Application Table of Contents

Deleted: *7*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: **9**
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 10
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 12
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 13
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *13*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *16*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *17*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *19*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 24
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 24
Mark Gerstein 5/12/2016 7:36 AM
Deleted: **25**
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 26
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 27
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 28
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *28*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *29*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *32*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *33*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 36
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *37*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 37
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 38
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 38
Mark Gerstein 5/12/2016 7:36 AM
Deleted: **39**
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 40
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 41
Mark Gerstein 5/12/2016 7:36 AM
Deleted: 41
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *41*
Mark Gerstein 5/12/2016 7:36 AM
Deleted: *42*
Mark Gerstein 5/12/2016 7:36 AM

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, including pseudogenes. Key well-established technologies for this aim include annotation based on protein, cDNA, EST, RNA-seq and mass spectrometry 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.

Mark Gerstein 5/12/2016 7:09 AM
Formatted: Highlight

Mark Gerstein 5/12/2016 7:09 AM
Deleted: |

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, creating in a effect a personalized GENCODE.

Mark Gerstein 5/12/2016 7:07 AM
Deleted: .

Aim 3: Extend annotation to a definition of the gene that include core regulatory regions and tissue specific enhancers from selected 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 sequencing of the human genome and the resulting reference human assembly is one of the great scientific achievements 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 understanding of a 'normal' genome has emerged and determining the range of normal is potentially an important part of defining what it means to have a genetic disease. Indeed, the variety of the genome has surprised many. We have discovered that structural and copy number variation is pervasive (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

The objective of the GENCODE consortium is to create this foundational reference genome annotation. Our overall goal is to identify and classify all gene features in the human and mouse genomes with high accuracy and based on defined biological evidence, and then to release these annotations for the 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 data types supporting genome annotation 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.

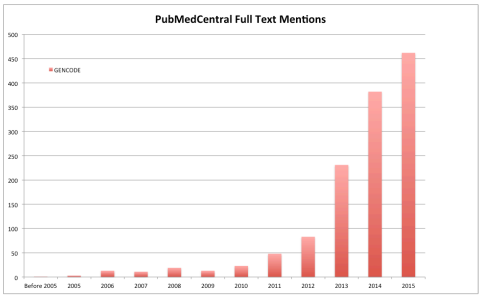

Figure 1: Number of times per year that the text "GENCODE" appears in PubMedCentral (PMC). This is a full text search of only the articles that are in PMC, and thus undercounts usage because only a fraction of papers are contained in PMC. Note that before the GENCODE project, there is apparently only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

GENCODE today

The GENCODE annotation is highly used in both large-scale and small projects. GENCODE is the default human and mouse gene set at Ensembl and the default human gene set for the UCSC Genome Browser (UCSC also provides the mouse GENCODE set and plans to switch to it as default once the mouse annotations are mature). GENCODE is the gene set used for major projects including the Exome Aggregation Consortium (EXAC), GTEx, 1000 Genomes Project, TCGA, ICGC and ENCODE. GENCODE engages directly with the Mouse Genome Informatics (MGI) resource at the Jackson Laboratory and with NCBI as part of the Consensus Coding Sequence (CCDS) project. The International Mouse Phenotyping Consortium (IMPC) uses the mouse gene set arising from this work.

Mark Gerstein 5/12/2016 7:11 AM
Formatted: Highlight

The growth of GENCODE usage has been dramatic over the past four years (Figure 1), and Google scholar has more than 2200 citations for the main GENCODE papers. These numbers undercount the true usage of GENCODE: the uses of GENCODE are not always correctly cited because some users cite the data source as a genome browser (Ensembl or UCSC) instead of the GENCODE project.

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 catalyzed improvements such as the introduction of GENCODE-Basic as a means of distinguishing between the comprehensive set of observed transcript products (including those annotated as incomplete, partially processed or degraded) and the smaller set of representative transcripts. This addressed a concern that the 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 they encode remain incomplete. This insufficiency applies to all classes of genic features including protein-coding genes, pseudogenes, long non-coding RNAs (lncRNAs) and small RNAs. The deficit manifests at multiple levels: complete absence of annotation, partial annotation, 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 occurs where incorrect structural or functional annotation is present. This can be attributable to error, although GENCODE's extensive QC seeks to reduce this to a minimum, or more likely the absence of a required orthogonal dataset at the time of annotation. For example at the time of annotation a locus may be annotated with a protein-coding "biotype", but additional available evidence would clarify the structure and functional potential and allow the biotype to be updated. All these modes apply to the primary genome sequence and patch and haplotype sequences created by the genome reference consortium (GRC), however, genes and transcripts not present

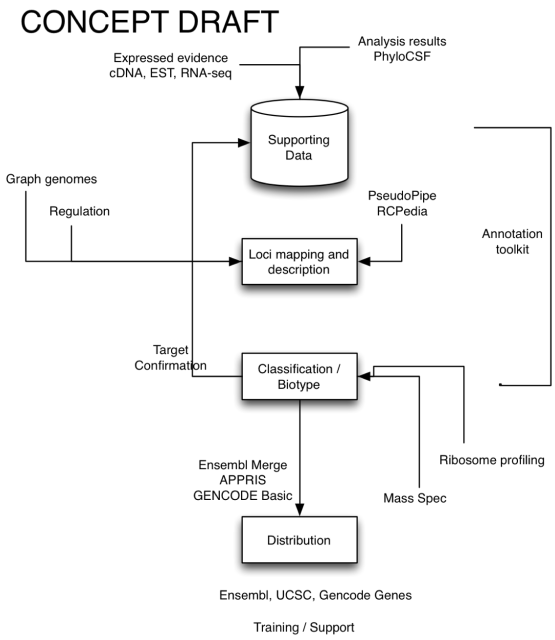

Figure 2: A schematic of the core GENCODE data and analysis flow

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 1

Paul Flicek 5/10/2016 9:21 PM
Comment [1]: At least one paper says that "discovery" GENCODE is better, but expressed concerns that the number of transcripts make RNA-seq analysis more complicated. We have

on these sequences eg in alternative haplotypes will not be captured.

The emergence and improvement of third generation sequencing technologies such as PacBio and Synthetic Long Read RNA-seq (SLRseq) together with the extension of recent techniques based on second generation short-read sequences such as ribosome profiling (Ribo-seq), Cap Analysis of Gene Expression (CAGE), RNA annotation and mapping of promoters for analysis of gene expression (RAMPAGE), polyAseq and mass spectrometry will individually allow us to reduce the degree of error and to target 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. Going forward, it is by building on this established expertise and utilizing multiple orthogonal datasets in combination that we will be able to shrink the gaps in annotation.

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). 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, well established and proven methods and processes are necessary for a comprehensive solution. Computational methods are extremely rapid, consistent and informative, but to date no automatic approach is able to achieve the depth of integration provided by an experienced and trained human annotator, especially in biologically complex regions. That a manual approach must be employed for at least part of the process is hardly surprising: while the practice of medicine has seen tremendous automation over the last half century, a future of automated computer diagnosis for every patient remains distant.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 2

GENCODE has developed since its founding into 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 critical for ensuring that GENCODE remains up-to-date and in line with relevant technologies. Over the course of this proposal we will follow major directions in genomics including graph-based genome representations, long-read transcriptome sequencing, connecting genes and regulatory regions affecting their transcription, and identifying genes that are not present on the current reference assembly. These pilots will determine whether and how each of these technologies contribute to the GENCODE reference annotation and, as appropriate, will be integrated into the core GENCODE processes.

Approach

The GENCODE consortium convened 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 to ensure that the right information is used in the right part of the project at the right time. The individual components and their integrated connections will be leveraged for the continued annotation of human and mouse and extended as appropriate based on the outcomes of the pilot projects. For all activities the focus and overall goal of GENCODE is the annotation of all evidence-based gene features at high accuracy.

Over the last four years, GENCODE completed a full first pass manual annotation of the human genome, conducted extensive QC on the annotation, and investigated promising 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 utilize them optimally and integrate them

into our workflows to identify missing features and improve and update the existing annotation. Combining multiple novel datasets will allow us to formalize our guidelines for edge-case resolution, while the large volumes of new data with direct relevance to gene annotation will require our continued development of methods to provide 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, mouse will benefit from further manual annotation as part of our continuing annotation effort. We will continue traditional manual annotation for mouse, both chromosome-by-chromosome and from targeted lists of genes and gene families, to ensure consistency with human annotation and support comparative analysis between the two species. However, we will also be able to rapidly adopt the updated methods piloted in human in order to retain as similar standards of annotation as possible for the two genomes, given the likely differences in the experimental datasets that are produced for them. In particular, human has much more experimental data, but mouse has access to tissues and developmental datasets 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 in 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 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 GENCODE genesets for human and mouse comprise a core of manual annotation for protein-coding, long non-coding RNA and pseudogene loci. These are supplemented by Ensembl annotation for small ncRNA genes, novel transcripts, and mouse genes in regions that are not yet manually annotated. Experienced human annotators using the Zmap/otter suite of annotation software define transcript structure and function by integrating a large number of orthogonal data types, computational predictions of genic features and literature. Transcript structures are predominantly determined based on refined alignments of transcriptomic data generated by first, second and third generation sequencing technologies. Transcription start sites are identified using CAGE and RAMPAGE data, while polyAseq perfoms the same role for transcription termination sites. The protein-coding potential of transcripts is investigated using protein homologies from reference databases, cross-species conservation as defined by PhyloCSF and PhastCons, and alignment of mass spectrometry and ribosome profiling data. Loci where no transcripts show evidence of protein-coding potential are classified as long non-coding RNAs. Pseudogene transcript models are annotated based on support from protein homologies and the identification of disabling mutation or retrotransposition and lack of locus-specific transcription. Annotators are directed towards regions of likely significance by computational approaches that report new or inconsistent data. These include Ensembl, PhyloCSF regions of conservation with protein-coding characteristics, and predictions of pseudogenes and retrotransposed loci by PseudoPipe, RCPedia and Retrofinder. Where a data type is not available to be displayed in Zmap, annotators access via UCSC and Ensembl browsers or custom browsers such as Zenbu.

Integrated approach to pseudogene identification and classification

The Yale group has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and HAVANA group, Yale 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 main 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 pseudogene variability across human individuals and, in doing so, help to understand the boundary between protein-coding genes and pseudogenes.

Computational methods to evaluate and enhance gene annotation

The Ensembl GeneBuild provides an automated, independent method to identify and annotate all genes including protein-coding genes, small and long non-coding RNA genes, and pseudogenes. Ensembl gene annotation has a reputation for high quality, as judged by community assessments of computational annotation methods, 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 non-coding 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 access human tissues and for regions of the mouse genome that have not yet had comprehensive manual annotation.

Mark Gerstein 5/12/2016 7:16 AM
Comment [2]: MG: link?

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 provides 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 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.

Bronwen Aken 5/10/2016 9:32 PM
Comment [3]: PMID: 21685081

Bronwen Aken 5/10/2016 9:32 PM
Comment [4]: PMID: 23161672 and others

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 region. CLS will be deployed for a series of complex tissues in both adult and embryonic time points. In addition, we will use mass spectrometry 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)

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

The GENCODE consortium has assembled a team of world experts in a variety of fields related to gene annotation to create and distribute this gold standard. We have been collaborating 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.

Since April 2013 we have made eight GENCODE releases for both human (V17 to V24) and mouse (M1-M8) Human updates have focused on lists of features while chromosomes lacking comprehensive manual annotation were targeted in mouse, with an emphasis on extending the annotation of pseudogene and lncRNAs. The number of human protein coding genes decreased significantly from V19 to V21 due to reanalysis done by the CNIO group in collaboration with the manual annotation group. However, more recently this trend has reversed following reannotation of putative coding features generated by PhyloCSF and aided by the CodAlignView tool. The review of changes to the human and mouse protein coding set continues to utilise the forum provided by CCDS collaboration between WTSI, EBI, HGNC, MGI and RefSeq. Around 50 protein coding genes are re-examined by the teams each month and discussed.

Manual annotation has been enhanced by integration of novel data types such as polyAseq, FANTOM CAGE and ribosome profiling (RP) to correctly identify transcription start sites, re-evaluate 3' UTR extensions and investigate translation. ENCODE 454 and PacBio data from the Snyder and Gingeras labs are beginning to be incorporated to improve the annotation of lncRNAs. Integrating mass spectroscopy (MS) data, we have identified highly reliable peptides for 64% of (non-read through) protein coding genes. The Choudhray group reprocessed three large-scale publicly available human proteomics datasets, made up of over 54 million mass spectra. Such evidence is frequently difficult to interpret and, contrary to other studies using the same data, we found evidence to confidently support the addition of only 16 novel proteins to GENCODE, although the same dataset supports the annotation of alternative splicing in 867 genes. The human MS data publicly available, only validates 10 novel peptides in 15000 lncRNAs, and thus did not validate reported ORFs previously described in other publications (PMID:26687005, PMID:23044541).

Pseudogene annotation in GENCODE draws on more than 15 years experience annotating and reviewing 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}. These techniques have generated within GENCODE 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 the most recent and up-to-date publication, Yale identified 14,505 pseudogenes in human, 911 in worm, and 145 in fly[deemphl] \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

Mark Gerstein 5/12/2016 7:19 AM
Formatted: Highlight

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}.

As part of the QC process of the entire geneset, a pipeline has been developed at UCSC to determine introns with low support including those that may be problematic due to alignment issues and those that have non-canonical splice sites. Intron support data is passed to the annotators for manual inspection and updates or removal of transcript models as appropriate. Together with Ensembl, UCSC has produced a set of transcript support levels (TSL) based on whole transcript support from sequences from the International Nucleotide Sequence Database Collaboration (INSDC). Both BLAT and Ensembl Exonerate alignments of the mRNAs and ESTs are utilized and annotations are assigned one of five levels, where level 5 is the worst and indicates no support. Annotations in the MHC regions and other immunological genes are not assessed since automatic alignments in these regions tend to be problematic. These levels can be highlighted in the UCSC Genome Browser and used for filtering the annotation. Ensembl and HAVANA teams collaborate to improve the quality of the GENCODE human and mouse gene sets, particularly by identifying and removing redundant or low quality models generated through automatic annotation. Other improvements include updates to the cross-referencing system to increase accuracy, adding functionality to BioMart so that APPRIS data can be queried, and a simplification of sequence ontology terms linked to each gene/transcript.

Following the release of GRCh38, GENCODE V20 was mapped to the new reference genome; 261 alternative loci were also manually annotated, including 35 LRC haplotype regions. Ensembl produced automated, genome-wide gene annotation on the new human assembly GRCh38 and for every release, Ensembl automated annotation continues to be merged with the manual annotation to produce the GENCODE geneset. Small noncoding RNA genes are produced solely by Ensembl for both human and mouse. From GENCODE V21 and M3 onwards support for GENCODE Basic was added to Ensembl, as were Transcript Support Levels and APPRIS tags. From V22, the UCSC Browser adopted GENCODE as their primary gene set, replacing ‘UCSC genes’; the GENCODE version will be updated periodically. To address the slow migration from GRCh37 to GRCh38 and provide users still using GRCh37 access to improvements in the GENCODE gene annotation, UCSC and HAVANA developed a methodology for mapping the GENCODE geneset to previous assemblies. We have made releases of GENCODE V23 and V24 mapped to GRCh37.

To systematically assess the quality of the GENCODE gene set, we experimentally verify the structure of all transcripts rated as novel or putative using RT-PCR-Seq. We tested 1,243 exon-exon junctions not supported by ENCODE or GTEx RNA-seq data, confirming support for 53%. In an equivalent analysis for 3,148 exon-exon junctions in mouse we confirmed ~49% of targeted exon-exon structures. To assess the completeness of lncRNA annotation using 3' and 5' RACE followed by 454 sequencing (RACE-seq), we initiated a pilot experiment targeting 400 GENCODE-annotated lncRNA loci lacking CAGE and GIS-PET support for 5’ and 3' ends in 7 different human tissues. The RACE-seq reads generated were manually inspected leading to the addition of ~2,600 previously unknown AS transcripts. Approximately 48% of the 5'-extended transcripts overlap a CAGE cluster, and 51% of 3’ extended transcripts contain polyA sites, suggesting RACE-seq is successful in identifying TSS and TTS. To extend of this workflow we used RNA Capture-LongSeq, a methodology based on RNA capture coupled with PacBio long read sequencing. We targeted the entire set of GENCODE annotated

long noncoding RNAs in human and mouse. A substantial number of other genomic elements were also probed, including other ncRNA classes (miRNAs, snoRNAs and snRNAs), enhancer elements and ultraconserved elements. The human design targets ~15,000 features, of which ~10,000 are intergenic lncRNAs. The mouse design targets ~9000, of which ~5000 are lncRNAs. We captured RNA from 4 common tissues in both species: brain, heart, testis and liver, two ENCODE cell lines (HeLa and K562), and two mouse fetal time points (embryo 7d and 15d). We obtained ~2 million Circular Consensus (CCS) pacbio reads with length up to 4 kb for each species, leading to the discovery of almost 100,000 completely novel, high-quality canonical introns in human, and more than 50,000 in mouse, increasing the number of splice junctions in each genome by ~27% and ~20%, respectively. Analysis of polyA sites revealed that ~60% of CCS reads are polyadenylated, the majority located more than 50 nucleotides away from any annotated transcript 3' end and are often preceded by strong polyA signals. In addition, we discovered in each species ~10,000 previously novel TSSs located proximal to FANTOM5 CAGE clusters. We are also working to complete the of annotation of alternative splicing at coding loci, using 3' and 5' RACE followed by PacBio long read sequencing (RACE-seq). Preliminary analysis of a pilot experiment targeting 541 UKGTN loci human testis and brain showed presence of approximately 8% novel canonical splice junctions in targeted genes. Figure 3, shows annotation completed since the start of the grant against our milestones, categorized by protein-coding, lncRNA or pseudogene biotype.

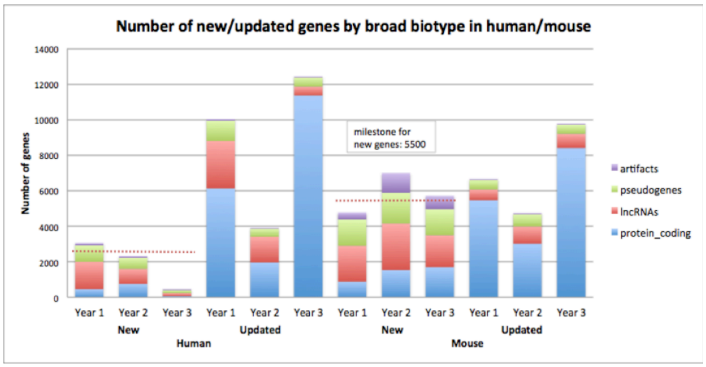

Figure 3: Annotation completed since the start of the grant against our milestones, categorised by biotype.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 3

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

As just described, 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 committed to supporting the genomes of a collection of 16 representative strains, thus effectively replacing the linear genome with a “graph-like” structure of 16 separate haplotypes. 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 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 group https://www.sanger.ac.uk/research/projects/vertebrategenome//havana/assets/guidelines.pdf).
Historically the HAVANA group produced transcript models largely based on the alignment of EST and cDNA sequences from the INSDC and protein sequence data from UniProt. These sequences were aligned to the individual BAC clones that make up the reference genome sequence using BLAST (PMID: 2231712), with a subsequent splice-aware realignment of transcriptomic data by Est2Genome (ref). The core GENCODE process (Figure 2) builds on this established method, and 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 viewed in the ZMap annotation interface (http://www.sanger.ac.uk/science/tools/zmap). Depending on the species and the locus there may be more than 400 datasets available to manual annotators including: gene and pseudogene predictions (including pseudogene predictions from the PseudoPipe (PMID: 16574694) and Retrofinder (PMID: 18842134) pipelines produced by GENCODE consortium members), cross-species conservation, transcription start and termination sites, regulatory features, mass spectrometry and Ribo-seq data, second (i.e. Illumina) and third (i.e. PacBio) generation RNA-seq data and transcript models and splicing feature predicted from them. Although the output of GENCODE is the final set of annotations, this collection of supporting evidence represents the full breadth of the data currently available within the GENCODE resource.

Bronwen Aken 5/10/2016 9:35 PM
Comment [5]: Could be either Exonerate or Richard Mott at PMID: 9283765
Not sure – check with Adam

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 2

Determining biotype GENCODE genes are assigned a “biotype” associated with one of four broad categories; protein-coding gene, lncRNA gene, small ncRNA gene or pseudogene. Genes derive their biotype from the biotypes assigned to their constituent transcripts. 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. Polymorphic genes are also labeled. All newly created transcripts and loci are initially assessed to determine their protein-coding potential and the assignment of a non protein-coding biotype is only made when the possibility of coding potential is eliminated. Protein-coding potential of transcripts is determined on the basis of similarity to known protein sequences, the sequences of orthologous and paralogous proteins, the presence of Pfam functional domains (PMID: 26673716), clear support of high quality peptides from mass spectrometry (MS) experiments and good evidence of translation from ribosome profiling (Ribo-seq) data. Deep conservation of a coding sequence (CDS) and unambiguous published evidence can be used to support manual annotation of a protein-coding transcript, even in the absence of any other support. Where a locus shares homology with UniProt accessions but its putative CDS is disrupted by in-frame stop codons, frameshifts, insertions or deletions it is assigned the pseudogene biotype. The broad pseudogene biotype definition has multiple subdivisions that describe the mechanism of creation and transcriptional status of the locus. 'Processed' pseudogenes are generated via retrotransposition events while 'unprocessed' pseudogenes are created by duplication. Unitary_pseudogenes are identified by the presence of an unambiguous functional ortholog in another species. Where a disabling mutation is not fixed i.e. it is polymorphic or segregating, the locus is annotated as a polymorphic pseudogene. These occur when a gene is protein-coding in some populations and 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 described above. Given our current inability to infer the functional potential or mode of action for most lncRNA loci, biotypes are assigned on the basis of genomic position relative to protein-coding loci. For example, antisense transcripts overlap the genomic span of a protein-coding locus on the opposite strand, and lincRNA transcripts are intergenic to protein-coding loci. The lncRNA annotation produced by GENCODE represent a core dataset underpinning the RNA Central lncRNA dataset (PMID: 25352543).

Mark Gerstein 5/12/2016 7:26 AM
Formatted: Highlight

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 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.

Updates to Annotation: Missing loci It is clearly one of the key objectives of GENCODE to represent all gene loci in human and mouse. The main data types we use to identify unannotated loci are transcriptomic data, and proteomics data. Transcribed loci are identified by Ensembl, ENCODE collaborators and good quality public methods such as PLAR (PMID: 25959816) using transcriptome data, and become candidates for targeted manual annotation. These loci are manually annotated when the RNA-seq based transcripts have ≥1 intron intersecting with an unannotated intron in splicing ESTs, cDNAs or third generation RNA-seq transcripts, or when introns from ≥2 independently created RNA-seq based transcripts overlap. We will continue to investigate transcription identified in RNA-seq datasets from previously inaccessible tissues and development stages, public third generation transcriptional evidence such as SLRseq (PMID: 25985263) and PacBio and Capture-Seq PacBio data.

Protein-coding loci are identified by analysis of reprocessed MS data from large-scale public shotgun proteomics datasets. In human this has led to the identification of ~20 novel protein-coding loci (REF). Mouse shotgun proteomic datasets are significantly smaller than human but the availability of samples in tissues and developmental stages inaccessible in human may support identification of novel loci. In addition to the MS data, we use cross-species conservation information from PhastCons (PMID: 16024819) and specific conservation of protein-coding sequence from PhyloCSF (PMID: 21685081) to

Bronwen Aken 5/10/2016 9:39 PM
Comment [6]: There might be a newer publication than this one PMID: 21460061

identify novel gene loci. For example detailed investigation of a refined set of thousands of high scoring PhyloCSF regions generated across the whole human genome yielded more than 100 novel protein-coding loci, many of which had very low expression support in human RNA-seq datasets. We are currently investigating the equivalent dataset in mouse. Although many of the novel loci have orthologs in both species, we have identified multiple loci where a gene has been lost in one lineage, emphasizing the importance of independent analysis in both species. PhyloCSF frequently highlights unannotated pseudogene loci. More than 200 unitary and unprocessed pseudogenes have been identified suggesting that there remain many unannotated pseudogenes in both human and mouse genomes. We will continue to use large public proteomics datasets and cross-species conservation information to identify putative novel protein-coding loci.

Partial annotation and underannotation Adding annotation of missing alternate splicing (AS) transcripts at annotated loci and extending partial AS transcripts to reflect their full length remains an important goal of the GENCODE project. Adding missing exons and splice junctions is essential in providing the best possible foundation for downstream analysis and interpretation. Even where all exonic sequence is annotated, the accurate extension of all transcripts to full-length is essential to describe the connectivity. It is also vital to extend all transcripts to full length to allow the proper interpretation of the functional potential of a transcript and promoter. Where a partial AS transcript cannot be annotated as protein-coding or NMD with certainty, it is annotated with an agnostic 'processed transcript' biotype, and manual annotation will be revisited when more data become available. In GENCODE 24 there are more than 10,000 processed transcripts annotated at protein-coding genes and a further 33,000 partial protein-coding transcripts tagged as incomplete (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. Recent reannotation of 70 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 was novel CDS.

Bronwen Aken 5/10/2016 9:42 PM
Comment [7]: Cumbersome sentence, what are we trying to say here?

It is also essential to extend transcripts of all biotypes to full-length, including transcripts with NMD and retained_intron biotypes. While both have historically been regarded as reflecting missplicing events and poor RNA preparation, 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 (REF). Extending transcripts at lncRNA loci allows additional exons to be identified, and allows a more informed determination of transcript and locus biotype, for example confirming that currently annotated lincRNAs are fully intergenic.

Missing and partial AS transcripts will primarily be detected and extended using third generation transcriptomic data, SLRseq and particularly PacBio and CaptureSeq. 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 identified 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 third generation transcriptomic data, we find that second generation RNA-seq, with its wider pool of cell-lines, tissues and developmental stages remains useful in identifying and annotating novel splicing features, particularly in combination with PhyloCSF, CAGE/RAMPAGE and polyAseq data. CAGE from the

ENCODE(REF) and Fantom(REF) Consortia and RAMPAGE data define the position of the transcription start site. PolyAseq data (REF) defines the polyA site i.e. the end of a transcript. Full-length transcripts combined with knowledge of transcript start sites (TSS) and transcript termination sites (TTS) give all the required information to make a determination of biotype, specifically certainty over the TSS allows the translation initiation site to be determined and identification of the TTS provides important context for the position of the stop codon and whether any post-transcriptional control (PTC) would be likely to trigger nonsense-mediate decay (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 definition of the GENCODE Basic set, which contains only full-length transcripts, will become redundant. 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, Ribo-seq, 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.

Integrated approach to pseudogene identification and classification (~1.5 pages)

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

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 retrotransposed or duplicated disabled copies of functional paralogs (referred to as pseudogene parents) . There is a good consensus overlap between the human pseudogene prediction set obtained with PseudoPipe and the set manually curated by the GENCODE annotators \cite{22951037}. Even more, the PseudoPipe predictions fueled the manual curation of pseudogenes in GENCODE \cite{22951037}.

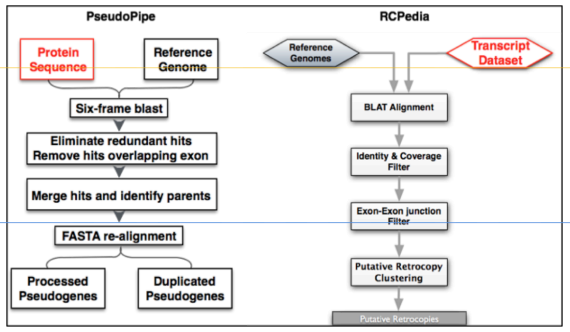

Figure 4: Automatic pseudogene annotation pipelines

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

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 4

Bronwen Aken 5/10/2016 9:48 PM
Comment [8]: Duplication of text in previous section. Do we want this again?

Sisu, Cristina 5/11/2016 12:07 PM
Comment [9]: I've removed this sentence but I just kept the part saying that the pipeline is able to distinguish the pseudogene biotypes

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 4

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.

Yale 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 \cite{24026178,26432246, 20210993}. 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, Yale 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 of pseudogenes 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}.

Mark Gerstein 5/12/2016 7:27 AM
Formatted: Highlight

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 (Aken et al 2016). The GeneBuild process automates the decision-making steps followed by manual curators, 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 non-coding RNA genes, and provides rapid access to novel transcript isoforms that are identified from new data in the archives. Regions of the genome that are biologically complex (i.e. immunoglobulins, major histocompatibility complex) or where input data are inconsistent are annotated manually. The GeneBuild is improved by adding new data types, and by feedback from the manual annotation team.

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 data in the public archives are re-aligned. This is a considerable undertaking taking several months to complete. Minor assembly updates, such as the update from GRCh38.p7 to GRCh38.p8, simply add additional alternate sequence alongside the primary assembly. These minor updates 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.

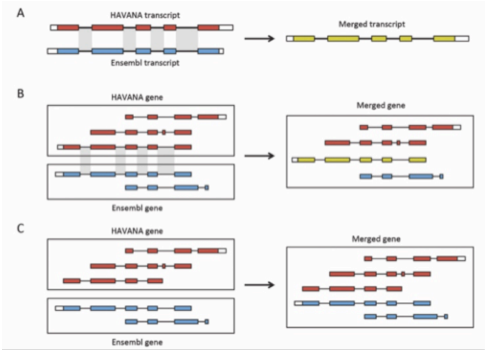

Figure 5: Annotation of patches. For both Ensembl and HAVANA models, transcripts with overlapping exons are grouped together into genes. A: If the intron-exon boundaries, excluding UTRs, of a transcript from HAVANA completely match those one from Ensembl the result is a merged transcript model, which is always based on the HAVANA annotation. If the intron-exon boundaries do not completely match then the two models are treated as separate transcripts belonging to the same gene. B: Exons for a HAVANA gene overlap with those for an Ensembl gene. All transcripts are grouped together in the same merged gene. The intron-exon boundaries for one HAVANA and one Ensembl transcript match perfectly so they are merged to create the merged transcript show in yellow. C: Exons for Ensembl and HAVANA transcripts overlap but there are no transcripts with complete matching intron-exon boundaries. We still group the transcripts together into a merged gene but no transcripts are merged.

Ensembl/HAVANA Merge. The comprehensive GENCODE gene set is created by including the entire annotation from HAVANA and selectively 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. This merge process, described in detail by Harrow et al. (2012), is performed genome-wide and involves pre-merge quality checks and comparison of all Ensembl transcripts against all overlapping HAVANA transcripts (Figure 5). Where the splicing structure of the Ensembl transcript matches the HAVANA transcript, they are merged and the alignments supporting the Ensembl annotation are combined with the HAVANA data. Novel genes and transcripts contributed by Ensembl are added. The HAVANA biotype takes precedence where data are inconsistent.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 5

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 For every update to the GENCODE gene set, a subset of representative transcripts are identified for each gene and labeled as 'GENCODE Basic'. This selection process prioritizes full-length protein-coding transcripts over partial or non-protein-coding transcripts within the same gene, thus highlighting those transcripts that are most useful in the majority of applications.

Add phyloCSF section here

Paul Flicek 5/5/2016 11:29 AM
Comment [10]: MIT to add. No more than approximately 0.5 pages.

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 annotates the likely effects of alternative splicing on protein features, and selects a single CDS as the main (principal) isoform based on these annotations (Tress 2008 PMID: 18006548). The annotations have an important quality control role and are at the core of the various CNIO quality

control pipelines.

The APPRIS principal isoform is generally the isoform with the most conserved protein features and the most evidence of cross-species conservation. APPRIS principal isoforms coincide overwhelmingly with the main protein isoform detected in proteomics experiments [Ezkurdia 2015 PMID: 25732134]. APPRIS selects a principal isoform for 73.4% of human genes and 82.1% of mouse genes. For genes in which APPRIS cannot choose a main variant, it selects the main isoform based on CCDS annotations [Farrell 2014 PMID: 24217909] and UCSC Transcript Support Level. Using information from these two methods APPRIS is able to select a principal isoform for 95.5% of human genes and 96.4% of mouse genes.

The APPRIS database [Rodriguez 2013 PMID: 23161672] houses annotations for seven Ensembl species including human and mouse. It is stable and is implemented as part of the GENCODE/Ensembl human genome annotation, and can be visualized in the UCSC Genome Browser as public track hub.

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. Nineteen features that correlated with lack of protein-level expression, including poor conservation, recent origin, poor supporting evidence and contradictory annotations were used to flag 2,001 coding genes from GENCODE v12 as potentially not coding [Ezkurdia 2012 PMID: 24939910]. Manual annotators have since revisited these genes and 1,026 have been reclassified as either non-coding or pseudogene.

The CNIO carried out a similar analysis on coding genes added between GENCODE v12 and v19, leading to the reclassification of almost 500 automatically predicted coding genes. In the most recent analysis (GENCODE v23) a further 2,050 protein-coding genes were labeled as unusual. The pipeline has also been applied to the mouse annotation and the initial analysis identified 4,841 mouse protein-coding genes that were potentially not coding.

The CNIO plans to automate the identification of these unusual coding genes to allow the pipeline to be run for each new release of GENCODE. This ought to be most useful for mouse , 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 Non-coding 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).

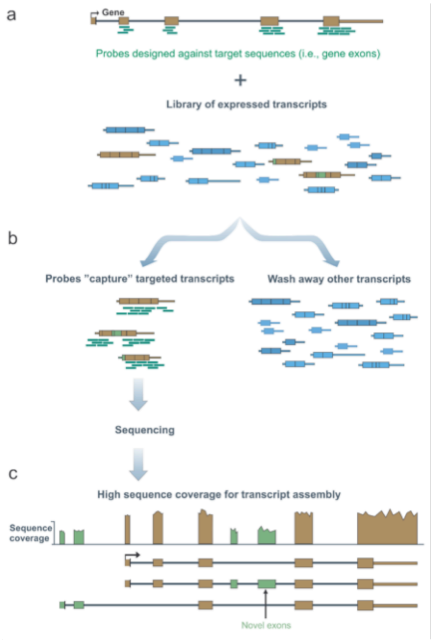

Figure 6: Seq

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.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 5

Paul Flicek 5/5/2016 12:45 PM
Comment [11]: CRG: Needs to be rationalized with other comments in application.

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) [Abascal 2015 PMID 26061177; Ezkurdia 2015 PMID 26496066].

The 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.

Reliable proteomics data identifies only a fraction of annotated alternative isoforms. Our proteomics analysis pipeline detected evidence of translated alternative splice variants in just 246 genes in the GENCODE v20 reference set [Abascal 2015 PMID: 26061177]. According to simulations, this number is considerably smaller than expected and strongly suggests that most protein-coding genes have a single main protein isoform [Ezkurdia 2015 PMID: 25732134]. This result contrasts with the abundance of alternative transcripts in microarray and RNA-seq experiments and is especially surprising since the CNIO protocol interrogated more than 100 different tissues, cell lines and developmental stages [Ezkurdia 2015 PMID: 25732134]. Those alternative splice events supported by proteomics evidence were significantly enriched in subtle splice events that did not disrupt Pfam functional domains and many of them were among the oldest splice events annotated.

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.

Protein samples will be processed according to established shotgun proteomics protocols and workflow and analysed by high resolution tandem mass spectrometry. Isotope labelling using tandem mass tags (TMT) that enables multiplexing of ten samples per run and quantitation using MS3 workflow on an tribrid orbitrap instrument will also be used to improve protein quantitation and sequence coverage. Using an off-line peptide fractionation method based on high pH reverse phase chromatography, we routinely generate quantitation of >10000 proteins per proteome. For quantitation, each identified gene a normalized spectral count value is calculated in each one of the 10-plex experiments by dividing the number of peptide spectrum matches (PSMs) of each protein with the total number of PSMs. Median normalized spectral counts per gene are computed across the different multiplex experiments. To attain deep sequencing and accurate quantitation we will apply approaches for sample fractionation, together with multiple enzymatic digestions for peptide generation and repeat mass spectrometry analysis. We will use synthetic peptides (see section xx) as reference standards to align and compare between sample sets as well as to target specific gene features. Spectrum will be assigned using our proteogenomics pipelines.

Proteogenomics Pipeline The Sanger proteomics group have developed a sophisticated and robust open access workflow for the analysis of large scale proteomics dataset (Wright et al. accepted Nature Communications 2016). Processing of 52 million spectra from the draft human proteomes published in Nature [Wilhelm 2014 PMID: 24870543, Kim 2014 PMID: 24870542] has led to evidence supporting genes that have been added to GENCODE as confident novel protein-coding genes; 8 were previously annotated as pseudogenes, while 8 were lncRNAs. We have also evidence for 867 alternatively-spliced genes which is considerably higher than other analyses. The pipeline is now being applied to other human tissue datasets

including the recent Human Proteome Project datasets [PMID: 26337862], and publically available data, such as the mouse tissue map [PMID: 23436904], as well as in house collected data.

There are several unique components to our workflow. Firstly, construction of a bespoke search database that encompasses the current known proteome of the sample species, 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. In addition, the sequences in the database are clustered as described by Nesvizhskii et al. [PMID: 25357241] improving protein inference. Our proteogenomics pipeline has been constructed around the open source OpenMS workflow platform [paper in prepartion] making use of the Sangers high performance computing clusters HPC for efficient large-scale analysis. We have also developed and in-house spectral clustering tool, MSSMIV (paper in preparation), which examines mass differences to identify spectra originating from peptides with modifications and amino acid substitutions. We ensure only proteotypic non-shared peptides are used for protein inference to avoid ambiguity in identifications. 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, UCSC or Biodalliance. The BED format is further converted into a bigBED format and used to create a proteomics track-hub. We are currently working on Gtf. Format for integration to IGV (paper in preparation).

Novel Peptide Analysis 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 peptide features that are not normally considered in PSM scoring by our novel priority annotation score [Wright et al, accepted Nature Communications) and 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. Our pipeline also implements multiple different quantification tools, which can be applied depending on the sample. The quantified genes and transcripts reported in the experiment and further submitted to ExpresionAtlas, all the data is deposited in PRIDE and ProteomeXchange repositories, enabling browsing and comparison of the quantified genes / transcripts across the multiple tissues and experimental samples.

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)

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 characterize the expression of rare isoforms over the course of the project period.

Paul Flicek 5/9/2016 3:35 PM
Comment [12]: Add justification why

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)

Structural misannotation i.e. inclusion of unsupported introns or exons is assessed by confirming the presence of the features in transcriptomic datasets. Initially we have used alignments of ESTs and cDNAs to check the complete structure of transcripts, and indicate the level of support in the transcript support level (TSL). Manual rechecking of ~10000 introns flagged as unsupported by EST and cDNA evidence suggest that with the exception of introns with non-canonical splicing, QC is very sensitive. More recently we have used RNA-seq data from multiple tissues produced by the Gtex project to identify introns with little or no support. We will tag transcripts containing such introns and remove them from the geneset. We will use third 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 is assessedin three ways. Biotypes of all transcripts are compared to other transcripts at the same locus to identify aberrant combinations. To identify unannotated protein-coding transcripts, PhyloCSF regions and proteomics data intersecting with lncRNA annotation are checked, as part of CCDS discussions or following literature review. To confirm true protein-coding functionality of annotated transcripts we utilise coverage by, shotgun and targeted proteomics and RP data. Quantitative proteomics will also give insight into the likely functional significance of loci that fulfil the essential criteria to be annotated as protein-coding ie have an intact CDS and evidence of transcription. Quantifying the abundance of a protein, 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.

Bronwen Aken 5/3/2016 5:14 PM

Comment [13]: Note that CCDS is talking about expanding inot the UTRs. This is actually something that TGMI would like. HAVANA spend time on CCDS and this is important to GENCODE as CCDS is used in the clinical community

Validating novel coding genes Proteogenomics, searching against databases of translated non-coding regions, is a growing field (Brosch 2011 PMID: 21460061, Gascoigne 2012 PMID: 23044541, Kumar 2016 PMID: 26773550). It is most effective when used with poorly annotated species [Nesvizhskii 2014 PMID: 25357241] because the large non-coding databases used to search against non-coding regions will affect the false discovery rate calculations and mean that many spectra will be falsely assigned to non-coding regions [Nesvizhskii 2014 PMID: 25357241].

The CNIO methodology for validating novel coding genes is centred on searching against known coding databases (UniProt, IPI, RefSeq) since we know that the vast majority of peptides identified reliably will map to conserved protein-coding genes [Ezkurdia 2014 PMID: 24939910]. These databases are much more likely to contain missing coding genes, although comparisons between Ensembl/GENCODE and the other databases suggest that few coding genes are missing from the human reference set. This protocol generates smaller numbers of novel coding genes but has a higher hit rate and allows us to use careful manual inspection to separate spectra that identify novel coding regions from false positive matches. This has proved a very profitable strategy; 61 missing coding genes have been added to the reference set from the CNIO proteomics analysis and a further 16 coding regions were found in conjunction with the Sanger group [Wright 2016 submitted].

Validating gene models 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 the genes had the same 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 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. All the main annotation tools (ZMap, otter and Blixem) are held in source code control repositories from which all development takes place using recognized code branching methodology for controlled introduction, release or if necessary rollback of new features and bug fixes.

Core annotators use tested releases of the software on Unix/linux or Mac OS X machines.

Annotation database access is controlled via a secure password based system with write access being controlled separately. An edit trail is kept for all database edits allowing roll back in case of error.

The databases are accessed via an http server whether the annotator is working externally or internally to the institute. All databases created by the otter pipeline code are duplicated across machines/machine rooms to provide a high level of fail safe redundancy. This is important because of the large volume of data involved.

Update cycle and process. Software **r**eleases are made every month via an anonymous automated build and unit test system to ensure robustness and independence of release production. Initial automatic testing is followed by full interface testing by the developers following scenario's developed to test all major aspects of the user interface and this is then followed with testing by experienced annotators for 2 weeks prior to release. The automated build system also automatically uploads all builds to the institute's public ftp site. In addition the same system is used to run nightly builds of the latest code to catch development problems on a daily basis. The build system also produces install packages for Unix (via autoconf) and Mac OS X (via a dmg package), MS Windows users are supported via a bootable virtual machine.

Annotation database updates are made as new data comes in and more rarely as the reference assembly changes. These updates are made in a carefully planned way so that normal annotation can continue as species are updated one by one.

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

Increasing computational support of manual gene annotation: In order to increase throughput when adding novel transcripts supported by RNAseq data we will create a GENCODE transcript based on the most representative RNA-seq-supported transcript (i.e. the one that contains the most introns supported in all overlapping RNA-seq datasets). Where available CAGE and polyAseq data are used to trim the ends of the novel transcript. These transcript models will be tagged to ensure they do not enter the GENCODE geneset until they are checked by an annotator, modified as necessary to achieve the best possible annotation at the locus and the tag removed. In this way the computational pipeline checks simple and precise intersections of features and adds all relevant attributes and metadata which frees the annotator to spend more time assessing the functional biotype of the locus e.g. whether it is protein-coding or a pseudogene and correcting or improving models where necessary. We will also implement a similar pipeline to maximize the potential of third generation transcriptomic datasets i.e. CaptureSeq PacBio and SLRseq. Again, we will use transcripts derived by mapping and clustering reads to create a GENCODE transcript tagged to ensure they do not enter the GENCODE geneset without specific authorization from an annotator. In addition, transcripts at protein-coding loci will have a biotype assigned and CDS added based on the existing CDS annotation at the locus where appropriate. The transcripts will 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.

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) that regulate the passage of the ribsome along the mRNA (Refs) and consequently play a role in controlling translation initiation (Refs). Variation affecting uORFs 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, specifically by extending the annotation of genes to include their regulatory features. uORFs are not comprehensively 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 the same contextual information. Integration of uORF annotation with other datasets such as CAGE and RNA-seq data will allow capture of consequences of alternative TSS usage on the regulatory repertoire of genes. By integrating RP and MS data we will build on the initial annotation of uORFs to provide information relevant to their functionality such as translation initiation efficiency and encoding of stable proteins.

Finishing the Mouse Pseudogene Annotation 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 [[mouse or human sect?]] 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.

Mark Gerstein 5/12/2016 7:33 AM
Formatted: Highlight

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 [[mouse or human???]] We are going to leverage our experience in pseudogene transcription analysis 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.

Mark Gerstein 5/12/2016 7:34 AM
Formatted: Highlight

Mark Gerstein 5/12/2016 7:33 AM
Deleted: in human, worm and fly

Annotation of individual and population data

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 analysis results \cite{26432246}.

A key aspect of personalized annotation is correctly connecting the annotation to loss-of-function events, polymorphic (pseudogenes present in most individuals in the population but functioning genes in some) and unitary pseudogenes. We have explored in detail their implications for the reference annotation \cite{21205862} and observed that in general, LOF events and polymorphic pseudogenes are just different versions of the same event, mostly depending on the major allele frequency. [[relate to earlier]]

Specifically for LOF variants, we have developed a tool - Variant Annotation Tool (VAT) \cite{22743228} - to catalogue loss-of-function 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. [[link to earlier]]

We characterized putative LOF events in individuals from 26 different populations using the 1000 Genomes Phase 3 data.\cite{26432245}. Some LOFs may impact only one individual, resulting in the inactivation of an essential gene and leading to disease, while other LOFs can become fixed in the population as nonfunctional relics through pseudogenization. We also surveyed the impact of LOFs on personal annotation\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 \cite{21205862}. 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 and at .... .

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

Mark Gerstein 5/12/2016 7:34 AM
Formatted: Highlight

Mark Gerstein 5/12/2016 7:34 AM
Formatted: Highlight

Mark Gerstein 5/12/2016 7:35 AM
Formatted: Highlight

Mark Gerstein 5/12/2016 7:35 AM
Formatted: Highlight

Mark Gerstein 5/12/2016 7:35 AM
Formatted: Highlight

Benedict Paten 5/5/2016 10:49 AM
Comment [14]: Ref to Gorilla paper.

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. [[connected to VAT & aloft]]

Mark Gerstein 5/12/2016 7:36 AM
Formatted: Highlight

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.

STOP

Pseudogene annotation of individual genome sequence 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.

Mark Gerstein 5/12/2016 7:36 AM
Formatted: Highlight

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.

Personal Proteomics Analysis

One particular workflow within our pipeline focuses 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.

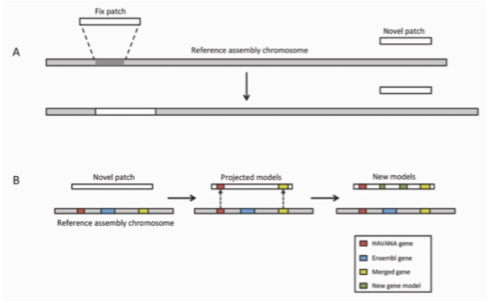

Figure 7: Patch

Pilot project 1: Graph genomes representation

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.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 6

Bronwen Aken 5/9/2016 3:37 PM
Comment [15]: Copied from previous grant

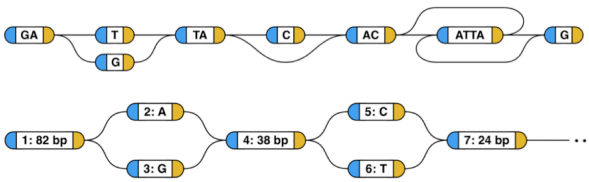

Figure 8: Example genome graphs: Each segment (node) holds some number of bases. A join (edge) can connect, at each of its ends, to a base on either the left (5', blue) or the right (3', yellow) side of the base. When reading through a thread to form a DNA sequence, you leave each base on the opposite side from which you entered, and reverse complement it if you enter on the 3' side and leave on the 5' side. The graph at the top 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 at the bottom 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)

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 that incorporate more common variation. Dealing with patches and alternative loci 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 walk 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.

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 9).

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 7

Paul Flicek 5/9/2016 3:36 PM
Comment [16]: Do decide on whether we want.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 8

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).

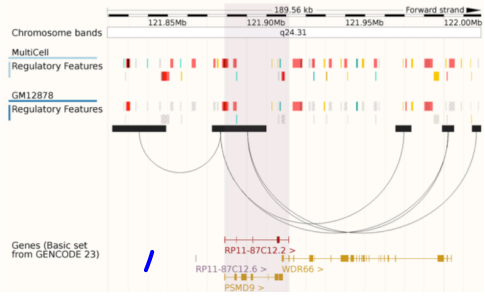

Figure 9: A prototype gene annotation. RP11-87C12.2 is currently represented as a set of exons and UTRs along the genome, however it is surrounded by cell type dependent regulatory elements. In addition, the ties between these regulatory elements and promoters (represented as arches) are also tissue dependent.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 8

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

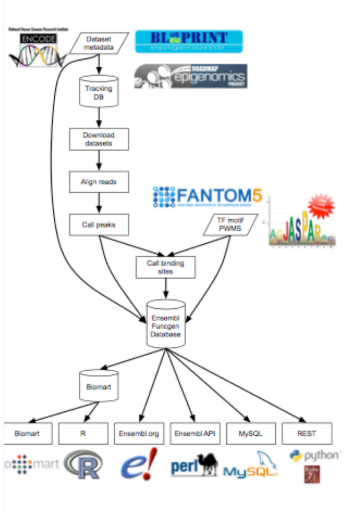

Figure 10: Summary of the existing Ensembl Regulatory Build pipeline: experimental data and external annotations are all integrated into a central annotation database.

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

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 9

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)

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 pseudogene annotation pipelines to produce an accurate map of pseudogenes in mouse strains.

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 and extract consensus proteins from Ensembl; 2) mark the coordinates of protein-coding genes; 3) six frame blast homology search to match the consensus peptides to the strain sequence; 4) refine results and eliminate

redundant hits; 5) merge hits and identify parents; 6) align parents and pseudogenes and check for disablements (e.g. premature stop codons); 7) assign pseudogene biotype.

RCPedia will 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 amazon machine images (AMI) compatible with Amazon AWS and other major cloud services so users can easily annotate additional genomes

To make annotation more efficient we will address the current bottlenecks of rapid data import, targeting manual annotation effort to where it is needed, automating new gene creation and significantly speeding up manual QC.

Data import Currently any data that requires remapping because it's from an earlier sequence assembly must be loaded into otter databases. Remapping will be added to ZMap so that the annotator can load data interactively. BAM/CRAM file support and Ensembl data access have already been added to ZMap but Trackhub support will also be added with full support for all common datatypes and full remapping support.

Directing Manual Annotation Effort To date we have not made effective use of gene build results to target annotator effort. We will use existing and augmented qualitative and quantitative information from otter and Ensembl gene builds to target genes that require manual annotation

Automating Gene Creation Annotators are now required to use vastly more data to create model genes and although ZMap currently allows the annotator to use "on the fly" alignment to create gene models from alignments too much of the process is still manual. We will augment current gene model building with code to aggregate isoform predictions and to automatically add supporting evidence meta data to the model.

Speeding up QC The otter/ZMap system has a prototype "quick open" system to allow annotators to move rapidly through check lists of regions requiring checking. This process will be speeded up by automatically producing the checklists from the same metrics used in the "Directing…" section above. The software will be modified to provide a "rapid open" option where only the minimum of data required for QC will be loaded.

QC checklists will also be produced from the Annotrack system used by Gencode collaborators to report issues with loci.

Paul Flicek 5/9/2016 3:50 PM
Comment [17]: From Ed Griffiths

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).

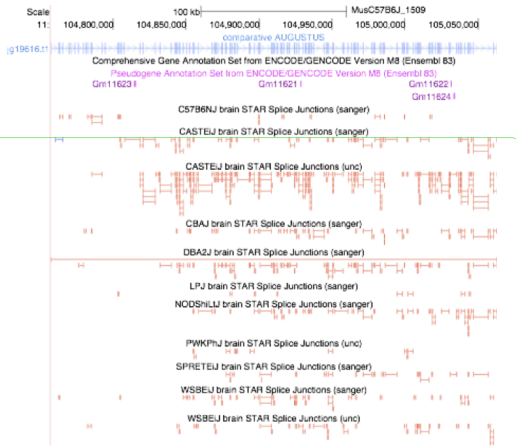

Figure 11: ***New Mus genes. Comparative Augustus identified a 138 exon transcript in a locus not previously annotated. This transcript has varying splice junction support, with the most coming from Mus castaneus. This transcript has been cloned and function is being investigated.***

Paul Flicek 5/9/2016 3:37 PM
Comment [18]: Decide

Paul Flicek 5/9/2016 3:37 PM
Comment [19]: Decide

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 10

How community annotation will be incorporated (~ 1 page)

Supporting external annotation efforts

GENCODE represents a world-leading infrastructure in the manual annotation of transcripts. 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.

However demand for manual annotation of transcripts across strains and species will outstrip our ability to provide such services via these existing mechanisms, therefore we will offer a two-tier system:

General Submission System We will support the submission of GFF3/GTF back to a new archive at EMBL-EBI to store external annotation. This archive will make use of existing otter technology to allow the archive to trace annotation calls back to individual submitters or consortia and as a way for automated processes to retrieve annotation. This linking will be accomplished via ORCID or another suitable authentication/authorisation system (see next section). Annotation will be QC'd upon submission and then, should it pass the checks, integrated into annotation builds as merge annotation. ZMap can already be used standalone but will be augmented with native support for CRAM, BigBED, BigWig and GA4GH APIs as these become important to quickly providing new primary evidence. Basic QC will also be integrated into ZMap allowing for high-quality annotation to be generated to submit to archives. Annotators can choose to use ZMap or to supply submissions from their own software.

"Trusted" Submission System otter uses a "Single Sign On" system (ORCID, Google etc.) to give access to its editing system to trusted users. These users can directly edit gene models whether working at the EBI or remotely. Currently the otter servers are located at the Sanger Institute but will be developed to run in other environments including cloud-based.

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

The GENCODE project plans a significant change in its institutional affiliation as part of this application as the project's center of gravity transitions from the Wellcome Trust Sanger Institute (WTSI) to the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) with the move of the Human and Vertebrate Annotation (HAVANA) group. This change is driven by a refocusing of the WTSI scientific strategy and represents the final phase in a series of moves that have seen several major informatics resources move from WTSI to EMBL-EBI over the last few years. These include Pfam, which moved in 2012, and Ensembl, which transitioned from a joint project of WTSI and EMBL-EBI to an EMBL-EBI only project in 2014. These transitions and any required reorganization in staff, project structure and computational infrastructure have generally been smooth and we expect that the process for the transition of the Havana team at WTSI will be similar. Moreover, Flicek, Aken, Yates and Zerbino were all directly involved with the transition of Ensembl and will call on this experience as needed. A more detailed description of the transition plan is below.

This renewal application will also see a change in the GENCODE lead PI to Paul Flicek (EMBL-EBI) as an immediate consequence of institutional changes above (Figure 12). He will be assisted in the administrative and reporting aspects of the grant by a Research Manager. The PIs at each of the performance sites will be responsible for project management and reporting for their site.

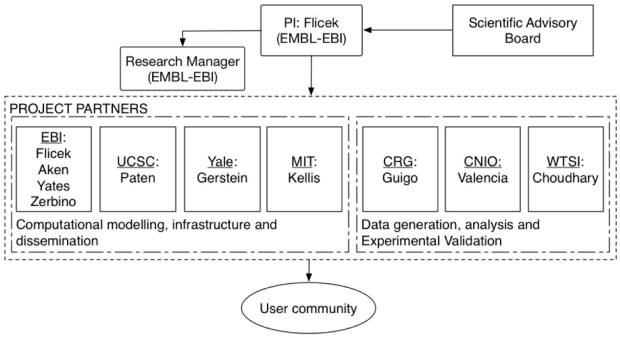

Figure 12: The Project's management structure with principal investigator names for each of the partner institutions.

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

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

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 11

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 11

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

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

Conflict resolution and transition planning All of the investigators have significant experience working together in a variety of projects over the last 10 years and no major conflicts are anticipated. Should any arise, we will aim to resolve any differences of opinion regarding the direction, process or strategy of GENCODE informally; if this is not possible, issues will be escalated first to the group of investigators and the lead PI (Flicek), then our SAB and the NHGRI. Should Flicek no longer be able to carry out the responsibilities of PI, a transition plan would be developed by the EMBL-EBI team and presented to the other investigators and then to NHGRI. As PI transitions were required twice in the current GENCODE funding due to staff departures at WTSI, we have an established model in the unlikely event this is needed.

HAVANA transition plan The intent to transition the HAVANA project from WTSI to EMBL-EBI was announced in early 2014 and the planning to ensure a smooth transition both scientifically and administratively has been actively underway since then. The date of final transition is set for 1 April 2017 coincident with the start of the proposed funding from this application and will follow an orderly process informed by previous resource transitions from WTSI to EMBL-EBI. Specifically, we will incorporate HAVANA into the Genes, Genomes and Variation cluster of resources (http://www.ebi.ac.uk/services/dna-rna), which is led by Paul Flicek and with the scientific responsibilities divided as described above. As the technical transition proceeds, we will move computational and software infrastructure from WTSI to EMBL-EBI through 2016 and early 2017 and provide HAVANA staff with guest logins to for testing before transition. The staff and physical transition will occur in 2017 although Adam Frankish's has already been offered an EMBL contract with start date of 1 April 2017 underwritten by EMBL core funds. Other staff will go through a selection procedure agreed by the Human Resources departments of WTSI and EMBL-EBI and that is compatible both with UK employment law and with the EMBL rules and regulations. This selection procedure is expected to be nearly identical to the process preceding the Ensembl staff transition. All HAVANA staff moving to EMBL-EBI will be offered EMBL contracts with start date of 1 April 2017, but may physically move to EMBL-EBI slightly before or after that date depending on various logistical considerations. Because WTSI and EMBL-EBI are separated by less than 100 feet this physical transition will incur a disruption of no more than 1 day.

Scientific Advisory Board

GENCODE will receive advice from a six member scientific advisory board (SAB) covering essentially all aspects of the project. The SAB will include three members of the SAB from the current iteration of the project for continuity and three new members The role of the SAB will be to provide advice on progress, priorities, new technologies, operational processes of the consortium and serve as representatives of our user community. The SAB will also assist in evaluating when the results of the GENCODE pilot project have reached appropriate maturity for scale up. They will also provide advice on any other improvements within their areas of expertise including operating in a cost-effective manner. The returning members of the SAB are Tom Gingeras, Cold Spring Harbor Laboratory; Ross

Hardison, Penn State University; and John Rinn, Harvard University. The new members of the SAB are Carol Bult, Jackson Laboratory; Lydie Lane, Swiss Institute of Bioinformatics;

The SAB will meet annually over the course of two days with the project PIs and may be called upon at other times for specific advice. The SAB meeting agenda will be set in discussions between the GENCODE PIs and the chair of the SAB. The format and process of the SAB meetings will follow a similar pattern to other EMBL-EBI project SABs including Ensembl and the NHGRI-EBI GWAS Catalog. In both of these cases, the SAB report is responded to formally via conference call six months after the SAB meeting (or sooner if required) and further updates are provided as part of the following SAB meeting.

Access and dissemination

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

Genome Browser Access

Access to the GENCODE annotation is primarily though the Ensembl and UCSC Genome Browsers, two of the most widely used resources for genome science. Both are funded separately for the majority of their activities and through this grant only for specific additional details directly related to GENCODE. This section largely describes how these resources will incorporate GENCODE into the rich genome information resources provided by Ensembl and UCSC to ensure that GENCODE is as widely accessible and as useful as possible.

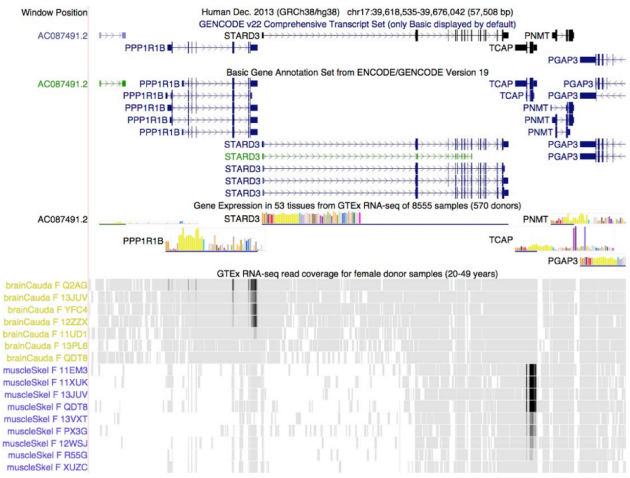

Figure 13: Gencode and GTEx in the UCSC Browser. View of a 56 Kbp region of human chromosome 17 where GENCODE annotates one non-coding and 5 protein-coding genes. Two genes in the region display tissue-specific gene expression as evidenced by GTEx RNA-seq including TCAP (titin cap protein) in muscle tissue and PP1R1B (a therapeutic target for neurologic disorders) in brain basal ganglia but not muscle. In this UCSC Genome Browser view, the main UCSC genes track (based on GENCODE V22, and colored by evidence strength) is configured to show a single isoform, while the earlier GENCODE V19 (used as the basis of the GTEx analysis shown here) shows all isoforms in the basic annotation set.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 12

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

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

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 12

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

The Ensembl team does the data merge that create the full GENCODE set from the HAVANA manual annotation and the Ensembl GeneBuild and, thus has GENCODE fully integrated at every step. The UCSC GENCODE group directly handles the ingestion of GENCODE data into the UCSC Browser, which frequently involves updating the browser source base to support GENCODE, 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. This data set is distributed both via UCSC and the genecodegenes.org web site. We will continue to support and enhance this approach and apply it to the next and subsequent versions of the mouse genome assembly.

New interfaces for genomic annotation display and access

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

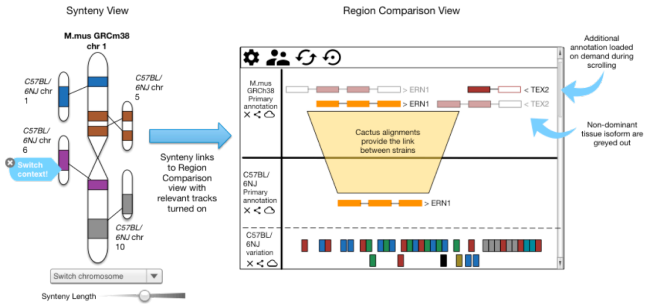

Figure 14: A view of mouse strain supported views moving from high-level synteny views to a configurable client side genome browser. Non-dominant tissue specific isoforms are greyed out.

Mark Gerstein 5/12/2016 7:36 AM
Deleted: 13

Mark Gerstein 5/12/2016 7:36 AM
Deleted: Figure 13

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

evidence interfaces will be modified to keep pace with the new sources of evidence being generated from this proposal.

Programmatic data distribution We seek to provide the GENCODE annotation through as many sustainable and modern distribution methods as practical. This will require the development of new publicly accessible APIs in addition to continued support for our more traditional pre-generated flat file freezes of the data sets. For example, to support interactive queries we will distribute GENCODE annotation via Global Alliance for Genomics and Health (GA4GH) compatible APIs integrated into the suite of APIs 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 separately as part of the Transforming Genetic Medicine Initiative (http://www.thetgmi.org). These checksums are intended to become the global unambiguous identifier for these sets. Change sets will also be released identifying new, retired and modified models with every GENCODE release.

We will maintain the current dedicate GENCODE portal (http://www.gencodegenes.org) for dedicated data download and specific project news. Annotation will continue to be made available over FTP and HTTP using common bioinformatics formats including GTF, GFF3, BED and BigBED. In addition we will continue to provide metadata as tab-delimited 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).

Software release

The primary output of GENCODE is genome annotation and not software. However, all software developed by this grant will be released publicly and generally via GitHub including via the existing Ensembl GitHub (https://github.com/Ensembl). Support for the use of the software will be though our existing our RT services linked from the GENCODE portal. The portal will be expanded to provide in-depth documentation and details of the processes used within the GENCODE consortium. We also plan to incorporate information on our software into face-to-face workshops and meetings.

All current GENCODE software produced by EMBL-EBI is open source and this will continue with the distribution of all project software under the Apache 2.0 license.

Training and outreach

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

In this application, 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.

Paul Flicek 5/11/2016 2:59 PM

Comment [20]: More details required on this. Where will in be, who will it target, etc.