APPLICATIONS OF NEXT-GENERATION SEQUENCING

The state of play in higher eukaryote gene annotation

Jonathan M. Mudge[1] *and Jennifer Harrow*[1,2]

Abstract | A genome sequence is worthless if it cannot be deciphered; therefore, efforts to describe — or 'annotate' — genes began as soon as DNA sequences became available. Whereas early work focused on individual protein-coding genes, the modern genomic ocean is a complex maelstrom of alternative splicing, non-coding transcription and pseudogenes. Scientists — from clinicians to evolutionary biologists — need to navigate these waters, and this has led to the design of high-throughput, computationally driven annotation projects. The catalogues that are being produced are key resources for genome exploration, especially as they become integrated with expression, epigenomic and variation data sets. Their creation, however, remains challenging.

Gene
Redefined for the modern era by Gerstein *et al.* (REF. 1) as "a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (that is, RNAs or proteins).

Genebuild
Used by GENCODE and Ensembl for a collection of transcript models generated by computational or manual annotation across an entire genome sequence. Protein-coding genes, long non-coding RNAs, small RNAs and pseudogenes may be included.

Transcript
Any form of RNA molecule that is transcribed from the genome sequence.

[1]*Department of Computational Genomics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK.*
[2]*Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Saffron Walden CB10 1XL, UK.*
jm12@sanger.ac.uk; jenniferlharrow@gmail.com

doi:10.1038/nrg.2016.119
Published online 24 Oct 2016

The core output of a gene annotation project could be described as an *in silico* transcriptome: a collection of 'models' referred to in this Review as a genebuild. However, genebuilds are found in a data server, not in the cell: they are only representations of the transcriptome that exists in nature. This fact has important implications for the study of biology: gene annotation is a key mechanism through which information is leveraged from genome sequences, and deficiencies in genebuilds will be propagated into downstream analyses. Thus, how close are our genebuilds to actual transcriptomes? Every publication seems to describe an entity that is larger, more dynamic and more functionally diverse than previously thought (as illustrated in FIG. 1) and this picture becomes even more complicated when considering the genomic sequences that regulate genes. In fact, the sheer complexity of the transcriptome may cause one to ask whether it is even possible that it could ever be completely described *in silico*. However, we are begging the question of whether we actually need to fully capture this complexity. A key annotation question concerns the proportion of the transcriptome that contributes to cellular function, and it could be argued that the goal of annotation projects should be to describe only this 'functional transcriptome': to extract the signal from the noise.

In this Review, we discuss the current state of play in higher eukaryotic gene annotation, and attempt to take genebuilds out of the 'black box' for the benefit of annotation users. Throughout, we use an updated definition of the term 'gene', beyond the traditional protein-coding criterion, as described further in REF. 1. First, we explain the key principles by which these resources are made, and why annotation projects are proceeding along alternative lines for different genomes. Inevitably, more work has been carried out on the human genome than any other, and many aspects of genome annotation are most effectively explained in this context. However, although human workflows are frequently re-used in the description of other genomes, such projects are not truly analogous. This is because their scientific goals are often substantially different — typically more limited in scope — but also because the resources available to support annotation have changed dramatically in recent years. Second, even human genebuilds have 'blind spots', and we wish to help users appreciate the biological information that is missing in such resources and how this can affect their work. The missing information generally reflects biological questions that remain unanswered, in particular, the issue of transcript functionality. Nonetheless, our biological understanding of the transcriptome is developing rapidly, and leaps in understanding are also being made in neighbouring fields of molecular biology, including proteomics, gene regulation and epigenomics. We explain below how gene annotation projects are coordinating efforts to combine such data sets into fully integrated views of genomic organization. Even so, it is clear that increasing genebuild complexity presents a considerable practical challenge to scientists, and we end this Review by discussing the problems that are faced by annotation projects in improving their usability.

What is gene annotation?

Annotation targets. FIGURE 2 summarizes the core principles of gene annotation workflows. Although numerous strategies have been used to describe different genomes and gene features, each ultimately represents the unification of two processes. First, annotation defines the structure of a transcript — for example, its

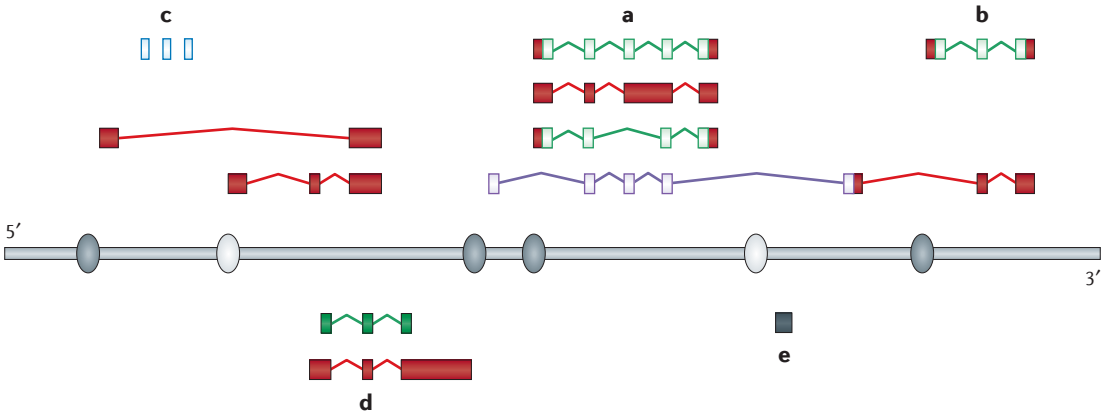

Figure 1 | **A modern view of the genomic landscape.** This hypothetical diagram illustrates the major types of genes and transcripts that are found in eukaryotic genomes. Two protein-coding genes are illustrated, gene **a** and gene **b**. Coding sequences (CDSs) are shown as open green boxes, and untranslated regions are shown as filled red boxes. Whereas locus **b** seems to generate a single CDS transcript, locus **a** generates two distinct protein isoforms through the differential incorporation of a central exon. Locus **a** also has an associated retained intron, and an additional 'read-through' transcript incorporates exons from both gene **a** and gene **b**. This transcript is subjected to nonsense-mediated decay (NMD; unfilled lilac boxes). Gene **c** is a long non-coding RNA (lncRNA) with two transcripts (red boxes), although three small RNAs are also transcribed from within one of its introns (open blue boxes). Locus **d** and locus **e** are unprocessed (filled green boxes) and processed (grey box) pseudogenes, respectively. Locus **d** is transcribed. A series of promoter regions (filled grey ovals) and enhancer regions (pale grey ovals) are indicated. Promoters are associated with transcription start sites for the various loci, whereas enhancers may be found some distance from the gene or genes that they regulate.

Functional annotation
The process of defining or predicting functional roles for transcript models during gene annotation.

Alternative splicing
Process by which a gene makes distinct transcripts through the use of different splice sites or exon combinations; these are known as alternative transcripts or transcript variants.

Pseudogenes
'Broken' genes that are derived from protein-coding loci. Can be formed by retrotransposition ('processed'), duplication ('unprocessed') or inactivation ('unitary', which may be polymorphic). All forms may be transcribed.

Long non-coding RNAs
(lncRNAs). Genes that do not contain protein-coding transcripts and that are not pseudogenes or small RNAs; a 200 bp size cut-off is typically applied to distinguish them from small RNAs.

Small RNA
A member of one of several known families of small RNA molecules. Includes the classic tRNA and rRNA families alongside more recent discoveries such as PIWI-interacting RNAs (piRNAs), microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs).

exon–intron architecture — and, second, it provides inferences into its potential function, for example, whether it is a protein-coding transcript. We refer to this second aspect as functional annotation. However, it is vital to appreciate that gene and transcript are not equivalent terms in annotation. This is illustrated by the fact that most genes generate multiple, distinct RNAs, especially through alternative splicing[2]. Transcripts are the major target of annotation projects; we regard 'gene annotation' as a process that creates 'transcript models'. As we discuss below, our modern understanding of transcriptional complexity within genes is driving the evolution of annotation strategies, as is the knowledge that eukaryotic genomes contain not only protein-coding genes, but also pseudogenes and long non-coding RNAs (lncRNAs), as well as small RNA families, including transfer RNAs, PIWI-interacting RNAs (piRNAs) and small nucleolar RNAs (snoRNAs)[3]. They may even contain RNA categories that remain to be discovered. In short, this complexity presents a substantial challenge to annotation projects (FIG. 1).

Annotation strategies. Numerous factors come into play when choosing an annotation strategy for a genome (FIGS 2,3). Obviously, financial considerations can place major constraints on the availability of human resources and computational power, as well as on the generation of experimental data to provide evidence for model construction (BOX 1). However, the strategy also depends on what it is hoped to achieve. For our species, genebuilds support scientific enquiries across a broad range of disciplines, and annotation resources are required to be as comprehensive as possible. The same is true for projects using classic laboratory species, such as *Mus musculus*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Other genomes may be sequenced to ask more specific scientific questions. For example, a common goal of sequencing projects within evolutionary biology is to find genes that have been subjected to positive selection. In this scenario, a high premium is placed on the identification of protein-coding sequences; features such as pseudogenes and small RNAs may even be completely ignored. Meanwhile, the Functional Annotation of Animal Genomes (FAANG) consortium plans to sequence and annotate livestock genomes in order to further our understanding of quantitative phenotypes[4].

Of course, it is broadly true that the more valuable a genome is to the scientific community, the more resources have been committed to its annotation. Therefore, humans[2,5], mice[6,7], *A. thaliana*[8], *C. elegans*[9] and *D. melanogaster*[10] have each been subjected to large-scale annotation projects over many years, involving numerous scientific institutes and sequencing centres (TABLE 1). In fact, the human and mouse genomes even have overlapping annotation resources that are independently produced, such as the genebuilds created by the RefSeq[5,6] and GENCODE[2,7] projects. Finally, we note that genome quality is an important factor when strategizing. One cannot create high-quality genebuilds using poor-quality genomes, and even modest genome assembly improvements can be massively beneficial to annotation projects, as demonstrated for the honeybee[11]. Indeed, annotation and sequencing have been carried out for both human and model organism genomes in a reciprocal manner, and we refer to them throughout as 'reference' genomes and genebuilds (FIGS 2,3).

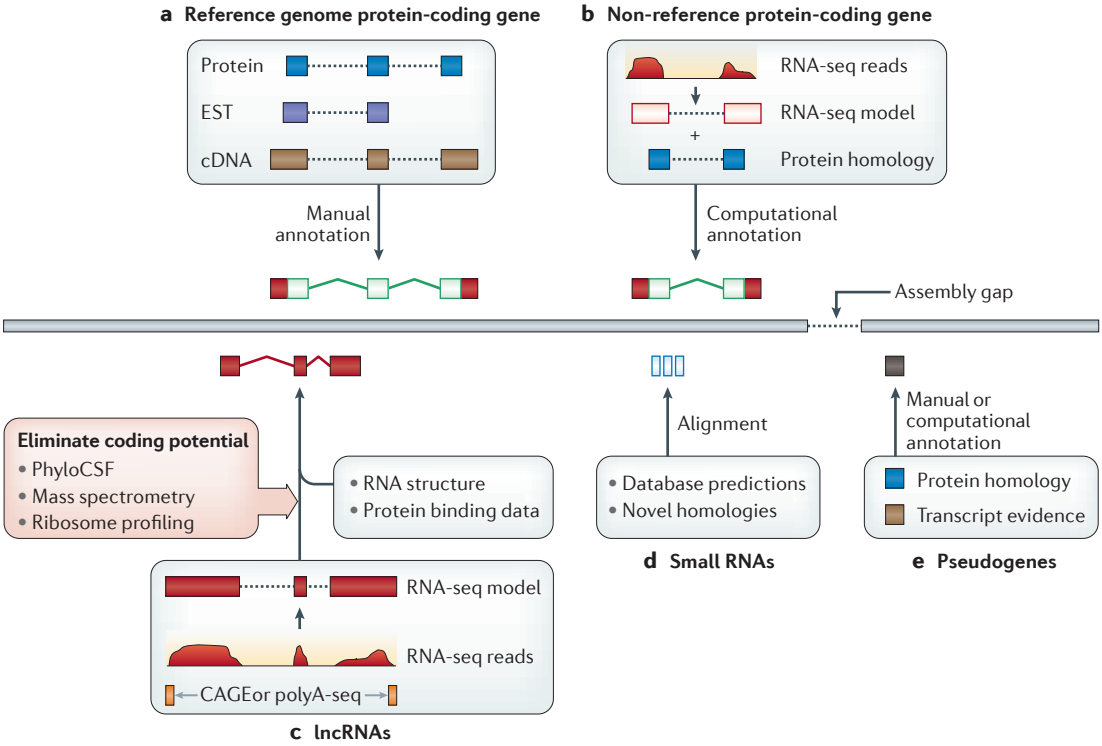

Figure 2 | **The core annotation workflows for different gene types.** These workflows illustrate general annotation principles rather than the specific pipelines of any particular genebuild. **a** | Protein-coding genes within reference genomes were generally annotated on the basis of the computational genomic alignment of Sanger-sequenced transcripts and protein-coding sequences, followed by manual annotation using interface tools such as Zmap[2], WebApollo[27], Artemis[127] and the Integrative Genomics Viewer[128]. Transcripts were typically taken from GenBank[129] and proteins from Swiss-Prot[13]. **b** | Protein-coding genes within non-reference genomes are usually annotated based on fewer resources; in this case, RNA sequencing (RNA-seq) data are used in combination with protein homology information that has been extrapolated from a closely related genome. RNA-seq pipelines for read alignment include STAR[130] and TopHat[131], whereas model creation is commonly carried out by Cufflinks[23]. **c** | Long non-coding RNA (lncRNA) structures can be annotated in a similar manner to protein-coding transcripts (parts **a** and **b**), although coding potential must be ruled out. This is typically done by examining sequence conservation with PhyloCSF[132] or using experimental data sets, such as mass spectrometry or ribosome profiling. In this example, 5′ Cap Analysis of Gene Expression (CAGE)[46] and polyA-seq data[47] are also incorporated to obtain true transcript end points. Designated lncRNA pipelines include PLAR[49]. **d** | Small RNAs are typically added to genebuilds by mining repositories such as RFAM[133] or miRBase[134]. However, these entries can be used to search for additional loci based on homology. **e** | Pseudogene annotation is based on the identification of loci with protein homology to either paralogous or orthologous protein-coding genes. Computational annotation pipelines include PseudoPipe[53], although manual annotation is more accurate[54]. Finally, all annotation methods can be thwarted by the existence of sequence gaps in the genome assembly (right-angled arrow). EST, expressed sequence tag.

Annotation evidence. Regardless of the scientific context of an annotation project, the most important factors to influence the genebuild that is produced are the evidence and methodologies that are used for model construction (BOX 2; FIGS 2,3). It is informative to consider how these elements have changed as reference annotation projects have become outnumbered by non-reference projects. In terms of evidence, the obvious difference is that Sanger-based transcript sequencing has been superseded by short-read RNA sequencing (RNA-seq)[12]. Thus, although the bulk of models in reference genebuilds were constructed on cDNA or expressed sequence tag (EST) evidence, such libraries are typically absent for other genomes. This fact is important when it comes to annotation. Most obviously, although RNA-seq is cheaper and more high throughput than earlier protocols, the RNA sequences obtained are shorter and more prone to error. This creates notable problems for annotation, as we discuss below, and it still remains easier to build accurate models on longer RNA sequences.

The second key source of evidence is protein sequences, but this situation is more complicated, as the field of experimental protein sequencing lags far behind that for RNA or DNA sequencing. Thus, the earliest annotation projects described coding sequences (CDSs) based on curated protein sequences from Swiss-Prot[13] (TABLE 1) and through the use of *ab initio* 'open reading frame (ORF)-finders' (REFS 14,15). The ORF-finding strategy sought to identify CDSs through a combination of codon frequency usage and ORF size, although many translations were subsequently judged to be spurious by manual curation. Currently, most non-reference

Coding sequences (CDSs). The regions of a transcript that are translated, that is, contain the information that encodes a protein sequence.

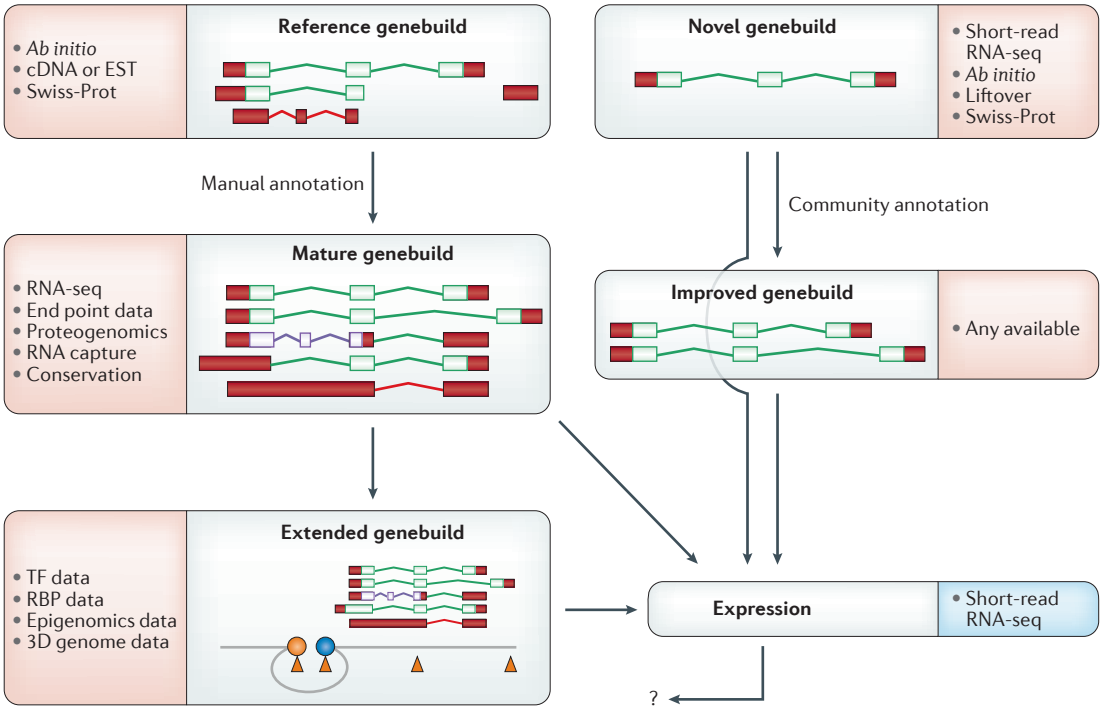

Figure 3 | **High-level strategies for gene annotation projects.** This schematic details the annotation pathways for reference and novel genomes. Coding sequences (CDSs) are outlined in green, nonsense-mediated decay (NMD) is shown in purple and untranslated regions (UTRs) are filled in in red. The core evidence sets that are used at each stage are listed, although their availability and incorporation can vary across different projects. The types of evidence used for reference genebuilds have evolved over time: RNA sequencing (RNA-seq) has replaced Sanger sequencing, conservation-based methodologies have become more powerful and proteogenomic data sets are now available. By contrast, novel genebuilds are constructed based on RNA-seq and/or *ab initio* modelling, in combination with the projection of annotation from other species (which is known as liftover) and the use of other species evidence sets. In fact, certain novel genebuilds such as those of pigs and rats now incorporate a modest amount of manual annotation, and could perhaps be described as 'intermediate' in status between 'novel' and 'reference'. Furthermore, such genebuilds have also been improved by community annotation; this process typically follows the manual annotation workflows for reference genomes, although on a smaller scale. Although all reference genebuilds are 'mature' in our view, progress into the 'extended genebuild' phase is most advanced for humans. A promoter is indicated by the blue circle, an enhancer is indicated by the orange circle, and binding sites for transcription factors (TFs) or RNA-binding proteins (RBPs) are shown as orange triangles. Gene expression can be analysed on any genebuild regardless of quality, although it is more effective when applied to accurate transcript catalogues. Clearly, the results of expression analyses have the potential to reciprocally improve the efficacy of genebuilds, although it remains to be seen how this will be achieved in practice (indicated by the question mark). 3D, three-dimensional; EST, expressed sequence tag.

genomes still lack substantial numbers of high-quality protein sequences, although ORF-finder efficacy has markedly increased as more genome sequences have become available. This is because a powerful way to find CDSs is through the ratio of synonymous to non-synonymous substitutions within a prospective ORF[16]; that is, to identify regions of DNA that are evolving as a protein-coding sequence.

Annotation workflows. The annotation 'workflow' chosen illustrates a second key difference between reference and non-reference genebuilds. Whereas all whole-genome annotation is highly dependent on computational processing, the projects for reference genomes have supplemented these processes with manual analysis. Generally, this involves teams of curators who either create transcript models from scratch, or else curate sets of computationally generated models[2,5,8–10], and can involve interactions with external groups such as UniProt[13] or gene nomenclature committees[17]. Manual annotation is regarded as the gold standard[18], and is one of the core workflows that allows genebuilds to be classified as 'mature' when performed to a significant degree (FIG. 3). Nonetheless, such labour-intensive work cannot cope with the number of species genomes that are becoming available, and most new genebuilds are generated entirely *in silico*.

Computational annotation has three core processes, depending on the resources available (FIG. 2). The first process is based on the alignment of transcript evidence. The second process is comparative annotation, by which the evolutionary closeness of two species allows for annotation — commonly the CDSs — to be 'projected' from one genome to another, or for evidence from one species to be used to build models on another. The third process is *ab initio* annotation, through which

Manual annotation
When a person constructs a transcript model *de novo* after appraising the available evidence (typically using software tools), or examines and potentially validates ('curates') a model that has been created computationally.

Computational annotation
The process of generating genebuilds through entirely *in silico* processes, that is, by the use of computational algorithms.

algorithm-based 'gene finders' such as GENSCAN[15] or AUGUSTUS[14] construct models on the basis of a priori knowledge of their likely sequence. Pure *ab initio* annotation is actually now uncommon in higher eukaryotic genomes, and these strategies are most often used in combination. The RefSeq Gnomon pipeline is a modified form of GENSCAN that can perform purely *ab initio* annotation, although it can also integrate RNA and protein homology data when available (see The NCBI Eukaryotic Genome Annotation Pipeline). Ensembl have adapted their pipeline in a similar manner, and the less species-specific evidence that is available for a given genome, the more annotation will be based on a combination of projection and *ab initio* modelling. Similarly, WormBase are combining projection from *C. elegans* with *ab initio* modelling in the annotation of other nematode genomes[9].

Even the largest annotation projects cannot yet describe genomes by the thousand, and researchers must often produce their own genebuilds. The Avian Genome Consortium, which aims to describe hundreds of bird genomes, is achieving this by working closely with Ensembl[19]. Annotation is being generated by the Beijing Genome Institute through the projection of existing bird and human Ensembl models, and is displayed in Avianbase, a modified form of the Ensembl schema[20]. RefSeq have also worked with external collaborators on specific genebuilds[5]. For researchers with fewer resources, numerous software tools can be used to carry out truly independent gene annotation. AUGUSTUS remains a popular choice; although it was developed as an *ab initio* tool for the Human Genome Project, its modern incarnation can incorporate transcript libraries and comparative evidence, albeit with a cost in terms of

Box 1 | **A description of gene annotation experimental data sets**

Transcript sequencing

Sanger-sequenced transcripts. RNAs obtained by traditional chain-termination methodologies. These are either cDNAs of approximately 1,000–2,000 bp in size (partly depending on the mRNA size) or expressed sequence tags (ESTs), which are single cDNA sequencing reads of approximately 500 bp.

Short-read RNA-seq. Whole-transcriptome shotgun-sequenced RNAs obtained as enormous libraries, typically on the Illumina platform[12]. The read length depends on the protocol used, although the most common data sets available are less than 200 bp. Reads are commonly generated as 'paired ends', in which the sequence is obtained from both ends of an RNA.

Long-read RNA-seq. The next wave of RNA-seq methods, generating longer sequences although at lower throughput. The Roche 454 platform provides reads of up to 1,000 bp, and the Iso-Seq methodology from Pacific Biosciences can capture whole RNAs.

Cap Analysis of Gene Expression (CAGE). Produces enormous ~27 bp fragment libraries extracted from the 5′ capped end of whole transcriptome RNA molecules when coupled to next-generation sequencing platforms[122].

***RNA Annotation and Mapping of Promoters for Analysis of Gene Expression (RAMPAGE)*[123].** Similar to CAGE, although provides longer paired-end reads as opposed to short sequence tags, with the size dependent on the short-read RNA-seq platform used.

PolyA-seq. Captures RNA sequence immediately upstream of the polyA tail. The protocol reported by Derti *et al.*[47] generates amplicons of 200–500 bp, although the size of the tags obtained will depend on the sequencing strategy used.

CaptureSeq. Uses strategically designed oligonucleotide probes to pull down target RNA from a sample. The captured RNAs can be sequenced using any common platform[44].

Functionality

Mass spectrometry (MS). Most commonly applied through the combination of liquid chromatography and tandem MS/MS, which produces large numbers of peptide spectral graphs based on their mass-to-charge ratio. Spectra are typically interpreted by comparison against a set of theoretical peptides extrapolated from an *in silico* coding sequence (CDS) database[75].

Ribosome profiling (RP). Identifies regions of transcripts that are undergoing translation. Cellular RNA is chemically degraded, allowing for RNA fragments that are 'protected' by ribosome binding to be recovered for high-throughput sequencing[81,82]. Also known as 'Ribo-seq'.

Ultraviolet (UV) crosslinking immunoprecipitation followed by sequencing (CLIP-seq). Ribosome-binding proteins are bound to their target RNAs, which are recovered and subjected to high-throughput sequencing[99]. It has the resolution to reveal binding sites within the RNA.

The extended gene

Hi-C. A massively high-throughput version of chromosome conformation capture methodologies. DNA is crosslinked across the sites of chromosome loops using chemical treatment, and these linkage sites are recovered and sequenced on next-generation sequencing platforms[105].

Chromatin Interaction Analysis Paired-end Tag Sequencing (ChIA-PET). An adapted form of Hi-C that enriches for specific DNA–protein complexes using chromatin immunoprecipitation followed by sequencing (ChIP-seq). It can thus be used to investigate the role of specific proteins in chromosome looping[124].

ChIP-seq. A method for analysing DNA–protein interactions in a cell[125]. It produces libraries of target DNA sites that are bound to a protein of interest, which are then mapped back to the genome to identify protein-binding regions.

Table 1 | **A selection of publicly available gene annotation resources for reference genomes***

Resource	Description	Primary institutions	URL
RefSeq	• Enormous integrated database of genome sequences, transcripts and proteins, covering all domains of life • Gene annotation is primarily based on the in-house computational Gnomon pipeline, and models for key species, such as humans, have been subjected to extensive manual curation	National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/refseq
GENCODE	• A multi-institute project providing gene annotation for humans and mice, in collaboration with the larger ENCODE projects • The genebuilds are a merge of manually annotated models produced by the HAVANA group with computational models generated by Ensembl • Further experimental and *in silico* validation for models is provided by other groups	• Wellcome Trust Sanger Institute • European Bioinformatics Institute • University of Lausanne • Centre de Regulació Genòmica • University of California, Santa Cruz • Massachusetts Institute of Technology • Yale University • Spanish National Cancer Research Centre	http://www.gencodegenes.org
Ensembl	• Multifaceted genome annotation resource, providing genebuilds alongside other annotations, such as regulatory and disease data • Also provides the Ensembl genome browser for integrated visualization • Gene annotation is based on the in-house Ensembl analysis pipeline	European Bioinformatics Institute	http://www.ensembl.org
UCSC Genome Browser	• Online tool supporting the visualization of genome annotations for numerous vertebrate and invertebrate species • Includes genebuilds from RefSeq, GENCODE and Ensembl alongside other gene annotations such as AUGUSTUS, CCDS and LRG • Certain groups have provided access to their own RNA-seq model collections as 'Track Data hubs'	University of California, Santa Cruz	https://genome.ucsc.edu
WormBase	• Database providing biological information, including genes and genome sequence, for the nematode *Caenorhabditis elegans* alongside other nematode species • Although all *C. elegans* gene models were initially computationally created, each has now been subjected to manual curation • Gene annotations for most other nematodes are generated computationally by the MAKER2 pipeline	• European Bioinformatics Institute • Wellcome Trust Sanger Institute • Ontario Institute for Cancer Research • Washington University, St Louis • California Institute of Technology	http://www.wormbase.org
FlyBase	• Central repository for genetics information relating to the insect family Drosophilidae, including a browser for gene annotations • Effectively all gene annotations have now been manually curated	• Harvard University • Indiana University • University of Cambridge	http://www.flybase.org
The *Arabidopsis* Information Resource (TAIR)	• Database of genetic and molecular data for the model plant *Arabidopsis thaliana*, including gene annotation • Models were initially produced by the *Arabidopsis* Genome Initiative, improved by the Institute for Genomic Research before being further improved and maintained by TAIR • The models have been subject to extensive manual curation, and community annotation is now facilitated via Web Apollo	Phoenix Bioinformatics	http://www.arabidopsis.org
UniProtKB	• A unified protein repository incorporating the Swiss-Prot and TrEMBL databases of protein sequences • Swiss-Prot is manually annotated by expert curators (based on the literature and manual gene curation), whereas TrEMBL contains computationally analysed entries mostly extracted from computationally derived transcript models	• European Bioinformatics Institute • Swiss Institute of Bioinformatics • The Protein Information Resource	http://www.uniprot.org
Roadmap Epigenomics Project	• Multi-institute collaboration developing a resource for the presentation and processing of human experimentally derived epigenomics data • Aims to generate reference epigenomes across a large variety of cell types • Includes data on gene expression, histone modification, DNA methylation and chromatin accessibility	The National Institutes of Health Epigenomics Mapping Consortium	http://www.roadmapepigenomics.org

Table 1 (cont.) | **A selection of publicly available gene annotation resources for reference genomes***

Resource	Description	Primary institutions	URL
The ENCODE encyclopedia	• Computational analysis pipeline being developed by the multi-institute ENCODE project to summarize the findings of experimental data sets across the genome sequence, including RNA-seq, Hi-C, ChIP-seq and histone marks (and incorporating data from the Roadmap Epigenomics Project) • For example, it can help users to extrapolate whether a given region looks like an enhancer	The ENCODE consortium	https://encodeproject.org/data/annotations
FANTOM (Functional annotation of the mamalian genome)	• International research consortium seeking to obtain further knowledge of the human and mouse genomes and transcriptomes • Since the year 2000, the project has shifted its focus from cDNA annotation, to transcription start and promoter analysis, and onto the description of lncRNAs	Coordinated by RIKEN Yokohama	http://fantom.gsc.riken.jp

CCDS, consensus coding sequence; ChIP-seq, chromatin immunoprecipitation followed by sequencing; ENCODE, Encyclopedia of DNA Elements; HAVANA, Human and Vertebrate Analysis and Annotation; lncRNAs, long non-coding RNAs; LRG, Locus Reference Genomic; RNA-seq, RNA sequencing. *This table is an entry point for exploring eukaryotic annotation resources in more detail, and is not intended to be comprehensive; the complete list of projects and groups that have contributed to gene annotation in the genome-sequencing era would be exceptionally large. Furthermore, it has not been possible to list individual groups contributing to the FANTOM and ENCODE projects owing to space limitations.

speed and ease of use[14,21,22]. For such practical reasons, researchers often annotate their genome using a simpler RNA-seq assembly pipeline such as Cufflinks[23]. Besides suffering from the RNA-seq assembly problems discussed below, such methods are severely limited by the fact that they do not produce true functional annotation (see below), and, in common with *ab initio* builders, will typically generate a single model per gene. We do not regard these catalogues as true genebuilds.

Community annotation. For genomes such as that of the rat it has become clear that computational genebuilds cannot meet the needs of the community, and yet adequate resources are not available to follow the RefSeq or GENCODE reference annotation model. One solution is to manually improve the annotation in a systematic, collaborative manner based on 'crowdsourcing' (REF. 24) (FIG. 3). Either the interested parties meet in person and carry out a large amount of annotation over a short period of time (known as a 'jamboree')[25], or else they work remotely over a longer of period of time, following the same annotation criteria[26] and using software such as WebApollo, which allows for 'live' annotation to be shared remotely[27]. This remote working strategy has been central to the annotation efforts of VectorBase, which is a community effort that seeks to describe the genomes of invertebrates that transmit disease to humans[28]. Nonetheless, the output of most projects cannot match reference curation teams in scale, and the focus is often limited to a particular biological theme, for example, the annotation of porcine immunology-related genes[29].

Annotation in population genomics. It is now commonplace to generate multiple genome sequences from the same species, especially to aid the study of variation. Human studies have inevitably led the way, with projects such as the UK10K generating genomes by the thousands[30], although 'population genomics' has now been carried out for species as diverse as rice[31] and killer whales[32]. Do these genomes require annotation? If DNA variation is of primary interest, single nucleotide polymorphisms (SNPs) can simply be extracted and displayed against the main assembly for that species. Furthermore, if users wish to 'browse' additional genomes then transcript models can be projected from the main assembly. Projection is part of the annotation strategy of the Mouse Genomes Project (MGP) — which has released 36 genome sequences of laboratory mice and wild-derived strains — in combination with *ab initio* modelling[33].

Nonetheless, the MGP also illustrates scenarios for which manual intervention is desirable. For example, when genes do not successfully project then manual curation can resolve whether this is due to variation or to genome sequence error[7], and it can also be used to judge the quality of *ab initio* models. Manual annotation can also be essential when investigating structural variants (SVs), which are of great interest to biologists owing to their association with disease and evolution[34]. The Genome Reference Consortium (GRC) continue to improve the human and mouse genome assemblies, and have created a series of 'alternative (alt) loci' for both species that target allelic variation, as well as SV regions containing genes that are subject to copy number variation (CNV)[35]. For example, the GRCh38 human genome assembly contains eight haplotypes for the major histocompatibility complex (MHC)[36] and 35 for the leukocyte receptor complex (LRC)[37]. The interpretation of CNV gene families can be difficult: gene copies are often highly similar or even identical, and a protein-coding gene in one genome may be pseudogenized in another. It is impossible to simply extract this information for display against a reference genome, and such regions can be difficult to resolve without manual intervention.

When is a genebuild complete?

Identifying missing transcripts. Having discussed progress in gene annotation, we now turn our attention to the limitations of existing genebuilds. Users should understand that even human genebuilds are works in

Transcription start site
(TSS). The base pair on the genome where transcription begins.

Polyadenylation tail
A sequence of adenosine monophosphates attached to the 3′ end of an RNA as transcription terminates, beginning at the polyA site.

progress, and we now consider how far into the distance the finishing line for such endeavours might be found. Logically, a complete genebuild would contain all the transcripts that a genome produces, with accurate functional information attached to each model. Certainly, an attempt to identify all the transcripts produced by a genome may be considered a key goal in the generation of a mature genebuild (FIG. 3). However, multicellular organisms have almost as many transcriptomes as they have cells, and an emerging goal for annotation projects is to provide information on where and when transcripts are expressed. In practice, this depends on the prior creation of unified transcript catalogues, that is, where transcripts from all sources are combined. Meanwhile, transcribed sequences may be absent from genebuilds for three reasons: existing models may be incomplete, that is, truncated at one or both ends; whole transcripts could be missing within existing genes; and entire genes could be absent. Obviously, the relevant RNAs may not be present in transcript libraries, which is most likely for transcripts with restricted expression. Nonetheless, additional transcripts clearly exist in libraries that are not yet incorporated into genebuilds; human RNA-seq projects routinely describe thousands of novel models[38], in common with targeted efforts on other reference genebuilds such as *A. thaliana*[39].

Unfortunately, RNA-seq continues to confound annotators[40]. The most common protocols generate short reads that are less than 200 bp in size (BOX 1), which is far shorter than the average mRNA. Reads are aggregated to predict full-length transcripts, although this process is challenging[12]. RNA-seq models are emphatically predictions, and have not been incorporated wholesale into most reference genebuilds owing to quality concerns[41]. As noted above, RNA-seq models have instead proved a frequent necessity for annotating genomes that lack Sanger-sequenced transcript libraries. Meanwhile, long-read RNA-seq libraries are becoming available to improve annotation (BOX 1). It is easier to align longer reads with accuracy, although the sequencing quality is still not comparable to that of Sanger protocols[42]. An interesting development is synthetic long-read (SLR)-seq, which circumnavigates the problem of short-read transcript assembly by generating synthetic long reads through the reconstruction of fractionated and barcoded short RNA fragments[43]. Efforts are also being made to complete transcript catalogues based on targeted methodologies. CaptureSeq involves the use of genomic hybridization arrays to 'pull down' portions of the transcriptome for sequencing[44]. It is effective at isolating transcripts that are expressed at low levels, which may otherwise be 'drowned out' in whole RNA assays[45]. CaptureSeq is typically used to identify novel genes (FIG. 4) and to target partial models for completion. The experimental set-up is laborious, however, and its use is currently limited to humans and mice.

Annotating transcript end points. How can one tell whether a model is precisely full length, that is, whether it contains the transcription start site (TSS) and transcript end site (TES) of the RNA? TESs can be identified from the 3′ polyadenylation tail, although there is no consistent

Box 2 | **Defining functionality within the genome and transcriptome**

There has been much debate about the proportion of the eukaryotic genome that is truly functional, generally through disagreements on how 'functionality' should actually be defined. Evolutionary biologists have traditionally placed a high value on the maxim that 'conservation equals function', and may thus doubt the functionality of non-conserved bases[87]. More recently, human experimental biology projects such as ENCODE have used a biochemical definition that is based on the use of high-throughput assays, including RNA-seq and immunoprecipitation techniques[104]. However, the proportion of the genome that participates in transcription and epigenomics is far larger than that which displays conservation, hence these definitions seem to be irreconcilable. The process of genome annotation can provide useful insights into this debate, as it approaches things from a different direction. In this case, the initial focus is not on individual base pairs, but on whole sequence elements such as transcripts, and it is of course transcripts that are the primary effectors of genomic information. The question can thus be restated as: which transcripts are functional and how do they function? In this context, the word 'functional' primarily concerns the role of the transcript in the cell; whether it is translated, for example, or whether it actually makes no contribution to physiology. It would thus seem reasonable to describe an mRNA as a 'functional transcript', and to describe a transcript that is simply stochastic noise as 'non-functional'. The more challenging ground is found between these two poles. For example, it is debatable whether the *AIRN* long non-coding RNA (lncRNA) transcript is a functional molecule given that it is ultimately a by-product of a regulatory pathway[93], and the same question could be asked of regulatory non-productive transcripts (NPTs) that are found within protein-coding genes. Certainly, the generation of these transcripts directly mediates functional processes, and for this reason we prefer to regard them as functional molecules.

The question of how transcripts function, has in fact, a further layer of complexity. A typical mRNA contains various sequence features: most obviously the coding sequence (CDS), but potentially also regulatory sequences such as *trans*-factor binding sites, secondary structures and upstream open reading frames (uORFs). These are all potential targets for annotation, which suggests that we should regard an mRNA more properly as a 'functional transcript that contains a number of distinct functional features'. It is also interesting to note that mRNAs may contain sequences that are not functional according to the strict evolutionary definition; this is in fact more dramatically the case for well-studied lncRNAs such as *HOTAIR*[126]. Although this discussion may seem esoteric, it is actually of great practical importance. Gene annotation is of particular value in the clinic, where it is often used to aid the interpretation of disease-associated variants. A clinician would like to know not only that a given mutation is associated with a functional transcript, but also which sequence features it affects within that transcript. Although sequence conservation can be a useful aid to the prioritization of variants, annotation processes are ultimately required to convert such information into actual biological features.

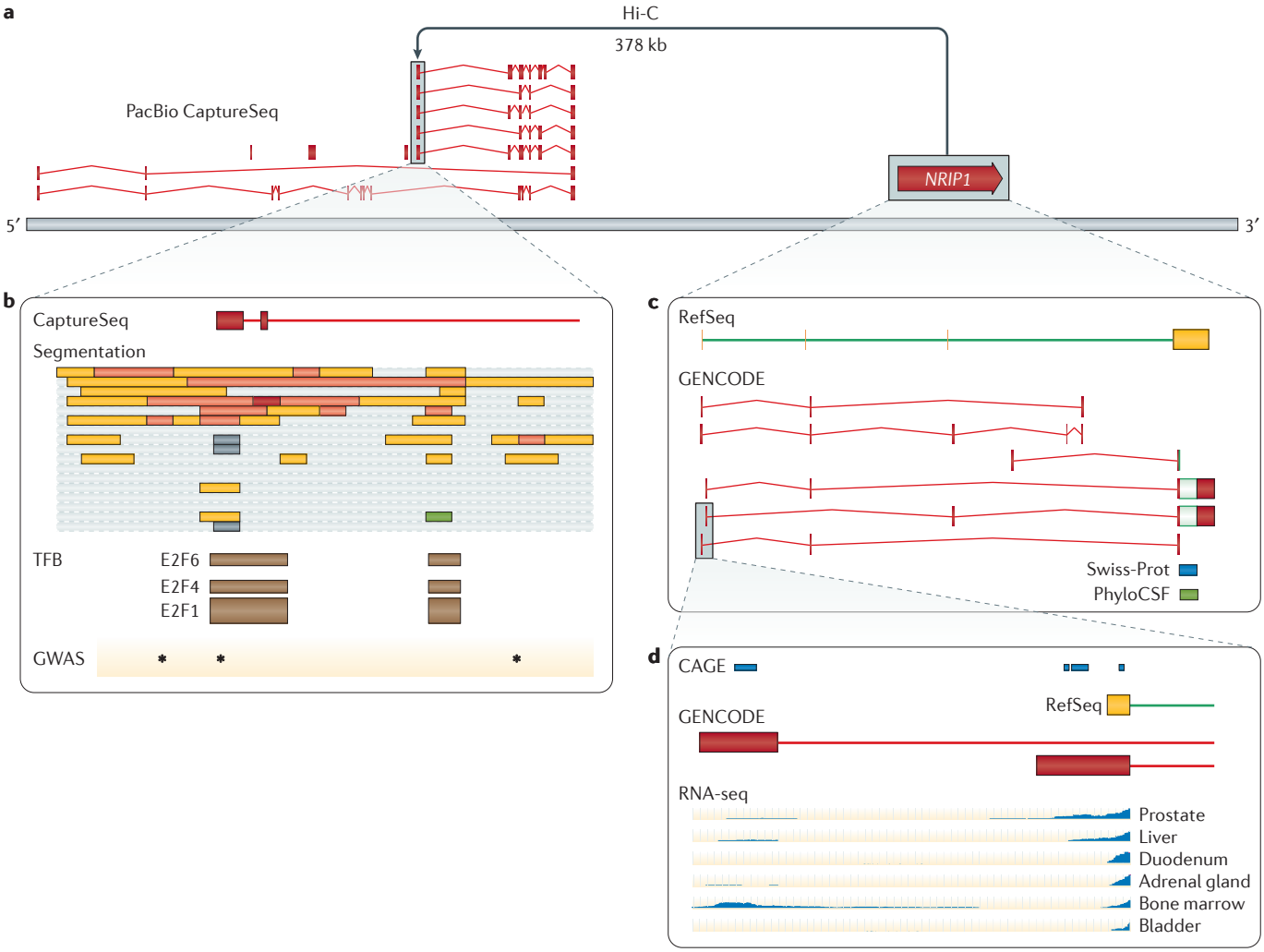

Figure 4 | **Transcriptional complexity in the *NRIP1* locus. a** | Capture Hi-C[107] indicates that the nuclear receptor interacting protein 1 (*NRIP1*) locus on human chromosome 21 forms a loop with a previously unannotated region nearby. Pacific Biosciences (PacBio) CaptureSeq data could be aligned here (R. Johnson, personal communication), leading to the annotation of lncRNA OTTHUMG00000488671 in GENCODE. **b** | A long non-coding RNA (lncRNA) transcription start site (TSS) falls within a sequence described as enhancer- or promoter-like (orange and red boxes) in different cell lines by the Ensembl-processing[135] of ENCODE epigenomics data[104]. Three transcription factor binding (TFB) regions — E2F1, E2F4 and E2F6 — colocalize based on ENCODE chromatin immunoprecipitation followed by sequencing (ChIP-seq) data[104]. In combination, these data suggest an 'extended gene model' for *NRIP1*, which may aid the interpretation of three genome-wide association study (GWAS) signals linked to Crohn disease (rs2823286, rs1297265 and rs1736020; shown as asterisks), as previously noted by Mifsud *et al.*[107]. **c** | *NRIP1* contains one curated transcript in RefSeq and six transcripts in GENCODE. The coding sequence (CDS; shown as an open green box) has Swiss-Prot support, and a PhyloCSF conservation signal[132]. (The untranslated regions (UTRs) are shown as filled red boxes.) **d** | Two distinct first exons of *NRIP1* are annotated, both supported by 5′ Cap Analysis of Gene Expression (CAGE) data[46]. RNA-seq from Uhlen *et al.*[117] indicates differential expression, with use of the upstream exon apparently limited to the bone marrow (and adipose; not shown). This TSS is dominant in white blood cells, which are bone marrow-derived cells. RNA-seq and CAGE support a more general expression profile for the downstream first exon, with evidence of TSS variability.

Translation initiation site
(TIS). The codon that is translated to give the first amino acid of a peptide; almost always ATG; also known as a START codon.

STOP codon
The final codon of a protein translation; almost always TAG, TAA or TGA; also known as a translation termination site or codon.

diagnostic sequence for the TSS so it is difficult to know whether a transcript is 5′ truncated. Such ambiguity is problematic, because confident functional annotation depends on accurate structures. Whereas a CDS may be obvious on a full-length transcript, it could be missed on a truncated version, especially if sequencing has not encompassed the translation initiation site (TIS) or STOP codon. The implications of this are particularly concerning in disease genetics, where CDS annotation is the key data set through which identified genetic variants are interpreted. This problem seems to be solvable, however, given the advent of modified RNA-seq assays to sequence end points (BOX 1). Notably, FANTOM5 has generated millions of 5′ Cap Analysis of Gene Expression (CAGE) sequences from more than 400 tissues or cell lines for humans and mice[46]. Although the major goal of this project is to study transcript expression, these data are also proving highly useful for manual curation efforts[7,40] (FIGS 2,3).

However, genes display considerable variability in their end points[46,47] — even within the same exon — which challenges our assumptions about the relationship between transcript models and cellular RNA. Annotation projects using these data sets can try to represent this diversity or else attempt to summarize it. The key issue is whether this complexity is biologically meaningful. This may not be the case for end point 'wobble', which could reflect stochastic variability in the binding of the RNA polymerase II or polyadenylation complexes. If a project favours simplicity, gene-boundary data can be converted into single base-pair sites and can be incorporated into computational workflows. For example, Boley *et al.* used CAGE and polyA-seq data in the generation of *D. melanogaster* RNA-seq-based models[48], and the PLAR pipeline incorporated polyA-seq in the annotation of 17 vertebrate genomes[49]. However, differential end point usage can have important functional consequences, especially linked to gene regulation (as discussed below).

Functional annotation

When RNA-seq protocols can produce accurate, full-length transcripts then the need to curate these structures will diminish. Instead, the legacy of genebuilds is likely to be their functional annotation. Traditionally, functional annotation centred on the question of which models encode protein. We now know that non-coding genes and untranslated transcripts can function in many different ways. Indeed, the definition of 'functional' remains controversial in genomics, as we discuss in BOX 2. Nonetheless, a survey for protein-coding loci remains a common starting point for the annotation of novel genomes, and efforts to annotate the complete set of translated regions are ongoing even in reference genebuilds.

Distinguishing protein-coding genes and pseudogenes. As discussed above, CDS annotation is typically based on the incorporation of curated protein sequences, as well the computational processing of protein homologies and conservation signals. However, a genuine signal does not confirm that a region is coding, rather it confirms that the region has been coding at some point in its history. This distinction is crucial, as eukaryotic genomes contain large numbers of pseudogenes[50] (FIGS 1,2). Pseudogenes are a major confounding factor for computational CDS annotation: they may contain large ORFs and are frequently transcribed, and duplicated or retrotransposed pseudogenes with high sequence similarity to the parent locus can complicate both CDS projection and RNA-seq mapping[51]. Even their manual interpretation is complicated, which is a major reason why GENCODE, RefSeq and UniProt do not agree on the number of human protein-coding genes. For example, retrotransposition can generate intact copies of the parental CDS[52], and whereas GENCODE have annotated more than 300 'retrogenes' as protein-coding, the functionality of those that do not exhibit conservation remains speculative. Alternatively, although duplicated copies of a parent gene may have disrupted CDS, it can be unclear whether this causes loss of function (LoF)[51]. These ambiguities are exacerbated in lower quality genome sequences: CDS disablements in prospective pseudogenes — and LoF mutations in resequenced genomes — could instead be sequencing errors. Although there are a limited number of dedicated tools for the computational analysis of pseudogenes, including PseudoPipe[53], manual annotation remains preferable[54].

The coding potential of alternative splicing. A second complication in CDS annotation is that protein-coding genes can make distinct proteins (isoforms) through alternative splicing[55,56]. However, although alternative splicing is ubiquitous among multi-exon genes, the extent to which it generates proteomic diversity is debatable[57,58]. Indeed, it should be emphasized that the bulk of CDS annotation in eukaryotes is based on extrapolation as opposed to experimental evidence, and this fact is likely to have important implications across the field of biology. Certainly, alternative splicing does not always generate isoforms, and we refer here to transcripts from protein-coding genes that do not generate mature proteins as non-productive transcripts (NPTs). Distinguishing coding transcripts and NPTs is a major goal of maturing annotation projects[40], although RefSeq and GENCODE approach the problem from different directions. RefSeq have traditionally focused on models that are considered likely to be protein coding on the basis of additional evidence, for example, from Swiss-Prot. GENCODE annotate such models along similar lines; however, they ultimately aim to provide functional annotation for all identified transcripts. The coding potential of these additional transcripts is judged in comparison with a model within the gene that is known to be protein coding. Thus, an exon-skipping transcript is likely to be annotated as coding if it does not contain a frameshift. Such first principles-based annotations are speculative. However, GENCODE reappraise their human CDS based on scoring that is provided by the annotation of principal and alternative splice isoforms (APPRIS) pipeline[59], which combines CDS conservation with predictions into the effects of alternative splicing on known protein domains. APPRIS has generated annotation for six mammals, as well as *C. elegans* and *D. melanogaster*. Finally, we note that GENCODE and RefSeq are in fact collaborating on the ongoing Consensus CDS (CCDS) project, the core goal of which is to produce CDS sets that are unified between different annotation projects[60] (TABLE 1).

Non-productive transcription and untranslated regions in protein-coding genes. If transcripts within protein-coding genes do not make proteins, what do they do? All cellular machines are error-prone, and intron retention (IR), for example, could simply be due to spliceosome failure[40,57]. Furthermore, the sequence motifs that govern transcription, splicing and translation are typically basic, and 'cryptic' sites throughout the genome can act as competitors to canonical sites. Such knowledge re-contextualizes the question of when a transcript catalogue is complete, and it is generally accepted that a proportion of the transcriptome is aberrant 'noise', although

Isoforms
Protein molecules that differ in their amino acid composition from other translations made from the same gene, for example, owing to alternative splicing.

Intron retention
Occurs when a transcript does not splice out one or more introns, that is, this sequence is left incorporated into the mature RNA.

the size of this proportion is debated[57,61–64]. However, NPTs can impart gene regulation. Many protein-coding genes reduce their protein output not by 'switching off', but by directing their transcription into non-productive pathways. The best characterized mechanism by which this occurs is alternative splicing-linked nonsense-mediated decay (NMD)[65]. Although NMD was originally understood as a mechanism for the degradation of aberrant transcripts, many genes use this pathway to dampen their output in a regulated manner[66], typically through the splicing of a poison exon that contains a termination codon. Regulation can also be imparted through intron retention, which is emerging as a key control mechanism in haematopoiesis[67]. In fact, up to three-quarters of mammalian genes exhibit systematic intron retention, especially in cell types in which expression is not anticipated; intron retention may be 'functionally tuning' these cells[68]. The contribution of intron retention to gene regulation is particularly well established in *A. thaliana*[39].

Regulatory NPTs can also be invoked through differential TSS usage, although in this scenario the transcripts are potentially less productive rather than non-productive. For example, the human granulin (*GRN*) gene produces two transcripts that have highly different rates of translation, even though they have the same CDS[69]. The weakly translated form has a longer 5′ untranslated region (UTR), which incorporates a short upstream ORF (uORF) that competes for ribosome binding with the regular CDS. TSS switching to the short 5′ UTR form thus increases protein production. Most human and mouse first exons contain multiple TSS regions according to FANTOM[46], and most 5′ UTRs contain uORFs (J.M.M. and J.H., unpublished observations). Could transcripts that differ in their 5′ UTRs have precisely the same translational efficiency? The regulatory importance of uORFs is certainly recognized[70], although they remain a blind spot for even reference genebuilds.

The situation for differential TES use is at least superficially similar, and there is evidence that this process can modulate RNA stability and localization by creating transcripts that differ in their secondary structure or in their response to *trans* factors[71,72]. Just as for TSSs, annotation projects generally extend models to the maximum 3′ distance that is supported by transcriptional evidence, and do not annotate additional models based solely on alternative TESs.

NPTs are not simply a late-stage target for mature genebuilds; such transcripts will also be sucked into the annotation pipelines for novel genebuilds, and are at risk of mis-annotation. Nonetheless, if such knowledge could be captured it may radically change the way users perceive their genes of interest. An obvious question is how to distinguish models that invoke NPTs as part of regulatory programmes from those that arise as stochastic noise. Currently, this is being achieved through low-throughput laboratory studies, and it is notable that the differential TSS usage in *GRN* is currently not represented in GENCODE or RefSeq. However, global insights can sometimes be gained from comparative analyses; poison exons, for example, are often highly conserved in vertebrates[66,73]. It may be that the blueprints for such phenomena can ultimately be read in the genome, for example, in the form of binding sites for *trans*-acting factors[74].

Annotating proteins with experimental data. CDS annotation is interpretive because the chemistry of the protein molecule makes it far less amenable than RNA to sequencing. However, recent advances in mass spectrometry (MS) have given birth to proteogenomics: the identification of CDS through the integration of peptide data and genomic or transcriptomic sequences[75]. The experimental parameters for this emerging technique are still being established[76–79]. Above all, it is a completely different paradigm from that of RNA sequencing (BOX 1): peptide identification depends not on mapping, but on the correlation between the spectra that are observed in the experiment and those predicted to be produced within a CDS search space defined *in silico*. The design of this 'search space' has a substantial bearing on the results, and the false-discovery rates for proteogenomics assays are notoriously difficult to gauge[75]. Peptides are also frequently too short to distinguish isoforms. Furthermore, not all proteins are amenable to MS owing to their chemistry or cellular location, and it is harder to capture proteins with low expression[75]. Nonetheless, the utility of this technique for CDS identification and validation is clear[80].

Ribosome profiling (RP) identifies RNA regions that are undergoing translation[81,82] (BOX 1). Currently, there is no community consensus on how RP data sets should be used in annotation, and there are technical questions remaining regarding their production and interpretation[83]. It seems that genuine RP regions do not necessarily highlight actual CDSs, that is, that RNA–ribosome interactions do not always lead to the production of mature proteins[84]. This could be because certain interactions are transient as opposed to truly functional[85]. Alternatively, there is evidence that lncRNAs and protein-coding genes can use ribosome binding to impart regulation, for example, via NMD[65]. Nonetheless, others suspect that RP data sets truly identify significant numbers of typically small proteins that do not display conservation or homology to known proteins[86]. The concept of 'lineage-specific' biology provokes strong opinions[87], and this debate is important from an annotation perspective, for which conservation is a key proxy for functionality. Although RP has been carried out on at least six other eukaryotic genomes so far — as collated by the RPFdb resource[88] — these data have not yet been incorporated into the computational annotation pipelines of reference genomes.

LncRNA annotation. LncRNAs present challenges to annotation projects similar to those presented by NPTs; they can have functional roles in mammalian cells[89], although it has been argued that many are transcriptional noise[90]. Pertinently, lncRNAs are typically weakly conserved in comparison with CDSs, and show high evolutionary turnover[91,92]. Nonetheless, it may be misguided to judge lncRNA functionality solely by analogy to protein-coding transcription, as the base-pair

Nonsense-mediated decay
(NMD). Cellular 'surveillance' mechanism that targets transcripts for destruction. Imprecisely understood, although transcripts featuring termination codons more than 50 bp upstream of splice junctions are thought likely to be substrates.

Poison exon
An exon that prevents correct coding sequence translation when incorporated into the transcript of a protein-coding gene, either by causing a frameshift or through the introduction of a premature termination codon.

Untranslated region
(UTR). Non-coding sequence on coding sequence transcripts found between the transcription start site and the translation initiation site (5′ UTR), and the STOP codon and polyA site (3′ UTR).

content of these transcripts is not always coupled to their functionality in an obvious way. For example, the *AIRN* lncRNA regulates the activity of the insulin-like growth factor 2 receptor (*IGF2R*) locus on the opposite strand not through the activity of its transcript — which is apparently a by-product — but through the act of its transcription[93]. This emerging perspective on functionality represents a paradigm shift for annotation projects (BOX 2).

It is difficult to infer lncRNA functionality through annotation alone; true understanding comes from the laboratory. Nonetheless, annotation does have an important role in judging translation, and most lncRNA models within genebuilds (or those generated by pipelines such as PLAR[49]) are simply transcripts that are not protein-coding, pseudogenes or small RNAs. It may also be useful to sub-classify models on the basis of their genomic location[94]. This could aid scientists in investigating particular lncRNA categories; enhancer-associated e-lncRNAs, for example, are of interest in the field of regulatory genomics[95], as is the bidirectional transcription that is commonly observed from protein-coding gene promoters[96]. However, lncRNA functional annotation may become more proactive: sequences such as microRNA binding sites[97] and RNA structures[98] are beginning to be described, and ultraviolet (UV) crosslinking immunoprecipitation followed by sequencing (CLIP-seq) can identify RNAs interacting with RNA-binding proteins[99]. In the meantime, genebuilds can incorporate laboratory-gained knowledge of functionality. LncRNAdb is a database that is attempting to catalogue functional lncRNAs on the basis of literature curation[100]. It currently contains entries for 287 lncRNAs from a variety of eukaryotic species. Other repositories seek to build larger consolidated lncRNA catalogues, including LNCipedia[101], which focuses on human lncRNAs, and NONCODE[102], which contains information from 16 species. Meanwhile, the RNAcentral database[103] contains 8.1 million RNA sequences, representing all major functional classes of non-coding RNAs from a selection of species; a key goal is to resolve the redundancy between the lncRNA data sets produced from different annotation groups.

Annotating the extended gene

Human genetics is faced with a substantial problem: trait-associated variants are commonly found outside gene sequences and thus defy interpretation. Gene annotation projects are, therefore, turning their attention to the genomic elements that control gene activity, the best studied of which are promoters and enhancers. Both are controlled by transcription factor binding, and each has its own characteristic (albeit imprecisely understood) epigenomic profile. ENCODE especially has provided enormous data sets on these sequences, mostly through the use of immunoprecipitation techniques[104] (BOX 1). Furthermore, it has been known for decades that chromosomes exhibit 'loops', which can indicate enhancer–promoter interactions. Modern assays such as Hi-C[105] and chromatin interaction analysis paired-end tag sequencing (ChIA-PET)[106] capture the DNA fragments that flank these loops, allowing them to be mapped onto the genome as topologically associated domains (TADs).

Such data sets offer the potential to create 'extended gene' models, as illustrated for nuclear receptor interacting protein 1 (*NRIP1*) in FIG. 4. From a human perspective, an obvious benefit of linking genes to regulatory elements is that it increases the space within which disease-associated variants can be interpreted, although it should be emphasized that such efforts are in their infancy. One problem is that Hi-C and ChIA-PET highlight enormous numbers of TADs (even when 'capture' methods are used to target known promoters[107]), raising questions about the signal to noise ratio[105,106]. This noise seems to be biological as well as artefactual[108], and it is unclear what proportion of genuine TADs actually demarcate enhancers. Currently, the ENCODE enhancer sets — extrapolated from biochemical data[104] — are far larger than those that have been functionally validated in the laboratory[109]. The fact that gene regulation is spatiotemporal complicates the situation, and it is known that genes can be controlled by multiple enhancers, and enhancers can control multiple genes[110]. Extended genes would be more useful if they could also integrate the transcription factor-binding sites that are found within enhancers and promoters. Transcription factor annotation has traditionally proved difficult: binding motifs are typically short (~6 bp) and imprecise, thwarting genome-mining efforts[111]. However, chromatin immunoprecipitation followed by sequencing (ChIP-seq) data sets are now available for dozens of transcription factors, highlighting *in vivo* regions of DNA occupancy while allowing for more accurate consensus motifs to be deduced[112]. If such information can be combined with chromosome conformation and chromatin immunoprecipitation data sets, more precise extended genes may be obtained. This has been well demonstrated for CTCF, a factor that is known to have a key role in loop formation[113,114]. The challenge for genebuilds is how to integrate and display such data alongside their transcript models. A description of extended genes is a core goal of the developing ENCODE encyclopaedia resource, and tools to visualize three-dimensional (3D) data sets on the genome, such as the 3D Genome Browser, are becoming available[112].

Improving the usability of genebuilds

The incorporation of transcript expression data. As genebuilds provide more precise representations of the transcriptome, they inevitably become more complex. This point has important repercussions for users. For example, many scientists are focused on human *BRCA1* owing to its association with breast cancer, and may wonder what to make of the fact that GENCODE has 30 transcript models of this gene, whereas RefSeq has six. In practice, human annotation resources especially are used in many different ways. Whereas some scientists would like to use all the transcripts that are associated with a given gene — for example, when designing hypothesis-driven experiments within a single locus — a common desire is for simplification. In fact, users

Enhancer
Sequence that regulates a promoter from a distal site on the chromosome, probably brought into close proximity through DNA looping.

Promoters
Regions immediately upstream of the transcription start site where the RNA polymerase complex attaches in order to initiate transcription.

often wish to work with a single transcript model per gene in order to streamline the experimental design of whole-transcriptome studies. One way to achieve 'transcript prioritization' is by measuring RNA expression, that is, to identify the 'dominant' transcript in a gene. Although it remains challenging to resolve individual transcripts based on RNA-seq, the fact that most human protein-coding genes have a dominant transcript indicates that there is a value to expression-based filtering[115]. However, dominance can 'switch' between cell types, and expression changes are typically analogue rather than simply 'on' or 'off' (REFS 115–117). An additional point of considerable importance is that RNA is ultimately a proxy for the measurement of protein output within protein-coding genes. In reality, the relationship between RNA and protein output remains imprecisely understood[118], and correlations between the two are frequently weak[119]. Although this may have striking consequences for RNA-based expression studies, the maturing field of quantitative proteomics does not yet provide precise guidance for annotation projects.

The prioritization of functional transcripts. Currently, reference genebuilds do not explicitly highlight principal transcripts based on quantitative evidence, and the description of spatiotemporal expression comes instead from 'downstream' endeavours such as the Genotype–Tissue Expression (GTEx) project[116]. One could anticipate that such information will soon be leveraged in reference genebuilds, perhaps influencing how models are displayed in genome browsers. Nonetheless, there is a limit to what expression data alone can tell us about transcript functionality (BOX 2): transcripts with lower expression are not necessarily non-functional (or even less functional), and the expression of numerous genes in fact seems to be dominated by NPTs[115]. When it comes to genebuild usability, it is this question of functionality that is of paramount importance, most obviously when it comes to CDS annotation. For example, it is important to predict the molecular and clinical consequences of variation within *BRCA1*, and the processing of variant data sets typically begins with a comparison against gene annotation[120]. This allows variants to be stratified according to their potential mechanistic consequences, such as whether they disrupt a CDS or fall within an intron.

Clearly, there is a close relationship between the quality of gene annotation and the accuracy of variant interpretation, and yet many aspects of annotation — especially functional annotation — remain putative. This is particularly true for genebuilds such as GENCODE, which attempt to annotate all transcripts. Putative functional annotation can introduce false positives into variant-calling workflows; for example, where LoF mutations are called in CDS exons that are not in reality coding. GENCODE attempts to reduce this problem by providing 'Basic': a ~50% reduced build in comparison to the 'Comprehensive' set, resulting especially from the removal of models with truncated CDSs. As discussed above, GENCODE also uses APPRIS to highlight coding models of probable functionality based on conservation[59]. By contrast, the use of smaller genebuilds could introduce false negatives into variant analyses, that is, where consequential variants are missed or misinterpreted because they fall outside the gene annotation. However, RefSeq allows users of its core genebuilds to work with sets of more prospective transcripts, in the form of its uncurated (XM) '*in silico*' models. Finally, Ensembl and the National Center for Biotechnology Information (NCBI) are collaborating on the Locus Reference Genomic (LRG) project[121]. The remit of this work is to standardize the gene annotation that is used in the clinic, with a key aim being to select a set of transcript models for core disease genes. These models are manually selected, in order to include what seem to be the key functional elements for a given gene.

Conclusions

The complexity of gene annotation projects reflects the complexity that exists in eukaryotic cells, and, as we do not currently fully understand the transcriptome, all of our genebuilds are incomplete. Current ambiguities are most keenly felt in our own species, where nothing less than a total understanding of biology is demanded. For other projects, the 'finish line' may not be so far into the distance, and the length of journey taken will in many ways reflect the value of that genome to science. However, all genebuilds face challenges in how they present their resource to the public; most obviously, they must find ways to make sure that increasing complexity does not correlate with decreasing usability.

1. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17**, 669–681 (2007).
This influential article attempts to rationalize a modern description of the gene in the context of transcriptional complexity.
2. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
This provides a detailed description of the GENCODE annotation pipeline.
3. Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* **10**, 126–139 (2009).
4. Andersson, L. *et al.* Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
5. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
This is an excellent starting point for exploring the NCBI annotation resources.
6. McGarvey, K. M. *et al.* Mouse genome annotation by the RefSeq project. *Mamm. Genome* **26**, 379–390 (2015).
7. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm. Genome* **26**, 366–378 (2015).
8. Berardini, T. Z. *et al.* The *Arabidopsis* information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**, 474–485 (2015).
9. Howe, K. L. *et al.* WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res.* **44**, D774–D780 (2016).
10. Attrill, H. *et al.* FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Res.* **44**, D786–D792 (2016).
11. Elsik, C. G. *et al.* Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* **15**, 86 (2014).
12. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
This provides a detailed description and comparison of various RNA-seq analytical pipelines.
13. Boutet, E. *et al.* UniProtKB/Swiss-prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.* **1374**, 23–54 (2016).
The UniProt and Swiss-Prot resources are outlined here.
14. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** (Suppl.2), ii215–ii225 (2003).

15. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
16. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).
17. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* **43**, D1079–D1085 (2015).
18. Guigo, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7**, S2 (2006).
19. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
20. Eory, L. *et al.* Avianbase: a community resource for bird genomics. *Genome Biol.* **16**, 21 (2015).
21. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-Seq-Based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
22. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644 (2008).
23. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
24. Loveland, J. E., Gilbert, J. G., Griffiths, E. & Harrow, J. L. Community gene annotation in practice. *Database (Oxford)* **2012**, bas009 (2012).
25. Pennisi, E. Ideas fly at gene-finding jamboree. *Science* **287**, 2182–2184 (2000).
26. Archibald, A. L. *et al.* Pig genome sequence—analysis and publication strategy. *BMC Genomics* **11**, 438 (2010).
27. Lee, E. *et al.* Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* **14**, R93 (2013).
28. Giraldo-Calderon, G. I. *et al.* VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**, D707–D713 (2015).
29. Dawson, H. D. *et al.* Structural and functional annotation of the porcine immunome. *BMC Genomics* **14**, 332 (2013).
30. The UK 10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
31. Guo, L., Gao, Z. & Qian, Q. Application of resequencing to rice genomics, functional genomics and evolutionary analysis. *Rice (N.Y.)* **7**, 4 (2014).
32. Foote, A. D. *et al.* Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat. Commun.* **7**, 11693 (2016).
33. Adams, D. J., Doran, A. G., Lilue, J. & Keane, T. M. The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mamm. Genome* **26**, 403–412 (2015).
34. Baker, M. Structural variation: the genome's hidden architecture. *Nat. Methods* **9**, 133–137 (2012).
35. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
36. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genom. Hum. Genet.* **14**, 301–323 (2013).
37. Hirayasu, K. & Arase, H. Functional and genetic diversity of leukocyte immunoglobulin-like receptor and implication for disease associations. *J. Hum. Genet.* **60**, 703–708 (2015).
38. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
In this study, thousands of human RNA-seq libraries are combined to generate almost 60,000 putative lncRNA genes.
39. Filichkin, S. A. *et al.* Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20**, 45–58 (2010).
40. Mudge, J. M., Frankish, A. & Harrow, J. Functional transcriptomics in the post-ENCODE era. *Genome Res.* **23**, 1961–1973 (2013).
41. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
42. Cho, H. *et al.* High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS ONE* **9**, e108095 (2014).
43. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742 (2015).
44. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9**, 989–1009 (2014).
45. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
46. The FANTOM Consortium *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
The leading publication of the FANTOM5 project, providing detailed analysis of hundreds of human and mouse CAGE experiments.
47. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183 (2012).
48. Boley, N. *et al.* Genome-guided transcript assembly by integrative analysis of RNA sequence data. *Nat. Biotechnol.* **32**, 341–346 (2014).
49. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).
50. Sisu, C. *et al.* Comparative analysis of pseudogenes across three phyla. *Proc. Natl Acad. Sci. USA* **111**, 13361–13366 (2014).
51. Frankish, A. & Harrow, J. GENCODE pseudogenes. *Methods Mol. Biol.* **1167**, 129–155 (2014).
52. Carelli, F. N. *et al.* The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**, 301–314 (2016).
53. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006).
54. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13**, R51 (2012).
55. Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 1–30 (2013).
56. Yang, X. *et al.* Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805–817 (2016).
57. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* **6**, e1001236 (2010).
58. Hao, Y. *et al.* Semi-supervised learning predicts approximately one third of the alternative splicing isoforms as functional proteins. *Cell Rep.* **12**, 183–189 (2015).
59. Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–D117 (2013).
60. Farrell, C. M. *et al.* Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.* **42**, D865–D872 (2014).
61. Bassett, A. R. *et al.* Considerations when investigating lncRNA function *in vivo*. *eLife* **3**, e03058 (2014).
62. Derrien, T., Guigo, R. & Johnson, R. The long non-coding RNAs: a new (P)layer in the "dark matter". *Front. Genet.* **2**, 107 (2011).
63. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* **9**, e1003569 (2013).
64. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
65. Peccarelli, M. & Kebaara, B. W. Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway. *Eukaryot. Cell* **13**, 1126–1135 (2014).
66. Lareau, L. F. & Brenner, S. E. Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol. Biol. Evol.* **32**, 1072–1079 (2015).
67. Wong, J. J. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583–595 (2013).
68. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24**, 1774–1786 (2014).
Demonstrates that intron retention affects three-quarters of mammalian genes, and suggests widespread involvement in gene regulation.
69. Capell, A., Fellerer, K. & Haass, C. Progranulin transcripts with short and long 5′ untranslated regions (UTRs) are differentially expressed via posttranscriptional and translational repression. *J. Biol. Chem.* **289**, 25879–25889 (2014).
70. Barbosa, C., Peixeiro, I. & Romao, L. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* **9**, e1003529 (2013).
71. Yeh, H. S. & Yong, J. Alternative polyadenylation of mRNAs: 3′-untranslated region matters in gene expression. *Mol. Cells* **39**, 281–285 (2016).
72. Barrett, L. W., Fletcher, S. & Wilton, S. D. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* **69**, 3613–3634 (2012).
73. Mudge, J. M. *et al.* The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.* **28**, 2949–2959 (2011).
74. Barash, Y. & Garcia, J. V. Predicting alternative splicing. *Methods Mol. Biol.* **1126**, 411–423 (2014).
75. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).
An obvious starting point to explore strategies for the analysis of mass-spectrometry data in genomics.
76. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
77. Wilming, L. G. *et al.* The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* **36**, D753–D760 (2008).
78. Ezkurdia, I., Vazquez, J., Valencia, A. & Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* (2014).
79. Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* **7**, 11778 (2016).
80. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
81. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213 (2014).
82. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
83. Jackson, R. & Standart, N. The awesome power of ribosome profiling. *RNA* **21**, 652–654 (2015).
84. Ingolia, N. T. Ribosome footprint profiling of translation throughout the genome. *Cell* **165**, 22–33 (2016).
A primer on the use of RP from one of the key developers of the technique.
85. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* **5**, e13328 (2016).
86. Mumtaz, M. A. & Couso, J. P. Ribosomal profiling adds new coding sequences to the proteome. *Biochem. Soc. Trans.* **43**, 1271–1276 (2015).
87. Graur, D. *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* **5**, 578–590 (2013).
88. Xie, S. Q. *et al.* RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.* **44**, D254–D258 (2016).
89. Goff, L. A. & Rinn, J. L. Linking RNA biology to lncRNAs. *Genome Res.* **25**, 1456–1465 (2015).
90. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front. Genet.* **6**, 2 (2015).
91. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8**, e1002841 (2012).
92. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **17**, 601–614 (2016).
93. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002).
94. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
95. Lai, F. & Shiekhattar, R. Enhancer RNAs: the new molecules of transcription. *Curr. Opin. Genet. Dev.* **25**, 38–42 (2014).
96. Scruggs, B. S. *et al.* Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
97. Furio-Tari, P., Tarazona, S., Gabaldon, T. & Enright, A. J. & Conesa, A. spongeScan: A web for detecting microRNA binding elements in lncRNA sequences. *Nucleic Acids Res.* (2016).

98. Novikova, I. V., Hennelly, S. P. & Sanbonmatsu, K. Y. Tackling structures of long noncoding RNAs. *Int. J. Mol. Sci.* **14**, 23672–23684 (2013).
99. Konig, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* **13**, 77–83 (2011).
100. Quek, X. C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **43**, D168–D173 (2015).
101. Volders, P. J. *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* **43**, 4363–4364 (2015).
102. Zhao, Y. *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **44**, D203–D208 (2016).
103. RNAcentral Consortium. RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res.* **43**, D123–D129 (2015).
104. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
105. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
106. Fullwood, M. J. & Ruan, Y. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell Biochem.* **107**, 30–39 (2009).
107. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
This study uses Capture Hi-C to examine the long-range chromosome interactions of 22,000 human promoters.
108. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
109. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).
110. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
111. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
112. Zerbino, D. R. *et al.* Ensembl regulation resources. *Database (Oxford)* **2016**, 1–13 (2016).
113. de Wit, E. *et al.* CTCF binding polarity determines chromatin looping. *Mol. Cell* **60**, 676–684 (2015).
114. Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
115. Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14**, R70 (2013).
116. The GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
117. Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
118. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
119. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
120. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
121. Dalgleish, R. *et al.* Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med.* **2**, 24 (2010).
This project provides insights into the relationship between gene annotation and the description of variation in the clinic.
122. Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods Mol. Biol.* **786**, 181–200 (2012).
123. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
124. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
125. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
126. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
127. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2012).
128. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192 (2013).
129. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
130. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
131. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
132. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
133. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
134. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
135. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).

Acknowledgements
The work performed by J.M.M. and J.H. on the GENCODE project is supported by the National Human Genome Research Institute of the National Institutes of Health (grant number U41 HG007234). The authors thank A. Frankish for informative discussions.

Competing interests statement
The authors declare no competing interests.

DATABASES
UniProt: http://www.uniprot.org/
WormBase: http://www.wormbase.org/#012-34-5

FURTHER INFORMATION
3D Genome Browser: http://promoter.bx.psu.edu/hi-c
ENCODE: http://www.encodeproject.org
The Genome Reference Consortium: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc
The NCBI Eukaryotic Genome Annotation Pipeline: http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

CORRIGENDUM

Organization and function of the 3D genome

Boyan Bonev and Giacomo Cavalli

Nature Reviews Genetics **17**, 661–678 (2016)

In the original version of this article, the statement that CCCTC-binding factor (CTCF) is conserved in most bilaterians was incorrectly referenced. Reference 58 has now been corrected in the online version of the article to cite Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. & Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc. Natl Acad. Sci. USA* **109**, 17507–17512 (2012). The authors apologize for this error.