseudoFinder system for finding processed pseudogenes as part of GENCODE pseudogene annotation process.

Plans to coordinate with related data resources (~ 1 page)

Sequence Ontology We are currently working with the Sequence ontology (SO) consortium (Ref) to identify the best SO terms relating for 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.

ZMAP – Regulation Comments from Adam:

We will develop the Zmap annotation browser to enable the display of diverse additional data types such as cis-regualtory and physical interactions. We will pilot the integration of these datatypes with

those on which we currently base annotation to annotate of regulatory features, annotate the connection of regulatory feature to genes and annotate transcripts and genes associated with regulatory features for example elncRNAs, bidirectional lncRNAs, alternative 5' UTRs originating from alternative promotor and enhancer sequences.

Research Strategy (Production Core): 12 pages

The central focus of the project should be the generation of a research resource using established, state-of-the-art technologies. All applications should include well-defined goals and milestones that describe what will be accomplished during the award period. Data production and curation should become more efficient over time; the application should describe how this will be achieved. Applications should include plans for distributing the data, software, or biological materials, since the major goal of this program is to provide wide access to broadly useful resources. Applicants should describe how they plan to provide outreach to the community to enable researchers to use the resource effectively. Applications should also include plans for maintenance and distribution of the resource beyond the period of the award. Such plans should acknowledge that, at the end of the project period, all data or resources generated by the project must be transferred to NIH or NIH-approved institutions if they were not already placed in lasting databases or repositories accessible to the broad scientific community. All software must be made widely available to the broad scientific community.

This section should describe in detail how the specific aims will be accomplished, production methods, expected outputs, metrics for production and quality control, and quarterly milestones for each aspect of the production activity.. For resources that are producing or curating data over several years, the application should describe how the approaches will become more efficient and cost-effective over time. The investigators who will be responsible for the project should be indicated and their roles described.

Complex applications for large awards should include all the elements below. Applications that are less complex may require fewer pages and may skip elements below that are not relevant.

Special requirements for informatics community data resource projects

1. The quality control procedures to be used: The development and use of various quality control methods are encouraged, but should be justified based on how they benefit the resource and avoid generating erroneous data and propagating errors to automatically annotated records. The proposed metrics of data quality should be described, including discussion of how comprehensive the data will be. The application should indicate the values of these quality metrics that the resource expects to reach.

2. The plans for maintaining the stability of the resource: Issues that should be addressed include, but are not limited to, the frequency of data versioning, API (application programming interfaces) and web services modifications, changes to data cross-referencing with other sources, and consistency of user interfaces.

3. Plans to improve curation: The application should describe the types of curation tools needed to generate the information for the resource, and how the curation process will be made more efficient. This includes ensuring that curators have the appropriate training and tools. The application may describe tools in use as well as propose the development of improved or new tools, which should be focused on the needs of the project.

4. Any plans to scale up the curation process: This includes the development of scalable methods to speed up both manual and computational curation processes and to incorporate large data sets. The development of these methods should be focused on the resource rather than being open-ended research activities. These methods should enable the resource to curate large data sets more efficiently.

5. If appropriate, how community annotation would be incorporated into the resource: Providing such a mechanism is recognized as a challenge, but efforts should be made to engage the user community and benefit from its knowledge. Applications should describe how this information would be solicited, vetted for quality, incorporated into the project, and attributed to submitters.

6. The plans for obtaining input on user needs: The application should describe how use of the elements of the resource will be monitored, and how, either continuously or periodically, user input and broad assessments of user needs will be done to inform priority setting.

Resource Sharing Plan: Individuals are required to comply with the instructions for the Resource Sharing Plans (Data Sharing Plan, Sharing Model Organisms, and Genome Wide Association Studies (GWAS)) as provided in the SF424 (R&R) Application Guide. The resource sharing plans should be provided only in the Overall Component.

All applications, regardless of the amount of direct costs requested for any one year, should address a Data Sharing Plan.

Quality control procedures to be used (~ 1 page)

Misannotation and QC GENCODE already have robust QC in place to detect errors in annotation of gene structures. Following the completion of first pass gene annotation in human annotators reviewed ~10,000 splice sites flagged as being unsupported or weakly supported by a UCSC QC pipeline based on the alignment of EST and cDNA evidence. With the exception of one class of intron, those with non-canonical splicing, error rates were low; the overcalling of errors likely due to the availability to manual annotators of alignment tools with far greater sensitivity than the pipeline used for QC. All non-canonical splice sites were checked and correct as necessary and all those that remained part of the geneset were tagged with an attribute to indicate the identification of the non-canonical splice site and the reason for its retention in the geneset for example where it was conserved across species. More recently we have undertaken using RNA-seq data from multiple tissues produced by the Gtex project to identify introns that are not found in any RNA-seq dataset, suggesting they unliklely to be expressed. We will tag all transcripts containing an intron that is not expressed and are currently determining the threshold for expression and number and range of tissues sampled at which non or low expressed transcripts are removed from the geneset. We will extend transcript structure QC to include novel RNA-seq dataset and utilise 3rd generation transcriptomic data to comfirm complete transcript structures as the scope and depth of such datasets permits.

Functional misannotation, i.e. annotation of transcripts and loci with incorrect biotytpe, the effect of which is that genuine CDS may be unannotated, while existing CDS may not represent actual translations can occur where there is non-standard annotation i.e. the annotation protocols are not adhered to, where annotation protocols are correctly applied but do not reflect the biology of specific gene locus and where there is insufficient data at the time of initial annotation to correctly define the biotype. In the first instance we undertake extensive QC to identify likely misannotated transcripts on the basis of the transcript properties compared to other transcripts at the same locus. The second and third classess of errors are more difficult to detect but are often found during list based checking for example PhyloCSF regions or proteomics data intersecting with lncRNA annotation, checking and reannotation carried out as part of the CCDS project or following literature review.

In order to confirm the true protein-coding functionality of annotated transcripts we will investigate coverage by RP data, and shotgun and targeted proteomics data. Analysis of RP data and shotgun proteomics will be applied genome wide but prioritisation for targeted proteogenomics analysis will be given to AS transcripts within genes of specific clinical interest that are well supported by transcriptional evidence.

Another important target set for targeted proteogenomics and transcriptional analysis are retrotransposed loci that have arisen recently in the primate lineage. The recent origin and consequent lack of conservation of these loci means that no information regarding their protein-coding potential can be inferred from whole genome alignment data. These loci are frequently transcribed and many contain potential coding regions in addition to regulatory and epigenomics profiles consistent with functionality. Where transcribed these loci are annotated as either processed pseudogenes, if they contain disabling mutations or protein-coding retrogenes where the parent CDS is intact. However, as these loci are generally single exons premature stop codons may result in truncated but still functional proteins, while the specificity of transcription for recently created retrogenes might be difficult to interpret. We will investigate the specific transcription of all these ambiguous loci using an RNA-seq alignment pipeline specifically designed to resolve mapping between pseudogenes and their parents. For all loci where transcription in confirmed we will investigate translation with our targeted proteogenomics pipeline.

Quantitative proteomics will also give insight into the likely functional significance of loci that fulfil the essential criteria to be annotated as protein coding. For example for a lineage specific retrogene of duplicated protein-coding gene with an intact CDS and evidence of transcription knowing the abundance of the protein in the cell, in both absolute terms and relative to parent and paralogous loci, will allow the discrimination of loci that produce a high level of protein providing a mechanism by which

they could have a possible functional role in the cell and those whose basal levels of protein product do not support such a role.

Ensembl QC The GENCODE gene set is compared to other sets (UniProt and cDNA alignments, and imported RefSeq data) to check for missing genes or transcripts. The most recent CCDS database is downloaded and the GENCODE set compared against that to ensure it contains the complete set of CCDS models. The alignments of cDNAs in the INSDC are updated for every Ensembl release. If annotation identified in these external data sets is missing and requires manual annotation, then this is stored in the AnnoTrack system (Kokocinski et al., 2010) so that a record is kept for the annotators to inspect these loci.

Validating novel coding genes

Proteogenomics, which involves searching spectra against databases made from translated non-coding regions, is a growing field (Brosch 2011 PMID: 21460061, Gascoigne 2012 PMID: 23044541, Kumar 2016 PMID: 26773550) Proteogenomics as a technique is most effective with poorly annotated species [Nesvizhskii 2014 PMID: 25357241]. As we have shown, the vast majority of peptides we identify reliably, map to conserved protein coding genes and there are will be few conserved coding genes missing from well annotated genomes like human. The danger of using proteogenomics is that large non-coding databases must be generated to search against non-coding regions [Nesvizhskii 2014 PMID: 25357241]. This affects the false discovery rate calculations and inevitably many spectra will be falsely assigned to these non-coding regions and the larger the non-coding database, the more false identifications. Separating those spectra that correctly identify novel coding regions from false positive matches can only be done by careful manual inspection of the spectra.

The CNIO methodology for validating novel coding genes has centred on searching against known coding databases outside of GENCODE to fill in the gaps in the annotation (UniProt, IPI, RefSeq). These databases are not much larger than the coding genes in GENCODE and are much more likely to contain coding genes. This protocol generates smaller numbers of novel coding genes, which is more manageable for the manual annotation teams, and has a higher hit rate. This has proved a very profitable strategy. Sixty-one genes identified as missing from the CNIO proteomics analysis have been added to the reference set.

Comparisons between GENCODE/Ensembl and the RefSeq and UniProt databases suggest that there now are few coding genes missing from the GENCODE annotation of the human genome. This strategy would bear most fruit with the mouse genome, which is not yet as well annotated as the human genome.

The CNIO and CNIC also participated in a large-scale proteogenomics analysis in conjunction with the Sanger Centre, where a further 16 coding regions were identified after much filtering and several rounds of manual validation [Wright 2016 submitted].

Validating gene models The CNIO has attempted to validate alternative splice isoforms at the protein level. However, reliable proteomics data identifies only a fraction of the annotated alternative isoforms. We used our proteomics analysis pipeline to identify translated splice variants annotated in the GENCODE v20 reference set [Abascal 2016 PMID: 26061177]. While we detected peptides for 12,716 protein-coding genes (64%), only 246 genes (1.2%) had reliable evidence for more than a single isoform. This number is considerably smaller than expected: according to simulations we would have expected to detect alternative isoforms for over 3,500 genes if all isoforms were equally detectable and 1,500 genes if one isoform per gene was 100 times more detectable than the others.

This strongly suggests that most protein coding genes have a single main protein isoform [Ezkurdia 2015 PMID: 25732134] and that alternative variants are not abundant at the protein level. This is in stark contrast to the abundance of alternative transcripts in microarray and RNA-seq experiments and is especially surprising in light of the fact that the eight large-scale experiments in the CNIO protocol interrogated more than 100 different tissues, cell lines and developmental stages [Ezkurdia 2015 PMID: 25732134].

Alternative splice events supported by proteomics evidence were those with the most evidence of cross-species conservation (this was true of genes too) and those that would have relatively minor effects on the structure and function of the main isoform. They were significantly enriched in mutually exclusively spliced homologous exons and in subtle splice events that did not disrupt Pfam functional domains.

Apart from the analysis of alternative splice isoforms the CNIO has also attempted to identify exons and transcripts that are not represented in the GENCODE human reference set. While the set of GENCODE human genes is close to complete, the gene models are not. A comparison of the UniProt and GENCODE reference sets showed that only half of the genes had the main isoform. The CNIO has used proteomics evidence and data from the APPRIS database to flag gene models in more than 300 human genes for the annotators.

The CNIO/CNIC proteomics pipeline will address not just the verification of annotated GENCODE coding genes, but will also incorporate protein coding sequences from the UniProt and RefSeq databases with the focus on improving the gene models, not just for human, but for mouse too.

Plans for maintaining stability (~ 1 page)

EXPECTING ADDITIONAL TEXT FROM ED GRIFFITHS / ANDY YATES HERE

The stability of GENCODE is important at a number of levels. At the most fundamental level, we must ensure that our computational and software infrastructure is well maintained to support the needs of the project for production and curation of data. At the next level, highly functioning processes are needed to ensure that we are able to update the GENCODE annotation following our current schedule of 4-5 releases per year. Finally, the project must ensure that new manual annotators joining the team are adequately mentored and trained to ensure that their decisions are consistent with their colleagues and with those made over the history of the project.

Software infrastructure. The annotation tool Zmap/Otterlace, updated via monthly releases that incorporate changes to enhance the annotation process e.g. better visualization of RNA-seq data and 3rd gen transcriptomic date in both the Zmap browser and Blixem alignment viewer, ability to import new datasets via trackhubs, improved filtering and highlighting options in Zmap, improved drawing speed, etc.

Update cycle and process.

Consistent manual annotation. Maintaining consistency of manual annotation across the HAVANA team starts as soon as a new annotator joins the team. Annotators receive training and feedback for their first year in the team and are mentored by experienced annotators subsequently. Annotators are expected to adhere to teams extensive, and publicly available, annotation guidelines (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf) unless a discussion with a more experienced annotator suggests otherwise. The guidelines are constantly updated to ensure they keep pace with the datasets we use routinely and in response to annotator feedback to make them clear. All annotators in the team have taken part in annotation consistency tests where the whole group annotates the same region over two days under ‘exam conditions’ and the annotation of each team member is compared to reference annotation produced by two senior annotators. All deviations from the expected annotation are fed back to the team, both in individual meetings and group discussions. To provide a more regular mechanism to detect and address inconsistency, a similar exercise is undertaken for a “locus of the month”. All team members provide annotation for the same (often complex) locus in parallel and their results are compared with reference annotation for the same locus. Again, all deviations are discussed in individual and group meetings and

annotator feedback is used as a basis to improve the guidelines. We continue to upload illustrative examples of annotation to the HAVANA internal wiki to provide a library of complex cases to help in the training of new annotators.

Plans to improve curation (~5.5 pages)

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 perspective. The first major focus will be completing the annotation to the extent technically and operationally possible and will include a focus on extending the existing human partial transcript models to full length, expanding the human lncRNA annotation improvement as well as the completion of the initial full pass of the mouse GENCODE annotation. The second major focus of will be the incorporation of individual genome representation and population data represented by available human variation data at both the sequence level (e.g. 1000 Genomes) and at the transcriptomic level (e.g. GTEx) and by the 17 mouse strain genomes produced by the Mouse Genomes Project led by the Sanger Institute. Beyond these fundamental efforts, two planned pilot projects will be undertaken with the goal of improving the process of annotation and to expand the overall utility of GENCODE.

Toward completing the GENCODE annotation

Projects In addition to clone-by-clone annotation we have undertaken wide ranging targeted annotation projects to both improve the existing annotation and identify missing loci and transcripts. For example we undertook a thorough review of manually annotated protein-coding associated with analysis from CNIO (REF) that led to the updating of hundreds of previously protein-coding loci to other biotypes. We have engaged in annotation of large gene families such as olfactory receptor genes in human and mouse using a small team of annotators and strict interpretation of guidelines to maintain consistency. We are now integrating RNA-seq datasets (REF) in both species to achive a consistent annotation of UTR sequences. Similarly we assigned a small team to complete the annotation of all immunoglobulin and T-cell receptor genes and pseudogenes in human and mouse. Further QC efforts have led to the definition and tagging of a set of ~350 retrogene loci in human and checking all annotated microexons smaller than 10 bases in length. We have updated all pseudogenes that were annotated without a biotype giving information about there mode of creation and have undertaken reannotation of the human processed pseudogene set to identify their sites of insertion into the genome. We have undertaken extensive QC of so called ‘readthrough’ loci whose transcripts overlap multiple independent loci. Prior to the release of GRCh38 we annotated patch and haplotype sequenes and following the release we lifted annotation across to the new reference genome, checking and correct ~600 loci that were mapped inconsistently and also undertook annotation of regions novel to GRCh38 that were never released as GRC patches. Given the large numbers of researchers who are still working on GRCh37(hg19) we have developed a pipeline to lift back improvements to annotation captured in GENCODE releases based on GRCh38 back to GRCh37.

In addition to these QC projects we have also worked to integrate new datatypes into the standard annotation process to aid discovery of novel features. For example using PhyloCSF regions, MS data to identify unannotated protein-coding loci and develop pipelines and strategies to prevent the false positive identification of novel protein-coding loci(Ref). We have investigated the use of CAGE and RAMPAGE data to confirm TSS and, in combination with RP data to identify TiS to identify loci with multiple TSS that affect the TiS and hence the CDS encoded. We have worked to extend 3’UTR sequences on the basis of support from RNA-seq and polyAseq data in addition to EST and cDNAs.

We have investigated the integration of regulatory data, promoter capture Hi-C data into annotation pipelines and annotated regions of the genome proximal to fine-mapped GWAS hits (Ref). We have worked with the output of the CaptureSeq pipeline as it has developed and annotated extensions and AS transcripts at lncRNA loci sequenced with 454 and PacBio data(Ref).

Gene models were manually extrapolated from the alignments using the otter annotation interface (ref). Alignments were navigated using the Blixem alignment viewer (Ref, Barson G and Griffiths E). Visual inspection of the dot-plot output from the Dotter tool (ref) was used to resolve any alignment with the genomic sequence that was unclear or absent from Blixem. Short alignments (<15 bases) that cannot be visualized using Dotter were detected using the ZMap DNA Search pattern-matching tool (ref). The construction of exon-intron boundaries requires the presence of canonical splice sites and any deviation from this is flagged with an annotation attribute to indicate the reason e.g. cross-species conservation for the inclusion of the non-canonical site. All non-redundant splicing transcripts at an individual locus are used to build transcript models. Transcript models are extended only as far as their supporting evidence allows and as such those based on partial evidence such as ESTs are annotation as partial models. Annotation attributes are again used to identify known partial transcript models. Once the correct structure had been ascertained, all transcript models are assessed to determine its most likely functional class or 'biotype'.

Finishing the Mouse Pseudogene Annotation [[CSDS to do]] Currently we are in the midst of completing the mouse reference genome pseudogene annotation, with plans to develop customized pseudogene annotations for the available mouse strains. Using Pseudopipe we are able to identify 18627 putative pseudogenes (9748 processed, 1940 duplicated and 6939 ambiguous) in the reference genome (MM8). Using RCPedia we find 9755 processed pseudogenes while Retrofinder predicts 18467 retrocopies. The tri-way consensus between the three pipelines with respect to the processed pseudogenes is ~80%. We will evaluate the annotation accuracy of our pipelines and refine the pseudogene identification and characterization process by using the manually annotated pseudogenes as a gold standard and comparing them with the automatic predictions.

Status on human-mouse pseudogene comparison Preliminary comparative analysis of human and mouse genomes have shown that they exhibit similar total numbers of pseudogenes while being dominated by processed pseudogenes. At family level we see that most of the pseudogenes are lineage specific and the majority of them arise from housekeeping genes (e.g. ribosomal proteins). By contrast, the age distribution of mouse processed pseudogenes closely resembles that of LINEs, while in human, the age distribution closely follows Alus (SINEs).

Annotating loss of function events in mouse We will build on our experience in identifying and analysing loss of function events in human\cite{20210993}, to develop a reliable annotation framework for unitary and polymorphic pseudogenes, and LOF variants in mouse.

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, and conducting a survey of gene disablements. In order to identify polymorphic pseudogenes we will extend our variant annotation tool to identify variants and frame shifts that revert disabling stop codons.

We will identify putative LOF variants by combining function based annotation, evolutionary conservation and biological networks data into a comprehensive pipeline. For this we will integrate resources such as PFAM and SMART functional domains, signal peptide and transmembrane annotations, post-translational modification sites, NMD prediction, and structure-based features (e.g. SCOP domains). We will calculate variant position-specific GERP scores and dN/dS values to evaluate evolutionarily conservation. We will also 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.

To understand the impact of putative LOF variants on gene function we will 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 are going to leverage our experience in pseudogene transcription analysis in human, worm and fly to study the pseudogene transcription in mouse, significantly improving on previous efforts. We will use RNA-seq data to calculate a RPKM value for each pseudogene as an indicator of transcriptional activity. Next we will highlight tissue and strain specific transcribed pseudogenes. Also, we will integrate tissue specific transcription information and regulatory data with the pseudogene annotation in order to characterize pseudogene activity. In particular, we will focus on the transcriptomics (ENCODE, BrainSpan, TCGA), epigenomics (ENCODE, Roadmap Epigenomics) and cis-regulatory interactions data (GTEx, PsychENCODE). Such information will be valuable for understanding the regulatory potential of transcribed pseudogenes.

Annotation of individual and population data [[TG to do + add the LOF stuff in above + less the genome and more the annotation]][[incorporate a little of LOF & vat]]
Current human genome annotations are based on the reference genome and as such do not provide an accurate representation for the large genomic diversity of the human population. We have developed approaches and tools\cite{21811232} to integrate personal variation data into the reference genome producing the individual's personal diploid genome and annotation. The latter is generated by mapping GENCODE annotations against the individual's personal genome. We have a large experience with constructing personal genomes, splice-junction libraries and personalized annotations and using them in functional genomic analyses \cite{22955620,22955619,24092746, 27089393}.

Using personal annotation allows us to account for differences due to impact of the personal variation on genes and other genomic elements between individuals as well as between haplotypes of the same individual. It has been demonstrated that using the diploid genome with individual's variants improves both mappability of the reads \cite{21811232} and downstream analyses results \cite{26432246}.

A key aspect of personalized annotation is correctly connecting the annotation to loss-of-function (LOF) events and polymorphic pseudogenes (pseudogenes present in most individuals in the population but functioning genes in some). We have explored the implications in detail of these for reference annotation [[cite http://papers.gersteinlab.org/papers/refgeneset]]. In general, these two categories are just different versions of the same event, mostly depending on the major allele frequency.

Specifically for LOF variants, we have developed a tool - Variant Annotation Tool (VAT) \cite{22743228} - to catalogue loss-of-function (LOF) events. It enables variant annotation with respect to a reference genome and a gene annotation model. VAT can identify pseudogenization events such as premature STOPs and polymorphic pseudogenes. 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. The identification and characterization of LOFs as disease related or pseudogenization precursors is important for personal annotation \cite{21205862}.

We characterized putative LOF events in individuals from 26 different populations using the 1000 Genomes Phase 3 data.\cite{26432245}. We also surveyed the impact of LOFs on personal annotation\cite{21205862} and found that LOFs variants that introduce premature STOPs 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. To this end, we have developed a pipeline to identify unitary pseudogenes in human\cite{20210993} and explored the functional constraints faced by different species and the timescale of functional gene loss \cite{20210993}. These results along with fully annotated pseudogene sets are deposited in our repository at pseudogene.org.

[[mostly cut underlined stuff]]

Protein-coding gene annotation in the mouse strains As part of the Mouse Genomes Project, UCSC has developed gene and transcript sets for seventeen additional mouse strains. These include

laboratory strains, as well as wild-type strains, such as *Mus Spretus*. To do this we worked to improve the three interrelated and interdependent problems of genome assembly, genome alignment and genome annotation, iterating on each repeatedly and progressively seeing global improvement in the three aspects. The result is a union gene set for Mus. Musculus that includes several hundred new genes (see Fix X. for an example of a substantial new gene), thousands of new isoforms and tens of thousands of novel gene haplotypes. This is leading to improvements in the existing B6 (mm10) GENCODE reference set - identifying genes that are actually polymorphic pseudogenes, finding unannotated B6 loci and identifying reference assembly errors. It is also providing new insights about the gene variation present across the pan-Mus species, and in conjunction with RNA-seq data, is allowing us to assess differential expression between strains using more accurate transcript sets. The methodological component of this work has resulted in a new, general purpose Clade Genomics Toolkit. This toolkit provides a range of tools that together provide a pipeline to simultaneously and consistently comparatively annotate many genomes, leveraging existing high quality annotations and RNA-Seq data. UCSC has now applied this toolkit to both mouse and primate genomes.

Annotation of individual genome sequence We have long experience of annotating non-reference sequence from GRC both haplotypes and fix patches, indeed manual annotation has proved to be essential for haplotype regions such as those produced for the LRC where a combination of tandem gene duplication and pseudogenisation events present a huge problem for automated pipelines which frequently misannotate transcript structures, joining distinct loci together and functional biotype, with pseudogenes misannotated as protein- same coding. Non-reference sequence is passed through the same analysis pipelines as reference genome sequence and adheres to the guidelines to ensure the annotation added to both is equivalent. The only significant difference occurs where a transcript extends beyond the boundary of a patch region. Where this occurs, an annotator adds as much structural and functional information as possible and adds an attribute to indicate the truncation. Transcripts tagged in this way are managed differently during the merge process with Ensembl annotation of patch regions for which the Ensembl genebuild pipeline sees the patch in its genomic context and is able to annotate full-length transcripts beyond the patch boundaries. Whereas generally the manual annotation is dominant during the merge process, in these circumstances the more complete Ensembl annotation takes precedence.

We also have experience of annotating regions of specific interest on genome sequence unique to individual mouse strains (Ref). Where private or lineage specific regions of the genome are identified, an Augustus based gene prediction pipeline is run and those regions with potentially interesting genic features are flagged for targeted manual annotation. The target regions are passed through the standard analysis pipelines, again to ensure a comparable annotation to the reference genome, however, where strain-specific transcriptomic evidence is available it can also be viewed to aid specific annotation.

Where high quality personal genome sequence becomes available, either publicaly or via collaboration, we will apply this pipeline to regions distinct to the human reference genome sequence, allowing us to identify and capture all novel loci.

HAVANA have developed experience in annotating LoF variation (Ref), 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 allow us to pull in variation data and effectively annotate on non-reference genome sequence, allowing us to represent the functional effect of the variant in its correct context within the transcript. This is significant as simply passing a variant through a variant consequence pipeline such as the Ensembl variant effect predictor (VEP) might indicate a variant 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 for example if the LoF variant was close to the 5’ or 3’ end of the CDS. We are further developing the Zmap software to make this annotation process more straightforward and are investigating alternative ways to save and distribute this information.

We will expand our annotation of personal genome sequence using variation and transcriptomic data from the same individuals to allow us to capture TSS, alternative splicing and TTS events and construct representative transcripts associated with specific variants. In this way we can fully investigate variants tagged as eQTLs, sQTLs, and LoF variants and describe them in their proper transcript context. Furthermore we will work with the Choudhary group to integrate Ntex proteomics data from the same samples, allowing us to compare the impact of variation on proteins and transcripts with particular reference to protein and transcript abundance.

Pseudogene annotation [[CSDS to cut]] We will develop a personal genome annotation resource containing a number of tools and utilities for constructing the diploid personal genome and the personal GENCODE annotation in order to produce an accurate representation of an individual's gene set.

In particular, given an individual's variation data, the proposed annotation resource will be used to identify and analyse GENCODE-annotated features characteristic to the individual, such as their distinctive set of functional genes or structures of variant-affected transcripts. Using our annotation pipelines we will create a comprehensive personal pseudogene complement. We will use the newly constructed personal annotations to identify LOF and pseudogenization events by comparison with the reference genome. We are going to assess the annotated personal SNPs for allele specific expression using the data from AlleleDB \cite{27089393}, an online repository that provides 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.

Improving the mouse strain pseudogene annotation The relatively small divergence time frame between the mouse strains \cite{25038446} allows us to map the reference mouse annotation on each of the strains using the UCSC LiftOver tool. In addition, we will develop extensions to the available annotation pipeline to use the UCSC strain dependent protein coding annotation as input in order to draft each strain's pseudogene complement. Using these two annotation sets will allow us to produce an accurate map of pseudogenes and loss-of-function events in mouse strains.

[[CSDS2MG I think that the paragraph above ~Improving the mouse strain pgene annotation~ will be more suitable in the **Plans to scale up the curation process (~2.5 pages)**

Extending Automatic Pseudogene Annotation Pipelines section below]][[OK make the change]]

Personal Proteomics Analysis

One particular workflow within our pipeline focusses on personal proteomics, whereby we compare samples from multiple individuals to identify differences is gene and transcript expression, determination of allele specific expression, differences in alternate splicing of genes, and to identify sequence variations such as SNPs. To achieve this we use multiplex TMT labelled samples which allows direct comparison of peptide abundance with in a spectrum and easily highlights cases where a peptide is not present in a particular individual. Currently we have processed a set of healthy and diseased human knee samples from 12 osteoarthritis patients. The example depicts quantitative analysis of NOS2 from proteomics and transcriptomics of three individuals. The protein appears more abundant in one individual, the combination of proteomics and trascriptomics resolves the coding isoform between 2 alternative transcripts. We are collecting TMT data for ENtex tissues samples which we will process through this personal proteomics pipeline.

Pilot project 1: Graph genomes representation

The human and mouse reference assemblies are no longer linear. Instead, they contain additional alternate sequences that provide new genomic sequence to either expand the haplotype represented in the reference genome ('novel' patches) or to provide improvements to known errors on the primary

assembly without changing the stable coordinate system ('fix' patches). These alternate sequences therefore provide new paths through the genome in the same way that is planned for graph genomes in the future. Dealing with patches in a sustainable and scientifically valid manner is therefore important as a preparation for graph genomes.

Human polymorphism is less extreme than that present between mouse strains, but complicated by the relatively higher rates of heterozygosity (most of the considered mouse strains considered are inbred, and therefore individual genomes are essentially haploid). To pilot an approach to population based genome annotation we will use our recent experience in developing population reference genome graphs (Fig Y). In brief, there is increasing recognition that a single, monoploid reference genome is a poor universal reference structure for human genetics, because it cannot include a significant fraction of human variation. Adding this missing variation results in a structure that can be described as a mathematical graph: a genome graph. Multiple groups are now collaborating to construct to complete reference genome graphs, annotated with rich haplotype information (see letter of support from Gil McVean). We propose to pilot mapping transcript structures into genome graphs. A genome graph data structure is a naturally compact way to associate genetic variations with specific isoforms, each isoform approximately corresponding to a specific path within the graph. Devising algorithms and a nomenclature to relate isoforms to these named variations is one way we might achieve a path to consistent population level annotation, which could avoid an unmanageable explosion in the annotation task that would result from instead attempting to independently annotate many individual human genomes.

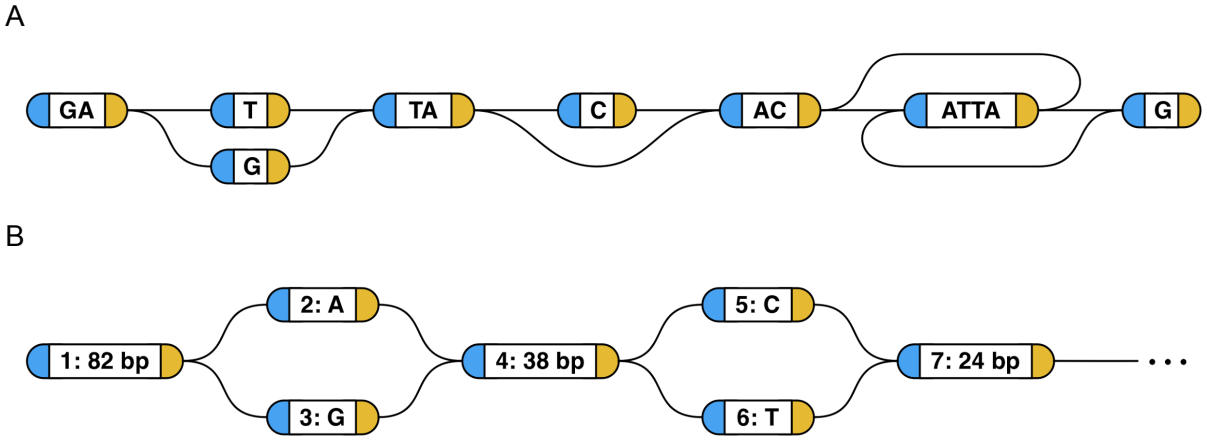
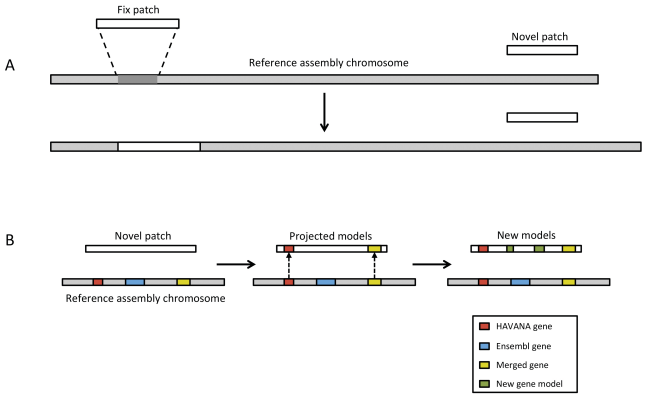


Figure Y: Example genome graphs. Each segment (node) holds some number of bases. A join (edge) can connect, at each of its ends, to a base on either the left (5', blue) or the right (3', yellow) side of the base. When reading through a thread to form a DNA sequence, you leave each base on the opposite side from which you entered, and reverse complement it if you enter on the 3' side and leave on the 5' side.

The graph in A is an example graph, showing the capabilities of the system. One thread that this graph spells out (reading from the left side of the leftmost sequence to the right side of the rightmost sequence, along the nodes drawn in the middle) is 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 in B is the beginning of the genome graph for BRCA2 derived from the 1000-genomes phase three 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)

Patch annotation. The Genome Reference Consortium (GRC) assembly patches are alternate sequences that fix problems in the reference assembly or add alternate loci. These regions potentially contain new genes, or allow improved representation of gene structures by fixing the genomic sequence underlying them. The GRC release a patch set approximately every three months for human. These new sequence regions require annotation. To create initial annotation on the complete set of patches Ensembl has developed a pipeline to rapidly automatically annotate these, mainly by projecting annotation from the reference assembly, but in places where the sequence has changed, by running a localized automatic gene annotation pipeline.



Pilot Project 2: Connecting regulatory regions to regulated genes

For a gene to function effectively, it has to produce the right molecular product in the right cell and at the right time. The regulatory regions that modulate its expression are therefore key components of the proper function of a gene. We propose to enhance the current annotation of GENCODE genes with their regulatory elements, depending on tissue and cell type (Figure 5).

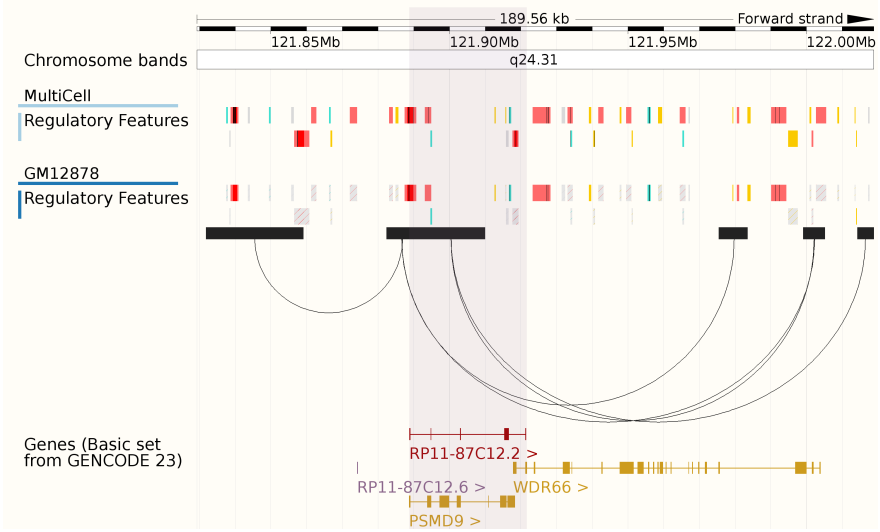


Figure 5: A prototype gene annotation. RP11-87C12.2 is currently represented as a set of exons and UTRs along the genome, however it is surrounded by cell type dependent regulatory elements. In addition, the ties between these regulatory elements and promoters (represented as arches) are also tissue dependent.

GENCODE has provided the scientific community with a very deep understanding of the coding regions of the human and mouse genomes, providing a key resource for genomics. In particular, it provided a stable and curated annotation of the coding regions of the genome, providing a common reference for genomic studies. It allowed the research community to create a bridge from variants in coding regions to changes in molecular product and downstream phenotypic effects. Although very rich, this annotation provides little indication of the dynamic function of genes, namely expressing products in the appropriate tissues, cell types and conditions. Alterations of these regulatory mechanisms have been shown to play significant roles in health, phenotype and evolution (Freedman, PMID 2161409; McLean, PMID 21390129; Levine, PMID 12853946). In

particular, it is estimated that a large fraction of the genome is directly involved in modulating the expression of genes, however it is still poorly understood (Kellis, PMID 24753594).

The epigenome is a dynamic feature of the cell that fluctuates depending on cell type, developmental stage, cell cycle, circadian rhythm, age, environmental conditions, etc. A large number of enhancers can be observed springing into action or returning to inactivity in a programmed fashion. Detecting regulatory elements therefore requires collecting as many datasets across as many conditions as possible, to detect even the most fleeting regulatory activity. Despite the drop of sequencing based assays, understanding the dynamics of epigenomes is still experimentally onerous. This is why the scientific community has started charting the non-coding regions of the genome within large consortia to better pool and coordinate resources. Regulatory regions are being characterized thanks to large consortia, such as ENCODE, Epigenomics Roadmap or Blueprint (Koch, PMID 17567990; ENCODE, PMID 22955616, Roadmap, PMID 20944595, Blueprint, PMID 22398613), all grouped under the umbrella of the International Human Epigenome Consortium (IHEC).

Having identified possible regulatory regions, we then need to determine their downstream effects. Various strategies have already been adopted to attach target genes to these regulatory regions: correlation of dynamic signal (Thurman, PMID 22955617; Andersson, PMID 24670763), genetic association (Dixon, PMID 17873877; Stranger, PMID 17289997) or physical proximity (Dostie, PMID 17446898; Fullwood, PMID 19890323; Leiberman-Aiden PMID 19815776; Schoenfelder, PMID 25752748). Large reference datasets have been produced by teams such as ENCODE (Koch, PMID 17567990; ENCODE, PMID 22955616), FANTOM5 (Andersson, PMID 24670763) or GTEx (GTEx, PMID 25954001). Despite their potential, these datasets have yet to provide us with a clear map of gene regulation, once again for technical and biological reasons.

In this aim, we propose to define the exact boundaries of regulatory regions, define the attributes of these elements and infer their target genes depending on tissue. Because the available data is extremely variable, both for biochemical and experimental reasons, our strategy consists in a) collecting as many datasets as possible, b) running machine learning software on them, then c) manually reviewing the results and comparing them to published results to better understand the limitations of the automatic pipelines then feedback improvements into the automatic annotation process.

Task 3.1 Collecting relevant motif, epigenomic and regulatory datasets

To identify regulatory regions, we initially developed the Ensembl Regulatory Build, which is focused on ChIP-Seq datasets, because of their high reproducibility and their well-characterized functional interpretation. In fact, IHEC chose a handful of chromatin marks to characterize epigenomes, alongside RNA-Seq and chromatin conformation. This will allow the major epigenomic data producers to focus on these core assays and create comparable datasets, also known as reference epigenomes. We are key members of IHEC and currently manage epiRR, the central registry of available epigenomes. As such, we will collect and process these reference epigenomes in a systematic way, enriching our catalog of cell type specific epigenomic datasets.

Alongside histone marks, other markers have been used to detect and characterize regulatory regions. In particular, DNA methylation, whether assayed with micro-arrays (Keshet, PMID 16444255) or bisulfite sequencing (Lister, PMID 19829295), has been abundantly measured across many samples. Although less informative than chromatin marks about regulatory function, these experiments are a cost-effective way of measuring activity of known elements. We will therefore extend our pipelines to integrate this alternative data source and extend our breadth of coverage.

Another mark of regulatory activity is enhancer RNA (eRNA), i.e. abortive RNA transcripts that occur on either side of enhancers. Already, the FANTOM5 consortium produced a massive compendium of CAGE-tag datasets across many human and mouse tissues (Andersson, PMID 24670763). We will integrate this information explicitly into our annotation.

Regulatory regions thus detected will be further annotated by their known sequence motifs. At the moment, Ensembl keeps track of known JASPAR transcription factor binding sites (Mathelier, PMID

24194598) and Diana TarBase miRNA target sites (Vergoulis, PMID 22135297). In the future, we will expand our annotation with SELEX (Jolma, PMID 23332764) and Uniprobe motifs (Robasky, PMID 21037262), as well as other relevant annotated sequence motifs that may be developed.

To attach these annotated regulatory regions to their target genes, we will finally collect as many cis-regulatory datasets as possible. We currently store GTEx eQTLS, but we plan to expand this storage to all available eQTL datasets. We will further integrate correlation calculations as computed from ENCODE or FANTOM5 data. Finally, we will collect physical proximity measurements, including Hi-C, Chia-PET or Promoter Capture Hi-C.

Task 3.2 Integration of experimental evidence

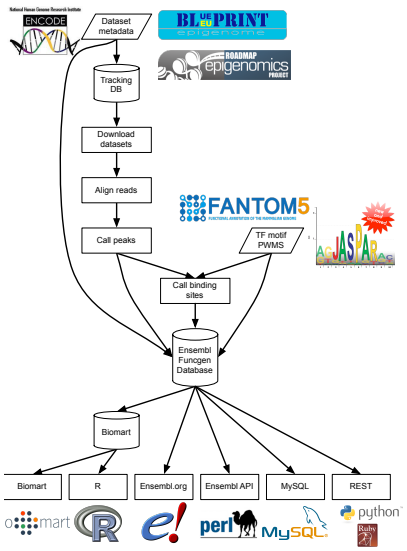


Figure 6: Summary of the existing Ensembl Regulatory Build pipeline: experimental data and external annotations are all integrated into a central annotation database.

Ensembl pioneered the integrative analysis of epigenomic datasets across consortia, developing the Ensembl Regulatory Build (Zerbino, PMID 25887522), a list of enhancer, promoter and promoter flanking regions across the human and mouse genomes. This pipeline takes as input large collections of ChIP-Seq datasets. We will extend our current pipeline to integrate as well DNA modifications and eRNA peaks to cover a greater breadth of experiments.

The different epigenomes will be annotated with rich metadata to characterize all possible parameters: cell type, tissue, treatment, age, etc. We will pursue our active collaboration with the Experimental Factor Ontology (EFO) (Malone, PMID 20200009) team to ensure that this fine-grained representation of our data can be efficiently searched.

<Manolis Kellis: clustering>

The interaction data will finally be integrated across evidence types using a Bayesian model that takes into account the local sequence and epigenomic data as well as the pairwise interaction data. We are currently collaborating with Oliver Stegle to develop such methods (See Letter of Support). This will produce an automatic gene assignment for all enhancers depending on tissue and cell type.

Task 3.3 Manual curation of results and feedback process

GENCODE has demonstrated the importance of manual curation in improving genomic annotations. Historically, genes were annotated using prediction algorithms. However, these systematic annotations were shown to have blind spots, typically difficult loci that contradicted the assumptions of the algorithm designer. With hindsight, a computer program can integrate all the lessons learned from this detailed examination, but in practice new exceptions crop up regularly.

Considering how new our current regulatory and cis-regulatory annotations are, we are certain that critical and detailed examination will reveal unexpected biases and edge-cases, which we will then feed back into our annotation process. Just like gene annotation over the past decade, we expect our regulatory annotation to improve asymptotically over the next several years, to eventually converge to a set of trusted and mostly stable annotations.

Manual curators have various tools available to detect possible annotation errors. The first approach consists in examining elements with extreme features, such as the longest or shortest enhancers. Certain regions such as the HOX cluster have such dense biochemical activity that naive algorithms struggle to break it down into components.

Another approach is the detailed examination of the genome, chromosome by chromosome. Curators will typically extract all available data on a given region and compare it to the annotation, looking for

anomalies. This data would consist of published results as well as pre-computed statistics such as PhyloCSF or XXX.

We will develop the Zmap annotation browser to enable the display of diverse additional data types such as cis-regulatory and physical interactions. We will pilot the integration of these datatypes with those on which we currently base annotation to annotate of regulatory features, annotate the connection of regulatory feature to genes and annotate transcripts and genes associated with regulatory features for example elncRNAs, bidirectional lncRNAs, alternative 5' UTRs originating from alternative promotor and enhancer sequences.

Post translational modifications

The CNIO's partners in the CNIC have recently developed an ultra-tolerant (500Da) Sequest database search based on the method described by Chick et al. (Chick JM PMID: 26076430). It allows the annotation of spectra previously unassigned by conventional database searches. Many of the newly annotated spectra belong to postranslationally-modified peptides (PTM) and SNPs. We will apply the ultra-tolerant search for identification together with our WSPP (weighted spectrum, peptide, and protein) model for statistical analysis to detect new uncharacterised peptides. This model provides a powerful approach for large scale unsupervised analysis of PTMs and SNPs. The number of total peptide matches detected using the ultra-tolerant search is more than twice of that of conventional non-modified search, suggesting that a large fraction of fragmentation events correspond to modified peptides and SNPs. Our newly developed method will allow the reconstruction of a complete dynamic map of the modified peptidome and SNPs.

Plans to scale up the curation process (~2.5 pages)[[FN to do]]

Extending Automatic Pseudogene Annotation Pipelines We are going to 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 the consensus protein coding genes; 2) extract the amino acid sequence of proteins from ENSEMBL peptide database; 3) mark and identify the coordinates of protein coding genes in the analyzed strain; 4) use a six frame blast homology search to match the consensus peptides to the strain sequence; 5) refine results and eliminate redundant hits; 6) merge hits and identify parents; 7) align parents and pseudogenes and check for the presence of disablements (e.g. frameshifts, premature stop codons); 7) assign pseudogene biotype.

RCPedia will 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.

Both Pseudopipe and RCPedia pipelines are broadly used by the pseudogene research community and both are available through our online resource pseudogene.org. In order to mitigate dependency, compatibility, installation and configuration issues, we plan to create docker images for the pipelines and make them publicly available. Docker images will contain all dependencies necessary to set up Pseudopipe and RCPedia. We will also create a amazon machine images (AMI) compatible with Amazon AWS and other major cloud services so users can easily annotate additional genomes

Expect text from Ed Griffiths here.

Clade genomics Toolkit

We will extend the Clade Genomics Toolkit (section xx), integrating enhancements to Augustus (working in conjunction with Mario Stanke, see letter of support) and more robustly combining information from multiple strains. This toolkit will be used to update annotations for mouse strains as user feedback leads to improved heuristics. Early in the grant period, we will be in a position to release per-strain GENCODE gene-sets for each available Mus Musculus genome, including the current 18 sequenced strains, which will be maintained by the Genome Reference Consortium (GRC). The result will be a much more complete, population gene set for Mus Musculus more useful for researchers using non-reference strains (see letter of support from Thomas Keane).

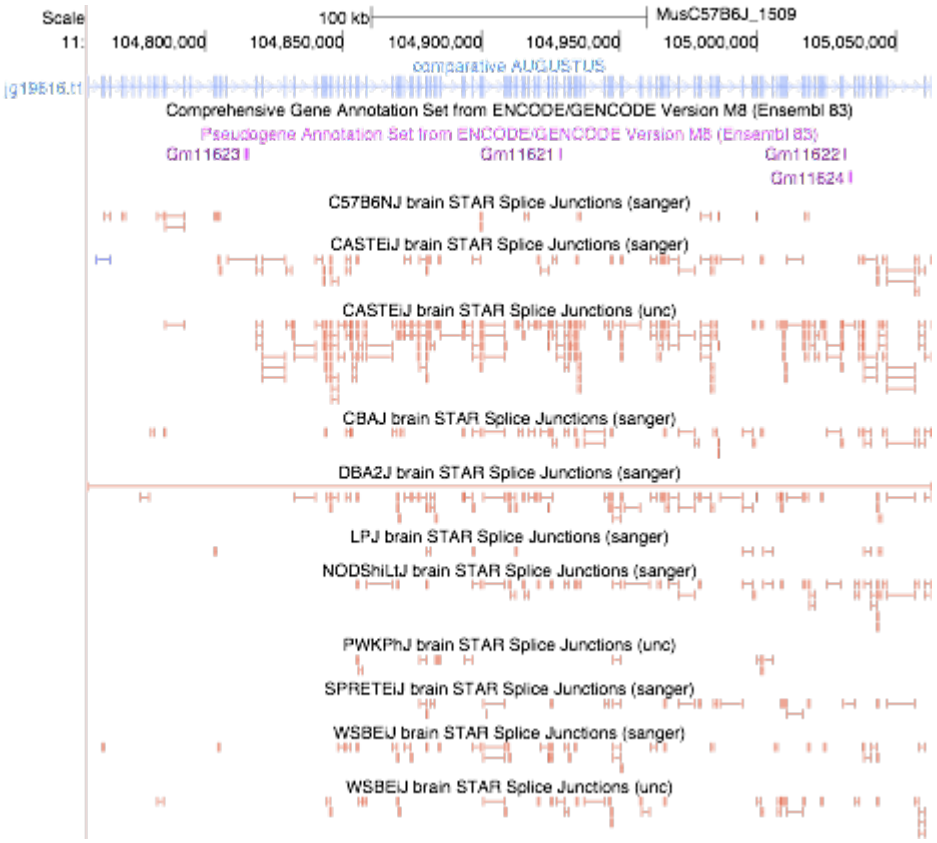


Fig X. New Mus genes. Comparative Augustus identified a 138 exon transcript in a locus not previously annotated. This transcript has varying splice junction support, with the most coming from Mus castaneus. This transcript has been cloned and function is being investigated.

How community annotation will be incorporated (~ 1 page)

Also text from Ed Griffiths here.

Supporting external annotation efforts

GENCODE represents a world-leading infrastructure in the manual annotation of transcripts. Demand for manual annotation of transcripts across strains and species will outstrip our ability to provide such

services. We will support the submission of GFF3/GTF back to a new archive at EMBL-EBI to store external annotation. This archive provides a way 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/authorisation system. Annotation will be QC'd upon submission and then, should it pass the checks, integrated into annotation builds as merge annotation.

GENCODE's primary annotation tools are ZMap and otterlace. ZMap is a GTK desktop application capable of creating new or editing existing annotation. Otterlace is a system for providing primary evidence, tracking models, users and providing a way to navigate target regions to annotate. We will continue to support the development of ZMap and otterlace ensuring the former is capable of running without otterlace and the latter is potentially installable within other compute environments including cloud environments. ZMap will aim to become a standalone tool. It will natively understand CRAM, BigBED, BigWig and GA4GH APIs as these become important to quickly providing new primary evidence and to store and write annotation out to the archives. Basic QC will also be integrated into ZMap allowing for high-quality annotation to be generated upstream of archives. We will distribute GENCODE software through. All software will be released under an Apache 2.0 license. In addition external contributions to the software will be accepted via GitHub's pull request system providing methods to perform code review before acceptance.

Targets

- Development of a transcript submission archive at EMBL-EBI
- QC and integration of external annotation into GENCODE gene sets where suitable to do so
- Developing ZMap into a standalone application capable of reading multiple data sources and writing annotation compatible with archives
- Develop otterlace into a system capable of deployment to cloud environments for non-GENCODE use-cases
- Release code to GitHub under an Apache 2.0 license

Plans for input on user needs (~0.5 pages)

Resource sharing plan (~0.5 pages)

Management, Dissemination and Training: 6 pages

The administrative structure of the project should be described. This section should include the organizational structure and staff responsibilities, progress reporting, and the Scientific Advisory Board (if one is proposed).

Organizational structure and staff responsibilities

The organizational structure of the project should be described. Issues that should be addressed include how the PD/PI and the project staff will be organized with respect to the project activities, how differences of opinion will be resolved, the scientific and technical expertise of the staff who will run the resource development activities, and their distribution of effort across their areas of responsibility.

Scientific Advisory Board (required for complex projects)

The PD/PI should appoint a Scientific Advisory Board (SAB) to advise on progress and priorities of the resource project. In consultation with the SAB, the project must set priorities for the types and depth of information to be included. The SAB should encourage continuous improvements as methods, data, and needs change with time. A strong emphasis on operating in a cost-effective manner should be established. Applicants should describe how they would appoint and use the SAB, and how they plan to organize advisory board meetings and agendas. Applicants should describe previous experiences with advisory panels, how advice was incorporated into a project, and how the advice contributed to a project's outcome.

New applications should not name the proposed SAB members or recruit members to serve on the SAB prior to the peer review of the application. However, they should describe the expertise to be included on the SAB.

Access and dissemination

The access and dissemination activities will vary according to the size and goals of the project. Applicants should include a well-described plan for access to and dissemination of the resource and its contents, consistent with achieving the goals of this program. The resource materials or data should be easily accessible by the scientific community. For some projects, the materials or data can be provided to established resources for distribution. For example, clone sets can be supplied to the appropriate model organism resource centers or commercial distributors, and structural variation data can be provided to the central genetic variation databases. Some projects, such as MODs or other genome informatics resources, will require a separate infrastructure for access and dissemination. In these cases, both the hardware and software components of this infrastructure should be described. The utility of the dissemination activity should be described along with the process for improving the resource in response to community needs and input. Any web-based dissemination activities should emphasize user-centric design.

For data resources, it would be appropriate to employ multiple methods of querying, including simple web interfaces for standard queries and tools such as APIs (application programming interfaces) or web services for more complex queries. The application should provide information about the applicant's experience with building interfaces, and statistics on their use. Applications should describe the plans to make the data, data schema, and tools in the resource downloadable by users.

In all cases, the materials or data should be made available rapidly after verification of their quality.

A robust web presence may be appropriate for informatics resources and some other types of resources. If appropriate, the web site should provide information about:

(1) the project's focus and capabilities, including research objectives if appropriate;

(2) how to interact with the resource and provide feedback;

(3) contact information;

(4) current newsworthy items;

(5) links to online tutorials, if appropriate;

(6) the availability of software, reagents, and other resources, as applicable; and

(7) links to related NIH-funded resources.

Training (if appropriate)

Some projects produce resources that require user training to maximize their utility. Where appropriate, the application must describe a plan and allocate sufficient resources for training both specialists and non-specialists to make the best use of the resource. Examples include presentations, short courses, or symposia offered independently or in conjunction with society meetings attended by the user community; web-based tutorials; and user manuals and training guides to describe the features of the resource.

The project may need to provide user support services with consultation and technical assistance to those using the resource. Applicants should describe their experience in providing user support, evidence of the quality of that service, and the plans to implement or continue this service.

Organizational structure and staff responsibilities

Paul Flicek (Genes, Genomes and Variation cluster, EMBL-EBI) will be the lead PI (Figure 1). He will be assisted in the administrative and reporting aspects of the grant by a Research Management post. The PIs at each of the performance sites will be responsible for project management and reporting for their site. Flicek will be assisted in Project Management at EMBL-EBI by three other EMBL-EBI Team Leaders, Bronwen Aken, Andrew Yates, and Daniel Zerbino: these managers are responsible for different activities within the Ensembl Project (gene annotation, data visualisation and dissemination, infrastructure and regulation, respectively), and are therefore well-placed to ensure that the aims are met.

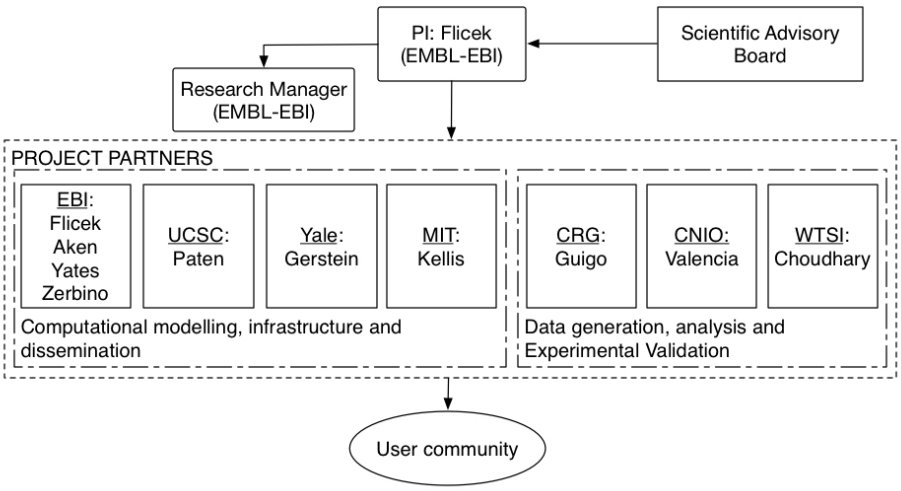


Figure 1: The Project's management structure with principal investigator names for each of the partner institutions.

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:

- Deploying 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 updatesupdate participants on their progress
- 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

We will aim to resolve any differences of opinion informally; if this is not possible, issues will be escalated first to the group of investigators and the overall PI (Flicek), then our SAB and the NHGRI.

the scientific and technical expertise of the staff who will run the resource development activities, and their distribution of effort across their areas of responsibility.

Scientific Advisory Board (required for complex projects)
The PD/PI should appoint a Scientific Advisory Board (SAB) to advise on progress and priorities of the resource project. In consultation with the SAB, the project must set priorities for the types and depth of information to be included. The SAB should encourage continuous improvements as methods, data, and needs change with time. A strong emphasis on operating in a cost-effective manner should be established. Applicants should describe how they would appoint and use the SAB, and how they plan to organize advisory board meetings and agendas. Applicants should describe previous experiences with advisory panels, how advice was incorporated into a project, and how the advice contributed to a project's outcome.

Access and dissemination
It is paramount to the project's impact that GENCODE be made available to as many researchers as possible. This includes making the annotation easily available and consumable but also to make all GENCODE developed software available for offsite use. Access to the GENCODE annotation is primarily though the Ensembl and UCSC Genome Browsers, two of the most widely used resources for genome science. 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. 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.

The UCSC Genome Browser, which receives more than 1.5 million hits per day and has over 150 thousand active individual users per month, serves as an important conduit for distributing GENCODE to a large number of users. With the conversion of the primary UCSC Browser gene set for GRCh38 to GENCODE, UCSC has helped establish GENCODE as one of the two main human gene annotations sets. As mouse GENCODE reaches full genome manual coverage, we will work with the UCSC browser group to switch mouse from the current UCSC genes to the new GENCODE set. This grant will support the interface between the UCSC Browser group and Ensembl, helping to ensure data consistency between the two browsers.

The UCSC GENCODE group directly handles the ingestion of GENCODE data into the UCSC Browser. This frequently involves updating the browser source base to support GENCODE, and avoids the UCSC browser group being a bottleneck in the development process.

To support users who have not migrated to the new human genome assemble, the UCSC and the HAVANA groups developed a methodology for mapping GENCODE from GRCh38 to GRCh37. We will continue to support and enhance this approach and apply it to the next and subsequent versions of the mouse genome assembly.

The UCSC GENCODE group recently computed GTEx expression quantifications of GENCODE genes and isoforms. Working with the UCSC Browser team these are being provided as individual tissue expression profiles for the GENCODE sets. 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 through the UCSC browser to the community.

Aim 4: GENCODE infrastructure, annotation distribution and engagement with community annotation efforts

All software developed by this grant will be released to a public GitHub projects alongside the necessary improvements to support new data types as highlighted in aim 1. External user support will be made available via our RT services and face-to-face workshops and meetings. Primary support will be via our public web interface and the current GENCODE website (http://www.gencodegenes.org/) will be expanded to provide in-depth documentation and details of the processes used within the GENCODE consortium. Frequent data freezes of the annotation sets will be made available via the Ensembl and UCSC browsers. Finally this support will culminate in allowing external third party annotators to contribute back to the manual annotation efforts via submission to EMBL-EBI with attribution back to the contributing author(s) or groups.

4.1: Developing new interfaces for genomic annotation display

A primary reason to annotate strains is to bring with them a way of visually highlighting these differences. Ensembl has a number of static views to view synteny data and viewing smaller regions of differences. Our aim is to provide new dynamic interfaces capable of quickly switching between the various strains and anchor regions. It is then possible to move between a re-designed region comparison view as part using the Genoverse scrollable genome browser. In addition we will also develop Genoverse to highlight/shade out non-dominant isoforms from view or hide them from display. Doing so will require the Ensembl website to allow selection of a panel of tissues and Genoverse to know which isoforms are not dominant by attaching meta-data to the isoforms as they are sent to the client indicating their status. In addition to this we will also alter our tools to further enforce this dominant view. VEP will be altered to prioritise 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.

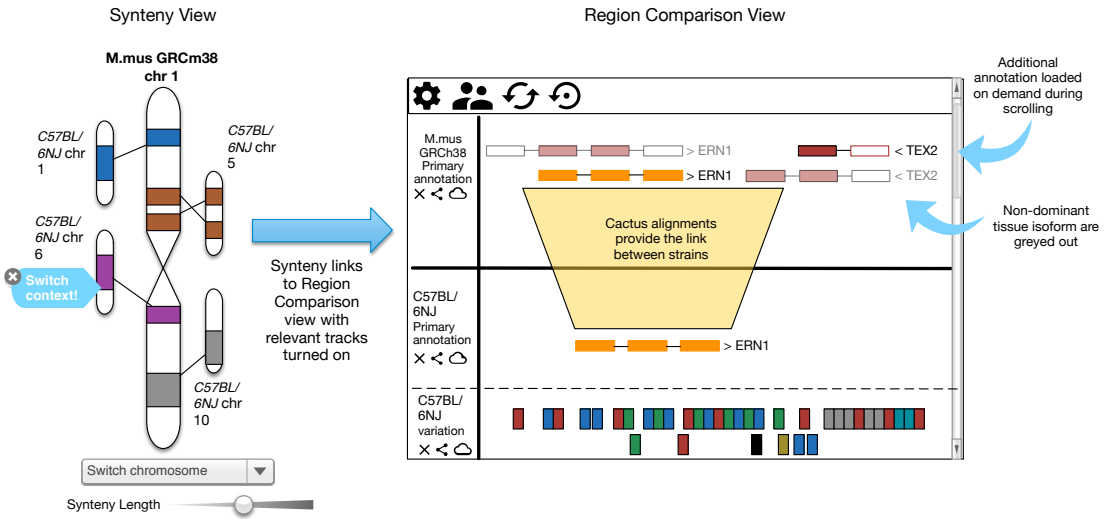


Figure XX: A view of mouse strain supported views moving from high-level synteny views to a configurable client side genome browser. Non-dominant tissue specific isoforms are greyed out.

Targets

- New views to support moving between large scale syntenic regions between strains to viewing comparative models against one and another
- Improved supporting evidence interfaces
- Enable the Ensembl website to promote tissue specific isoforms in views and tools including VEP and sequence search

4.4: Electronic dissemination of annotation

Our goal is to provide GENCODE annotation through as many sustainable and modern distribution methods. This will require the development of new publically accessible APIs and traditional pre-generated flat file freezes of the data sets. 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 to 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. GENOCDE annotation data freezes will be accompanied by unique identifiers based on annotation checksums generated as part of the TGMI award at EMBL-EBI. These will become the global unambiguous identifier for these sets. Change sets will also be released identifying new, retired and modified models. We will maintain the current GENCODE portal site. 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 data sets 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).

Targets

- Disseminate GENCODE annotation and metadata via GA4GH annotation APIs
- Create change sets, to be distributed over HTTP, in both programmatic and human readable formats
- Continue to make flat-file data freezes of annotation available via FTP and HTTP and supplement with tab-delimited and structured metadata
- Distribution of GENCODE annotation tracks using UCSC Track Hubs

Training

4.3: Outreach, training and support

Ensembl and EMBL-EBI already provide an established worldwide outreach program. These activities are primarily in support of the Ensembl WT award. Ensembl hosts over 100 workshops a year across the US, Europe and Asia. The workshops focus not only on describing and teaching GENCODE analysis methods but also using GENCODE data sets for downstream analysis. Here, we will establish a program presenting one workshop per year focused on GENCODE annotation and training. 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. We will host a biyearly workshop on how to use GENCODE annotation tools to annotate genomes and how to submit annotation back to archives. User support will be made available via RT hosted at EMBL-EBI.

Targets

- Hosting a yearly workshop on using GENCODE annotation
- Hosting a bi-yearly workshop on using GENCODE tools
- Continued user support via RT
- Continued focus in Ensembl and EMBL-EBI workshops on how to integrate GENCODE annotation into analysis