nature reviews genetics

https://doi.org/10.1038/s41576-026-00937-3

Review article

Check for updates

Annotating genomes at increased scale and resolution

Hyun Joo Ji[1,2], Mihaela Pertea[1,2,3] & Steven L. Salzberg[1,2,3,4]

Abstract

Genome annotation captures the essence of a genome by cataloguing its genes, transcripts, proteins and other functional elements of the DNA sequence. Accurate annotation serves as the foundation for a wide range of downstream analyses and discoveries, ranging from basic biology to an understanding of the linkage between genes and disease. Over the past two decades, advances in high-throughput sequencing techniques have enabled faster and more accurate capture of diverse genomic features, generating data at an unprecedented scale. Concurrently, computational methods for translating these data into evidence for genome annotation have steadily improved, leading to better automated genome annotation systems. As such, the growing number of sequenced genomes provides a positive feedback loop, in which database searches become more effective and shared sequence patterns emerge more clearly. These advances are promising steps towards annotating the functions of many poorly understood genes, particularly non-coding RNA genes, for which more research is needed.

Sections

Introduction

Gene structure annotation

Functional annotation

Annotation of newly assembled genomes

Conclusions and future challenges

[1]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA.
e-mail: salzberg@jhu.edu

Introduction

The completed human genome sequence announced in the year 2000 was hailed as a breakthrough that would "revolutionize the diagnosis, prevention and treatment of most human diseases"[1]. However, scientists recognized that a genome sequence alone was just the starting point, and even the US President acknowledged that the next steps were to "sort through this trove of genomic data to identify every human gene … [and to] … discover the function of these genes and their protein products"[1].

Annotation turned out to be harder than expected: 25 years later, the scientific community is still refining the human gene catalogue, with multiple competing annotations that have not yet converged[2]. The genomes of thousands of additional animals, plants and other species have also been sequenced[3–8], and annotating all these additional genomes remains a major scientific challenge. Early efforts to keep up with all these data recognized the need for automated genome annotation, which led to the development of methods that focused primarily on protein-coding genes and relied heavily on ab initio gene finding[9–13] and protein-sequence alignment[14]. However, even the best of these early methods had high rates of false positives, could not identify precise exon–intron structures for many genes and were unable to detect non-coding genes (Box 1).

Recent advances in experimental and computational methods are providing new types of evidence to guide and improve genome annotation. For example, newer high-throughput sequencing assays combined with computational tools indicate which portions of a DNA sequence are transcribed[15–18], where the resulting RNA transcripts are spliced[19–27], and what portions of the RNA are translated into proteins[28]. Furthermore, recent advances in machine learning algorithms, combined with ever-growing amounts of available training data, have produced major breakthroughs in protein structure prediction[29,30], which can be used to enhance annotation[31]. Beyond protein-coding annotation, new methods have also been developed to detect and annotate non-coding RNAs[32–35], including transfer RNAs[36] (tRNAs) and ribosomal RNAs (rRNAs) (both of which were captured in the early days of genomics), but now also long non-coding RNAs (lncRNAs), small nuclear RNAs (snRNAs), small interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), Piwi-interacting RNAs (piRNAs), microRNAs (miRNAs) and others. Finally, the rapid growth in the number of sequenced genomes, some of which have been annotated more thoroughly (for example, *Homo sapiens*[37–39], *Mus musculus*[40], *Drosophila melanogaster*[41] and *Arabidopsis thaliana*[42,43]), has also led to the emergence of the powerful liftover approach. In this method, genes and other features from a source genome are aligned directly to a target genome, and the annotation is transferred over.

In this Review, we highlight recent progress in the sequencing and computational techniques that guide genome annotation. Initially, we cover tools for annotating gene structure, including approaches that leverage various types of RNA-sequencing (RNA-seq) data, other high-throughput sequencing data types or deep learning-based modelling. Next, we discuss how to annotate the function of protein-coding and non-coding genes, as well as pseudogenes. Finally, given the increasing amounts of data, we overview tools for automated gene annotation and contrast the strengths of these methods with manually curated catalogues. Overall, we focus on emerging tools, evidence-based strategies and remaining challenges in the annotation of large, complex eukaryotic genomes.

Gene structure annotation

Genome annotation requires the identification of various types of genomic elements, of which protein-coding genes have received the most attention. Protein-coding gene annotation involves the

Box 1 | Early efforts at genome annotation

Early genome annotation used statistical and machine learning models to identify protein-coding sequences, translational start sites and intron-splicing boundaries. Although these methods were somewhat effective at locating portions of protein-coding genes, they performed poorly at identifying the precise exon–intron structure of most transcripts, and they could not identify the untranslated regions for any genes. Ab initio gene finders also required specialized training for each new species, which in turn required creating a high-quality set of training data, which was not always possible. The most successful of these methods[9–13] used hidden Markov models, in which a DNA sequence was modelled as a series of nucleotides output from hidden states. The DNA within coding exons was constrained to remain in-frame by creating three parallel models for each of the reading frames. Donor and acceptor splice sites were modelled within the intronic regions, usually as Markov chains, to enforce the constraint that introns begin and end with GT and AG, respectively, and to capture the conserved pattern of sequences near each splice site.

Independently of ab initio gene finding, the use of protein sequence alignment has long been the most reliable means of identifying genes. The BLAST program[14] can translate a DNA query in all six frames and compare it with a pre-compiled database of protein sequences. Given that protein sequences are conserved over much longer time spans than DNA sequences, homologous matches found through BLAST searches can provide strong evidence that a sequence encodes a protein-coding gene, even when the query sequence is evolutionarily distant from the homologue. This method has become more effective as protein sequence databases have expanded: the public database UniProt has grown exponentially since its creation in 1987, containing over 252 million entries as of 2025[114]. However, when a gene is spread into multiple exons separated by large introns, as is common in many eukaryotic genomes, it can be difficult to annotate it precisely using BLAST alone.

The first annotation pipelines, dating from the mid-2000s, combined ab initio predictions and protein alignment with limited sequencing evidence from expressed sequence tags[206], an early version of RNA sequencing. These pipelines included the automated annotation system within Ensembl[201] and the Gnomon system developed by NCBI (National Center for Biotechnology Information)[167], as well as standalone systems including JIGSAW[207] and MAKER[202]. Pipelines continue to be widely used for automated annotation, but modern ones can draw upon far larger protein sequence databases, vast RNA-sequencing data resources, and other types of annotation not available prior to the invention of next-generation sequencing.

identification, for each transcript, of the transcription start and end sites, translation initiation and termination codons, and all intron splice sites (donors on the 5′ end and acceptors on the 3′ end of the introns). After protein-coding gene annotation, two classes of RNA genes are expected in any annotation: rRNAs and tRNAs, the structure and function of which are well known[36]. Some annotations also include lncRNAs, which may number in the thousands for some species[44], although much less is known about their functions. Other classes of genes that might be annotated but that are not part of standard annotation pipelines include miRNAs, snRNAs, snoRNAs, siRNAs and piRNAs, which play distinct roles in RNA splicing, gene regulation and other biological functions.

Advances in RNA-seq technologies for transcript structure annotation

The emergence of RNA-seq in 2008[15–17] revealed that most human genes contain at least one alternative splicing event[45]. By capturing RNA molecules that have undergone splicing, RNA-seq experiments offer direct evidence of transcript boundaries and splice sites. Analysis of RNA-seq data typically involves aligning raw sequencing reads to a reference genome using a spliced aligner such as HISAT2[46] or STAR[47], followed by transcript reconstruction using assemblers such as StringTie[48–52] or Scallop[51,52]. This approach has led to a large increase in the number of annotated transcripts for many well-studied species. For example, in contrast to the number of protein-coding genes in the current human genome annotation, which has remained relatively stable at just under 20,000, the number of transcript isoforms has grown substantially, ranging from 99,201 in CHESS[38], to 129,740 in RefSeq[39] and 211,446 in GENCODE (v49)[37]. Similar expansions in transcript isoform annotations have been observed in highly-studied species such as mouse[37] and *A. thaliana*[42,43,53]. For less well-studied species without a reference genome assembly, de novo transcriptome assemblers such as Trinity[54] have been used to reconstruct and annotate transcripts directly from RNA-seq data.

These advances in genome annotation were provided by RNA-seq data typically generated on Illumina platforms producing relatively short reads ranging from 75 to 150 base pairs with high base-level accuracy. However, such short reads rarely span more than two exons and therefore have limited ability to reconstruct isoforms. By contrast, long reads generated by instruments from Oxford Nanopore Technologies and Pacific Biosciences cover larger portions of RNA molecules, enabling more accurate identification of full-length transcripts. However, long-read RNA-seq generally has higher error rates and lower per-base coverage (than short-read RNA-seq) and can vary across sequencing platforms and protocols[55–57], which has motivated the development of updated aligners[47,58,59] and improved transcriptome assemblers[49,60–63]. Hybrid approaches combining both short-read and long-read RNA-seq can produce more accurate transcript annotations[64], in which long reads are used to determine overall intron chain structure, and short reads are used to correct potential sequencing errors and refine splice junction locations[50,65].

Single-cell RNA-seq (scRNA-seq) enables transcriptome profiling at the cellular level[66–70], and rich repositories of such data sets are being produced through initiatives including HubMap[71] and the Human Cell Atlas[72]. Notably, although scRNA-seq data are currently used primarily for gene expression analysis, transcript assembly from scRNA-seq holds the potential to identify and annotate cell type-specific isoforms. However, most scRNA-seq protocols rely on short-read sequencing technologies, which have some of the same limitations inherent to bulk RNA-seq experiments, plus additional challenges unique to single-cell data, including low sensitivity, transcript-end bias, high dropout rates and sparse coverage[73]. Some of these issues are being addressed with the development of several transcript assembly tools, such as RNA-Bloom, scRNAss and Beaver[74–76]. Alternatively, the use of long-read sequencing technologies in scRNA-seq protocols such as Smart-seq2[77] and Smart-seq3[78] has shown promising results, as several studies have successfully used long-read scRNA-seq data to identify isoforms specific to certain cell types or disease conditions[79–83]. Notably, however, computational tools for accurate transcriptome reconstruction from long-read scRNA-seq are still under active development[84,85], and long-read and scRNA-seq are currently much more expensive than bulk RNA-seq methods, which limits the accessibility of the former for some projects.

Additional technologies for refining transcript features

Advances in RNA-seq technologies have improved our ability to capture exon–intron structures, but RNA-seq alone often fails to capture the full transcript. More specifically, coverage is often reduced near the transcription initiation and termination sites, which limits annotation ability at these important regions. As such, additional experimental evidence — from cap analysis of gene expression (CAGE)[18], 3′ end sequencing and ribosome profiling (Ribo-seq)[28] — can be used to refine certain transcript features.

The CAGE experimental protocol targets the 5′ cap sites of transcripts and generates short tags (20–30 nucleotides) that correspond to the 5′ ends of transcripts. CAGE sequencing reads can then be mapped back to a reference genome using an aligner such as STAR[47]. The resulting alignments can be clustered together and used to infer transcription start sites[86,87]. To further refine these regions, statistical models have been developed to separate overlapping CAGE signals into distinct, non-overlapping initiation units[88], which sometimes indicate multiple transcription start sites for the same gene. CAGE data are already integrated into some annotation databases, including FANTOM[88] and ENCODE[89] for human and mouse.

To define the other (3′) end of transcripts, several protocols have been developed to identify transcript termination and polyadenylation sites. Such methods include direct RNA sequencing[90], 3′ region extraction and deep sequencing (3SEQ)[91], poly(A) site sequencing (PAS-Seq)[92] and others[93]. Most of these approaches target the poly(A) tails of mature transcripts, typically using oligo(dT) priming. The rapid expansion of 3′ end sequencing methodologies has generated a substantial volume of data for transcriptome-wide profiling of polyadenylation sites. However, these protocols have limitations, and several computational strategies[93–95] have been developed to ensure greater consistency in poly(A) site identification. These computational tools typically involve aligning reads to a reference genome and clustering adjacent sites to define high-confidence poly(A) sites, also referred to as cleavage sites. Internal priming, which can occur for any sufficiently long sequence of adenines naturally within a transcribed gene, requires customized processing pipelines for its removal[96]. The APASdb[97] and PolyASite[96,98,99] databases provide curated repositories of predicted poly(A) sites derived from a wide range of experimental protocols and analysis pipelines.

Ribo-seq enables genome-wide analysis of translation by sequencing ribosome-protected mRNA fragments, thereby capturing regions of mRNAs potentially engaged in active translation[28]. A range of computational tools have been developed to analyse Ribo-seq data[100–103] and identify translated regions of the genome. These tools include PRICE[103], which uses expectation-maximization algorithms; SPECtre[101],

which takes advantage of the trinucleotide periodicity characteristic of translated sequences; and ORF-Rater[102], which integrates linear regression with random forest classification. Translated regions are typically called either open reading frames or coding DNA sequences; however, these terms do not capture precisely what Ribo-seq experiments measure. More recently, the unifying term 'translon' has been proposed to describe any region that is translated by the ribosome, regardless of whether or not it produces a functional protein[104].

Recent Ribo-seq studies across diverse model organisms, including human, mouse and zebrafish, have uncovered thousands of putatively novel translons, most of which are very short. Separate efforts have focused on identifying translons that had orthogonal evidence supporting them, such as evolutionary conservation[105] and mass spectrometry data generated from both conventional methods[106] and immunopeptidomics[107,108]. Some novel translated regions have been recently integrated into reference annotations, most notably by the GENCODE consortium[37,109]. Other efforts have used Ribo-seq combined with RNA-seq evidence to discover alternative translation initiation sites and novel protein-coding exons that represent new isoforms of existing genes[110,111].

Deep learning for identifying splice sites

Splice site prediction has long been a fundamental challenge in gene annotation, predating the recent surge of interest in deep learning approaches[112,113]. Accurately identifying splice sites from genomic DNA is essential for reconstructing the exon–intron structure of transcripts, especially in cases in which RNA-seq-based splicing evidence is limited or ambiguous. Early tools implemented traditional machine learning techniques such as hidden Markov models (HMMs) (Fig. 1), whereas more recent tools (for example, SpliceAI[19], Pangolin[20] and Splam[21]) have incorporated deep learning models with improved accuracy on selected species.

Machine learning models typically predict splice sites within a window of genomic sequence, returning a score that reflects the confidence that a position is a donor or acceptor splice site. A variety of deep learning architectures have been explored for this task, including convolutional neural networks[22,23] and transformer-based models[24]. These models have been trained and evaluated using annotated splice sites from human reference annotation databases such as GENCODE, and their accuracy has been measured by performance on splice sites that were not included in the training data. The overall training and evaluation frameworks are broadly similar across splice site predictors, but notable variation in performance can arise from differences in how the problem is framed or how the data are curated. For example, one of the first deep learning models to predict splicing, SpliceAI, only predicts a single splice isoform for each protein-coding gene. By contrast, Splam evaluates pairs of donor–acceptor sites and has no limit on how many splice isoforms it can predict at a single gene locus. Another factor that distinguishes these models is the size of the input sequence window, with SpliceAI using a window of more than ten kilobases and Splam using two 400-base pair windows centred on the donor and acceptor sites.

Machine learning models for splice site identification are increasingly used to augment genome annotation pipelines. For example, SICILIAN[25] uses a logistic regression model to assign statistical confidence to junctions in scRNA-seq spliced alignments, enabling detection of cell-type-specific junctions missed in bulk RNA-seq experiments. Splice prediction models are also being used to identify sequence

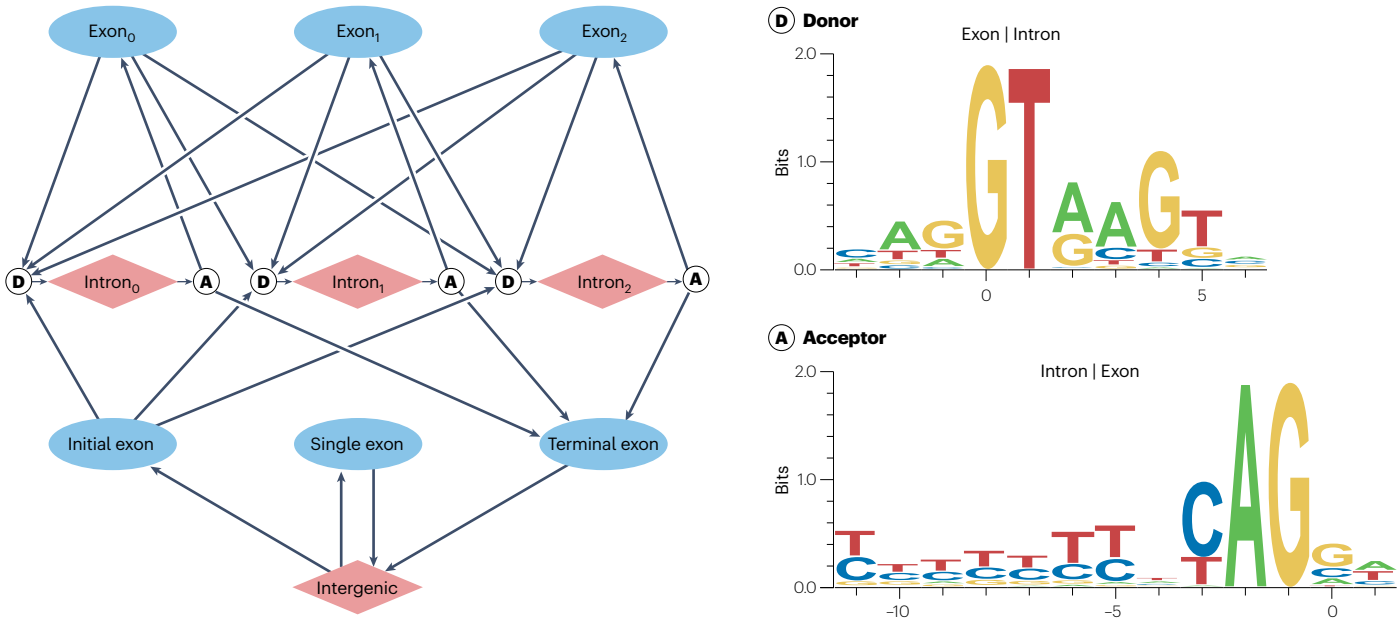

Fig. 1 | Predicting splice sites with deep learning. A generalized hidden Markov model (HMM) to predict protein-coding exons, with states representing exons and introns in each of the three possible reading frames, based on GlimmerHMM[13]. Transitions from exons to introns go through a donor site (D) Markov chain, shown schematically as a sequence logo[197]. The initial exon is constrained to begin with a start codon (ATG), and the terminal exon ends with a stop codon (TAA, TAG or TGA). Transitions from introns to exons go through an acceptor site (A) model. The use of three intron states and three internal exon states allow the model to keep track of the reading frame; for example, if an intron begins on a codon boundary (frame 0), then the following exon must begin in frame 0. When processing a DNA sequence, the HMM algorithm finds an optimal path through the model, which corresponds to a predicted protein-coding sequence.

variants that are likely to have adverse effects on splicing[19,20,26,27], although the accuracy of these predictions is currently hard to evaluate.

Functional annotation

Once the sequence and genomic structure of a gene have been determined, the next step in annotation is to describe the function of the gene. Functions can be described at multiple levels, ranging from the molecular role of a protein within the cell to its effect upon observable phenotypes.

Homology-based functional annotation of coding sequences

For many years, the functional annotation of protein-coding regions has relied on the use of BLAST[14] to align predicted proteins against databases of known proteins, such as UniProt[114] and the NCBI (National Center for Biotechnology Information) database[39]. This approach is based on the assumption that functional proteins are likely to be conserved across multiple species, which has led to sequence-level conservation scores[105] often being integrated into the annotation process. As these protein databases have grown over time, the use of alignment has become more effective simply because so many proteins are now known, which means that an ever-growing proportion of newly annotated genes will have a match in the database. Aligning proteins against these large databases provides two distinct pieces of information about the function of a gene. First, one can infer from sequence conservation that a gene is indeed functional; that is, if a sequence has been conserved in two species that diverged tens of millions of years ago, the conservation is almost certainly due to evolutionary selection[115]. Therefore, for the annotation task of identifying most protein-coding genes, sequence alignment provides very strong evidence. Second, when a query protein sequence closely matches a known entry in a reference database, then the query protein likely has a similar function. Most annotation pipelines use this assumption to assign a name to a gene; for example, if a query gene matches the human keratin gene *KRT1*, it might be given the name '*KRT1*', or, alternatively, it might be called '*KRT1*-like' or simply be annotated as 'similar to human *KRT1*'.

Inferring function and naming genes based on homology is widely used and practical, but sequence similarity does not necessarily imply functional equivalence, as even proteins with high sequence identity can perform different functions. For example, the yeast proteins Gal1 and Gal3 share more than 70% amino acid identity but differ in function: Gal1 acts as a galactokinase whereas Gal3 serves as a transcriptional inducer[116]. This problem of high sequence similarity masking functional divergence can be particularly acute in large gene families with many paralogues, such as the keratin gene family. Human, canine and equine genomes all contain multiple keratin paralogues with overlapping sequence but specialized, tissue-specific roles. Some keratin paralogues are even pseudogenized, such as equine *KRT9P*, which appears functional in dogs and humans but has lost activity in horses owing to a frameshift mutation[117]. Therefore, relying solely on sequence-based homology risks propagating inaccurate or overly generic annotations, particularly in gene families in which subtle sequence differences reflect important functional divergence.

Alternative approaches to functional annotation of coding sequences

Although homology-based annotation remains the primary strategy for assigning function to protein-coding genes, a growing number of alternative approaches have been developed. One such approach is co-expression analysis, which infers gene function through 'guilt by association'[118]. By examining coordinated expression patterns across diverse tissues, developmental stages or environmental conditions, researchers can cluster genes into functional modules. Methods such as Weighted Gene Co-expression Network Analysis[119] enable the construction of co-expression networks that group genes into modules enriched for known biological functions, offering clues to the function of uncharacterized genes that share similar expression profiles. Another more direct approach to identify essential genes involved in specific biological pathways involves high-throughput loss-of-function experiments, such as CRISPR–Cas9 and RNAi-based screening, which have been performed in both mammalian[120–125] and non-mammalian[126,127] model systems. These functional genomics approaches can empirically link genes to phenotypes, providing valuable insight into their biological roles.

Protein structure prediction represents another powerful tool for functional annotation. Advances in deep learning, particularly the development of AlphaFold2[29] and RoseTTAFold[30], have enabled accurate prediction of 3D protein structures directly from amino acid sequences, even without clear sequence homology. Structural alignment tools[128–130] can facilitate functional annotation by searching predicted structures to databases of experimentally validated protein structures[131] or domains[132]. Structure prediction also provides an effective method for distinguishing between functional and non-functional isoforms of protein-coding genes, particularly in cases in which some isoforms have much more stable structures than others at the same locus[31].

Additional deep learning methods[30,133] have been developed to predict not only individual protein structures but also interactions between proteins, and between proteins and nucleic acids, potentially broadening the utility of structure-based functional prediction. However, caution is warranted, as recent studies have pointed out that some models make erroneous predictions[134] and generate physically implausible conformations[135].

Functional annotation of non-coding RNAs

The widespread availability of RNA-seq technology has catalysed the discovery and characterization of a broad range of non-coding RNA (ncRNA) gene types. Given that many ncRNAs are not polyadenylated, sequencing ncRNA requires methods other than the standard poly(A) selection protocols used for most RNA-seq experiments. For example, total RNA sequencing after rRNA depletion enables the detection of both polyadenylated and non-polyadenylated transcripts[32,33], including circular RNAs and a variety of small ncRNA types. Furthermore, small RNA populations can be enriched through specialized protocols, such as size selection or adapter ligation[34,35], which have contributed to the identification and categorization of miRNAs, siRNAs, piRNAs, snRNAs and snoRNAs. Additionally, CAGE and ChIP-seq experiments have been used to assign cell-type specificity[136] or gene regulatory functions[137] to ncRNAs.

Most ncRNA classes are defined based on their biological functions. For instance, rRNAs are essential components of the ribosome, snoRNAs direct chemical modifications (such as methylation or pseudouridylation of other RNAs[138,139]), and both siRNAs[140,141] and miRNAs[142,143] mediate post-transcriptional gene silencing through RNAi pathways. By contrast, lncRNAs are defined primarily by their transcript length, usually constrained to be at least 200 nucleotides, and by the absence of a protein-coding open reading frame[144]. Although tens of thousands of lncRNA transcripts have been annotated in humans and other organisms[145–149], only a small fraction have been experimentally

validated with a defined biological function. lncRNAs that have been characterized include *XIST*, which mediates X chromosome inactivation in female mammals[150]; *HOTAIR*, which mediates transcriptional silencing of the distant homeobox D cluster genes[151]; *MALAT1*, which regulates pre-mRNA splicing[152]; *NEAT1*, which is essential for paraspeckle formation[153]; *Evf2*, which enables differential transcriptional regulation of adjacent genes that share common DNA regulatory elements[154]; and *BC200* (*BCYRN1*), which has a role in cell migration and proliferation, and which has been linked to several types of cancer[155]. These well-studied lncRNA genes highlight their diverse functions in cellular processes.

Although ncRNAs are not routinely annotated as part of genome annotation pipelines, several specialized ncRNA databases have been developed, including lncRNAdb[145], LNCipedia[146], circBASE[156] and miRBase[157]. These resources serve as centralized platforms for consolidating experimental findings and guiding future research into specific RNA subclasses within the broader field of genome annotation.

Annotation of pseudogenes

Pseudogenes are usually defined as non-functional or otherwise defective copies of other genes[158]. These elements are quite numerous in many metazoan species (GENCODE lists 14,701 pseudogenes in its current human genome annotation)[159], with most being either processed pseudogenes, which are created by retrotransposition of spliced mRNAs[160], or unprocessed pseudogenes, which are created by segmental duplications of genomic DNA. The process of segmental duplication of genomic DNA typically creates intact copies of genes, including their upstream regulatory sites and introns, that are initially functional but over time tend to accumulate mutations and become pseudogenes (although in rare cases these extra copies can acquire new functions and become fixed in the population)[161]. Processed pseudogenes tend to be the larger class in mammals, for example comprising 72% and 74% of all pseudogenes currently annotated in human and mouse, respectively[37]. Notably, many regions annotated initially as pseudogenes were later shown to have functions and re-labelled as genes, as occurred with a truncated copy of *NOTCH2* that plays a direct role in human brain development[158,162,163].

The annotation of pseudogenes is complex and not a routine step in most annotation pipelines. When included, it is typically accomplished by a combination of automated and manual methods, the former of which detect processed pseudogenes by comparing putative protein-coding transcripts against others in the same genome[159,164]. If one locus contains introns and another locus is without introns, then some (but not all) pipelines will re-classify the latter as a processed pseudogene.

Annotation of newly assembled genomes

When a genome is assembled for the first time, the next step is to annotate the entire sequence as thoroughly and accurately as possible. Computational methods have been developed to automate this process, and the best strategy depends on the information available.

Automated annotation pipelines

Automated pipelines such as BRAKER3[165], MAKER2[166], Gnomon[167], the Ensembl annotation system[168], FINDER[169], GALBA[170], GeMoMa[171] and EviAnn[172] can be applied to annotate almost any eukaryotic genome, and they are particularly useful when no similar species has been annotated. These tools rely on various combinations of RNA-seq data, protein-to-DNA alignment and ab initio gene finders (Fig. 2). They continue to improve, and newer methods that use RNA-seq data have advantages over approaches that do not, particularly in their ability to identify non-coding genes and the untranslated regions (UTRs) of protein-coding genes.

Automated gene annotation tools that use RNA-seq data are well-suited to capture both protein-coding and non-coding genes (which ab initio gene finders and protein alignment methods cannot), as well as diverse alternative splicing events. Additionally, methods that use RNA-seq data are especially useful for organisms without well-annotated close relatives; for such organisms, homology-based annotation methods may be less effective. To annotate a genome using RNA-seq data, programs such as HISAT2[46] or STAR[47] are used to align the raw reads to the genome, which can then be assembled using programs such as StringTie[48–50], Scallop[51,52,60] and the earlier Cufflinks[173] system (Fig. 2, upper left). Additional steps (see below) are needed to find the open reading frames for protein-coding transcripts.

Another source of evidence for protein-coding gene structure comes from alignments of proteins to the DNA of a target genome. This process of aligning proteins to genomic DNA requires an aligner that can translate the DNA in all six reading frames and then perform spliced alignment (allowing for large insertions in the form of introns). Tools for spliced protein-to-DNA alignment include Exonerate[174] (used by MAKER2), miniprot[175] (used by EviAnn) and Spaln3[176]. When a protein-coding gene is identified with this approach, the alignment helps determine exon–intron boundaries as well as the positions of the start and stop codons, although it does not capture the UTRs for each transcript.

Another strategy for determining exon–intron structure utilizes alignments of annotated mRNA transcripts from another species to the target genome. If the source transcripts are from a closely related species (for example, human transcripts aligned to the chimpanzee genome), then a spliced aligner such as minimap2[59,176], specifically designed for mapping long query sequences, can align the transcripts and identify the boundaries of most exons. This strategy, which is included in the EviAnn pipeline, has the advantage of identifying UTRs as well as exon–intron boundaries, although its sensitivity declines as the evolutionary distance between source and target genomes increases.

The final category of automated annotation pipelines uses ab initio gene finders to annotate protein-coding genes. These gene finders include the tools AUGUSTUS[177] and GeneMark[178] (both used by BRAKER), which use machine learning methods such as HMMs and thus usually require training on a set of known genes, preferably of the same or similar species. After training, they attempt to identify a protein-coding region by finding a start codon, a series of alternating donor and acceptor splice sites that maintain the open read frame, and then a stop codon to mark the boundaries of the protein-coding region. More recently, deep learning algorithms have been applied to this task, as in Tiberius[179], which uses a convolutional neural network that includes an HMM to enforce reading frame constraints across the sequence. The deep learning approach might offer advantages in the future, although it is fundamentally constrained by limits on the number of high-quality genes available for training. Notably, these ab initio methods (by design) cannot find UTRs, which is a task that requires other evidence, such as RNA-seq data. Another limitation of the ab initio methods is that they generally find just one isoform per gene locus, whereas eukaryotic genomes usually contain many more (for example, the latest annotations of the human genome contain more than five transcript isoforms per protein-coding gene[2]).

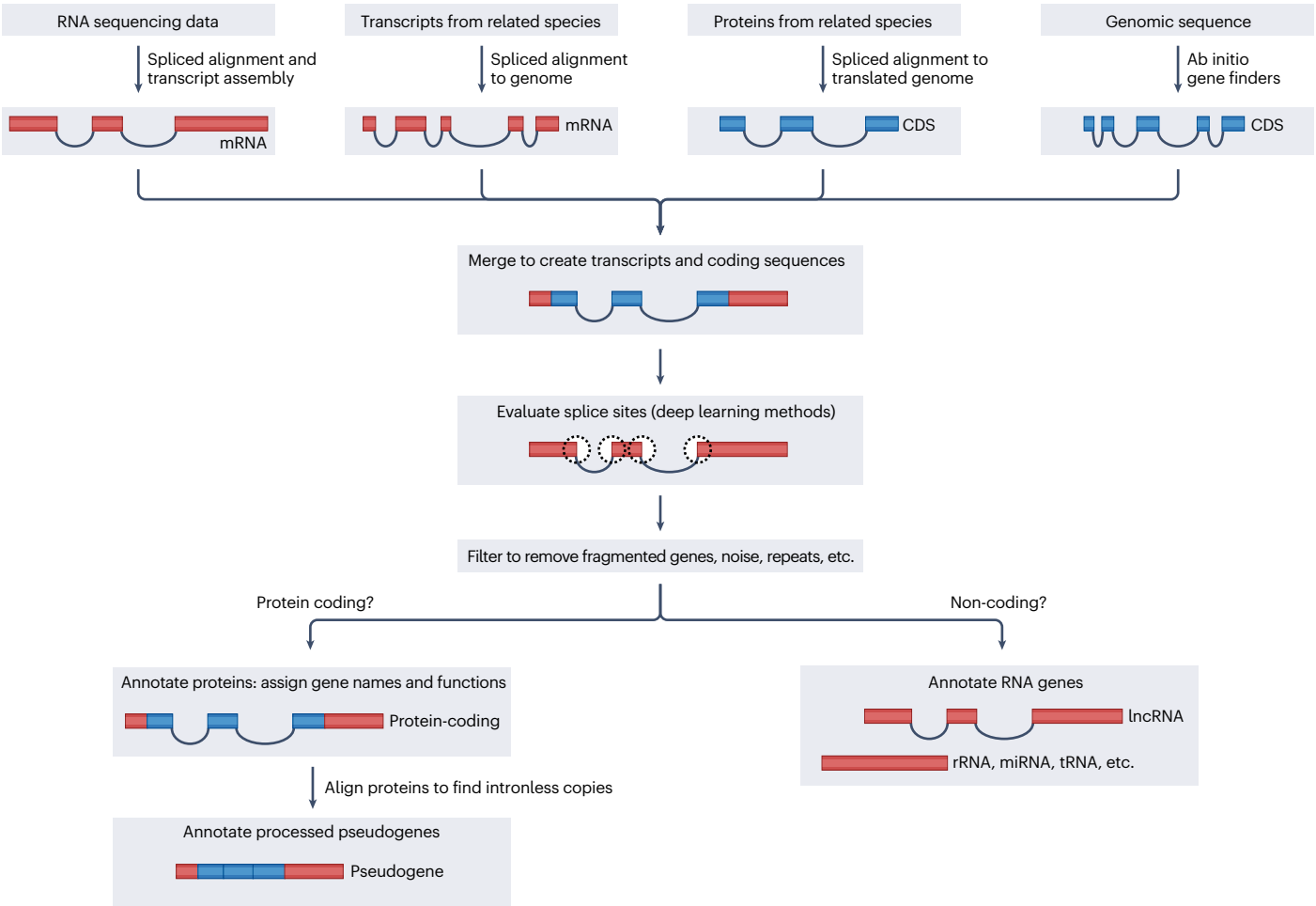

Fig. 2 | General workflow in automated genome-annotation pipelines. Types of evidence used for annotation include RNA-sequencing experiments, transcripts and protein sequences from related species, and the genome itself. These can be processed to predict the exon–intron structure of transcripts and the positions of protein-coding exons, which can then be merged to determine the protein-coding portions of transcripts. Note that some pipelines do not merge predictions in a separate step but may use alignments as an input to ab initio gene prediction. After filtering the predictions to remove various types of noise and computational errors, a pipeline can assign gene names and sometimes functions to each transcript. Some pipelines align all annotated protein-coding genes in a final step, and if any transcripts are found to be intronless versions of other transcripts, these are re-labelled as processed pseudogenes. Some of the evidence types shown across the top might not be available for some genomes, which may have an adverse impact on the quality or completeness of the resulting annotation. In particular, either RNA-sequencing data or transcripts from related species are needed to annotate non-coding genes and the untranslated portions of coding transcripts. CDS, coding DNA sequence; lncRNA, long non-coding RNA; miRNA, microRNA; rRNA, ribosomal RNA; tRNA, transfer RNA.

Lifting over annotation from a reference genome

Once a high-quality reference annotation exists for a species, annotating a new genome from a similar species does not require starting from scratch. Instead, existing annotations can be transferred based on sequence similarity, using tools such as Liftoff[180,181], LiftOn[182] and CAT[183]. This 'liftover' process is generally much faster than even the fastest automated annotation pipeline; for example, Liftoff can lift over annotations from one human genome to another in approximately 30 minutes on a single server, including the time required for alignment. When a well-curated annotation already exists, liftover methods can also be much more accurate than de novo approaches.

Perhaps the best case for using a liftover method arises when annotating another genome from a species (such as human) for which we already have a well-annotated genome. In this circumstance, the new genome is likely to share very high sequence identity (over 99%) with the reference genome. Liftover tools such as Liftoff and CAT align the DNA sequences of annotated genes from the reference to the new genome and transfer the associated annotation. In addition, whenever the new genome has copy number variation, the Liftoff system can find additional copies of annotated genes. Examples of this liftover strategy include the annotation of the Ashkenazi human genome, Ash1[184], the first telomere-to-telomere complete human genome, T2T-CHM13[185], and the bread wheat genome, *Triticum aestivum*[186], each of which relied on existing high-quality annotation. Notably, although the hexaploid bread wheat is extremely repetitive, with 85% of the genome occupied by retrotransposons, the LiftOff package was not only able to annotate it but also discovered >5,700 additional protein-coding gene copies[186]

that had been missed by the original annotators, who used the TriAnnot automated pipeline customized for wheat[187].

If the genome being annotated is not a member of the same species as the reference, DNA-based liftover methods can still be effective, but their accuracy declines as the sequences diverge. To map genes across larger evolutionary distances, the LiftOn[182] system augments DNA alignment with spliced protein alignment using miniprot[175]. This strategy enables LiftOn to transfer annotation between genomes that are more distantly related, such as mouse and rat.

Curated reference annotations

For the human genome, many years of hand-curated annotation efforts have produced several high-quality genome annotation databases, although the leading sources, namely GENCODE[37], RefSeq[39] and CHESS[38], do not yet agree with one another, particularly on the number and structure of lncRNA genes, but also on the isoforms of protein-coding genes. The Matched Annotation from NCBI and EMBL-EBI (EMBL-EBI, European Bioinformatics Institute) (MANE) project[188] is a carefully curated 'universal standard' for human protein-coding genes, with the stated goal of identifying just one isoform for each protein-coding gene, and for which the leading annotation databases all agree on the positions of splice sites, transcriptional start and termination sites, and start and stop positions of the protein-coding sequence. Release 1.4 (late 2024) of the MANE database included 19,338 proteins on the GRCh38 human reference genome, which comprises >97% of the human protein-coding genes in each of the leading databases. However, even with all this effort, it is widely agreed that the annotation of the human is far from complete[2].

Table 1 | Summary statistics from recent gene annotation catalogues

Organism	Number of genes	Number of transcripts*	Annotation source	Release date
Homo sapiens, GRCh38 (human)	41,731	141,100	CHESS[38]	2024
Mus musculus (house mouse)	78,258	278,299	GENCODE[37] release M37	2025
Caenorhabditis elegans (nematode)	49,173	67,055	WormBase[195] release WS296	2025
Danio rerio (zebrafish)	32,506	59,876	Ensembl[193]	2018
Macaca mulatta (macaque)	35,432	64,228	Ensembl	2019
Canis lupus familiaris (dog)	31,121	55,335	Ensembl	2020
Drosophila melanogaster (fruitfly)	17,872	35,109	FlyBase[41] Release dmel_r6.64	2025
Arabidopsis thaliana	32,833	54,013	Araport11[42]	2016
Oryza sativa Nipponbare (rice)	45,870	53,144	RAP-DB[192] Release IRGSP 1.0	2025

*Gene and transcript counts include all gene subtypes (for example, protein-coding, long non-coding RNA or pseudogene) as listed by each source; the Wormbase transcript count was computed from the annotation files by summing the counts across all RNA subtypes (for example, mRNA, microRNA or non-coding RNA); the RAP-DB gene count was computed by summing the number of lines in the predicted and experimentally validated gene locus annotation files (locus.gff, available from the RAP-DB website), and the RAP-DB transcript count was obtained by summing the counts of RNA entries in predicted and experimentally validated transcript annotation files (transcripts.gff).

GeneCards[189], a resource that provides uniquely rich annotation for the human genome, illustrates how genome annotation might expand in the future. The GeneCards website and database aggregates and summarizes information on every human gene from diverse sources, including the scientific literature[189]. Each gene has its own dedicated page that includes technical details about its exon–intron structure, genomic location, known mutations and their effects, drugs that target the gene, interactions with other genes, and paralogues and orthologues, as well as summaries about any diseases and other phenotypes linked to the gene. Although all these features could potentially be annotated for any species, the GeneCards database goes far beyond the basic annotation file that comprises the annotation for most genomes today.

Many other species have manually curated genome annotations that can be used as a source for annotation liftover, including mouse[40], rat[190], *D. melanogaster*[41], *Caenorhabditis elegans*[191], *A. thaliana*[43] and rice[192]. The annotation catalogues for these genomes continue to be updated, and gene numbers change as new genes are discovered and annotated (Table 1). In addition to these manual annotations, high-quality automated annotations have been produced by the NCBI RefSeq database for 1,365 genomes[39] and by the Ensembl project for 2,919 genomes[193].

Efforts are underway to generate high-quality genome sequences for even larger collections of organisms, such as the Vertebrate Genomes Project[3], the 5,000 arthropod genomes project[4], the 10,000 Plant Genomes Project[5] and the overarching Earth BioGenome Project[6,7]. All these genomes will need to be annotated, and once completed, they will be available as references for mapping annotations onto new genomes from additional species. Note that although this growth does create an ever-richer resource, it also comes with the risk of error propagation, in that incorrectly annotated genes may have their functions transferred to many other species, a problem that has been labelled as "transitive catastrophe[194]".

Conclusions and future challenges

Improvements in sequencing technology and new types of genome-based assays have greatly improved genome annotation over the past two decades (Fig. 3). However, except for a relatively small number of intensively studied organisms, the annotation of most genomes has not changed since the original publications. Updating those annotations with the benefits of improved computational methods and better data will substantially enhance the quality of downstream analyses that depend on them (Table 2).

Historically, annotation of gene function has focused on the molecular function of the gene in the cell, but a gene can play many roles in the biology of an organism. Some gene functions are annotated based on diseases they cause, including human genes linked to breast cancer and cystic fibrosis, although those same genes also have well-documented roles in the molecular biology of normal cells. The function of a gene cannot be captured simply by assigning it a name, as is often done using sequence homology. Future gene-level annotation might involve much richer descriptions, possibly along the lines of what is today provided by the GeneCards database for human genes[189], and further extended with additional layers of context such as developmental stage, tissue-specific expression and regulatory interactions. Machine learning and large-scale data integration are likely to accelerate these efforts, enabling annotations that are not only more comprehensive but also more predictive of biological outcomes.

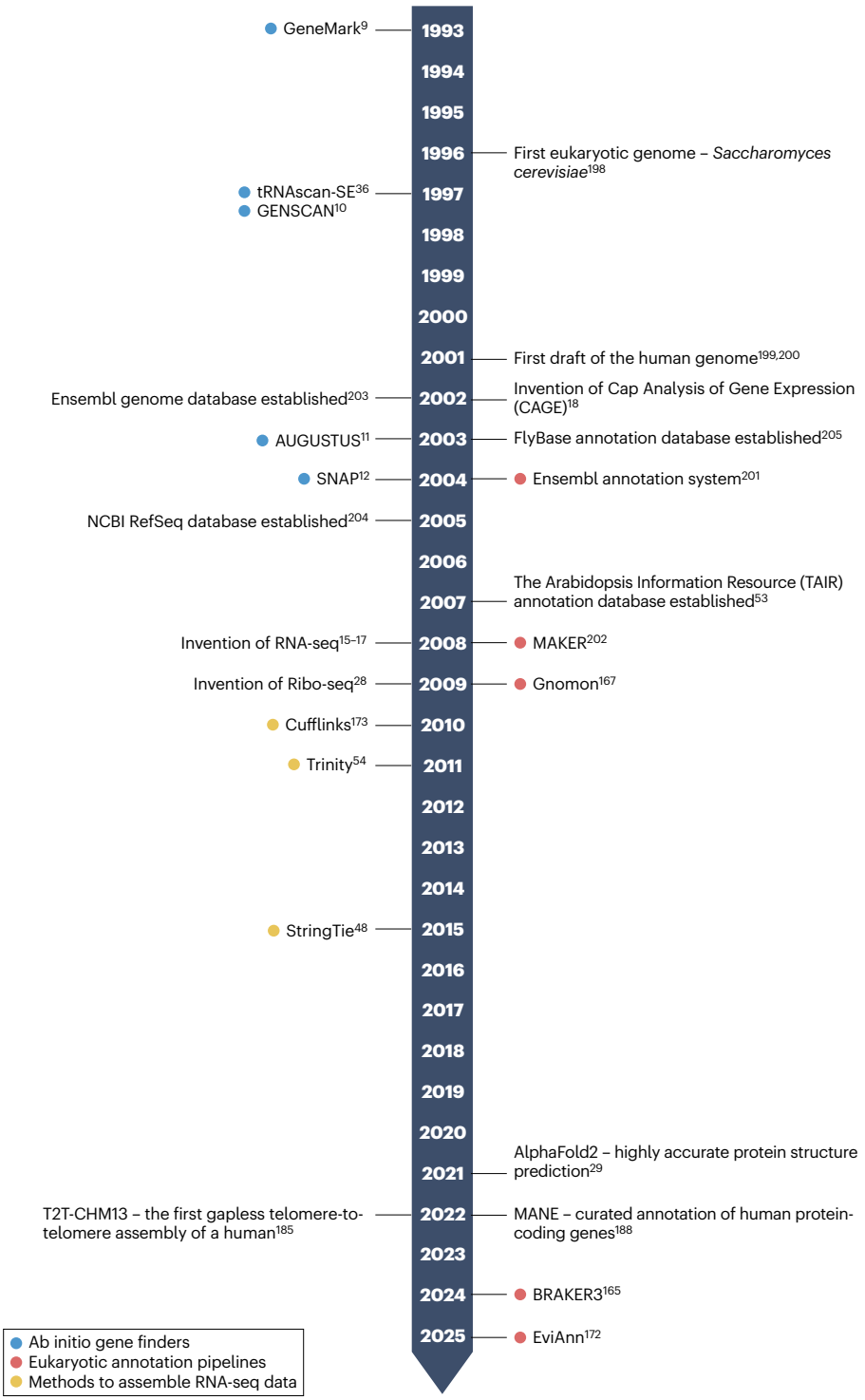

Fig. 3 | Timeline of advances in eukaryotic genome annotation. The earliest genome-scale efforts, which were largely manual, began with the publication of the first eukaryotic genome, *Saccharomyces cerevisiae*[198], and the first draft of the human genome a few years later, in 2001[199,200]. In the early 2000s, ab initio gene finders[9–12,36] (blue dots) were developed to predict protein-coding gene transcripts, and these were later augmented by eukaryotic genome annotation pipelines[165,167,172,201,202] (red dots), which sought to automate annotation while integrating multiple lines of experimental and computational evidence. In parallel with these developments, major genome databases emerged as repositories of annotation, including Ensembl[203] and National Center for Biotechnology Information's (NCBI's) RefSeq[204], which contained annotation of a wide range of genomes, together with model species-focused efforts such as FlyBase[205] (for *Drosophila melanogaster*) and The Arabidopsis Information Resource (TAIR)[53] (for *Arabidopsis thaliana*). The advent of high-throughput sequencing technologies spurred the development of new protocols, including RNA sequencing (RNA-seq)[15–17], cap analysis of gene expression (CAGE)[18] and 3′ end sequencing and ribosome profiling (Ribo-seq)[28], that greatly enhanced the annotation of transcripts and protein-coding genes. These protocols in turn drove the development of computational methods to assemble RNA-seq data[48,54,173] (yellow dots), which capture a more complete picture of the transcriptional landscape for many species. More recently, deep learning-based methods for protein structure prediction[29] have begun to contribute to annotation by helping to decide which splice isoforms are functional[31]. Although these advances have collectively led to substantial improvements in automated annotation (red dots in 2024–2025), the highest-quality annotation of human protein-coding genes available currently, the Matched Annotation from NCBI and EMBL-EBI (MANE) database, still involves extensive manual curation[188], and that database has not yet been fully mapped onto the recently completed human genome assembly[185]. EMBL-EBI, European Bioinformatics Institute.

Thousands of new genome sequencing projects are likely to be completed over the next few years, and all of them will need annotation. As the tree of life is annotated, the availability of similar, already-annotated species will likely make liftover methods the preferred choice, but these methods are far from perfect. Among other challenges, liftover methods struggle to distinguish closely related gene family members, and they are unable to identify the functional consequences of sequence changes among related proteins. More generally, annotation methods will continue to rely on highly curated reference genomes as their universal standard, and continuing efforts

Table 2 | A selection of software systems for genome annotation

Method	Software	URL
Automated genome annotation pipelines		
	BRAKER3[165]	https://github.com/Gaius-Augustus/BRAKER
	FINDER[169]	https://github.com/sagnikbanerjee15/Finder
	GALBA[170]	https://github.com/Gaius-Augustus/GALBA
	GeMoMa[196]	https://www.jstacs.de/index.php/GeMoMa
	EviAnn[172]	https://github.com/alekseyzimin/EviAnn_release
Ab initio gene finders		
	AUGUSTUS[177]	https://github.com/Gaius-Augustus/Augustus
	GeneMark[178]	https://genemark.bme.gatech.edu/
	SNAP[12]	https://github.com/KorfLab/SNAP
	GlimmerHMM[13]	https://ccb.jhu.edu/software/glimmerhmm/
	Tiberius[179]	https://github.com/Gaius-Augustus/Tiberius
Annotation liftover methods		
	Liftoff[180,181]	https://github.com/agshumate/Liftoff
	LiftOn[182]	https://ccb.jhu.edu/lifton
	CAT[183]	https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit
Spliced alignment of RNA sequences to genomic DNA		
	STAR[47]	https://github.com/alexdobin/STAR
	HISAT2[46]	https://daehwankimlab.github.io/hisat2/
	minimap2[59]	https://github.com/lh3/minimap2
Spliced alignment of protein sequences to genomic DNA		
	miniprot[175]	https://github.com/lh3/miniprot
	Spaln3[176]	https://github.com/ogotoh/spaln
Transcriptome assemblers		
	StringTie[48–50]	https://ccb.jhu.edu/software/stringtie/
	Scallop[51,52,60]	https://github.com/Shao-Group/scallop2
	Trinity[54]	https://github.com/trinityrnaseq/trinityrnaseq/wiki

to improve the annotation of reference genomes will propagate across and benefit a wide range of animal and plant genomes. Ultimately, the future of annotation will depend not only on technological progress but also on sustained collaboration within the community to ensure that the rapidly expanding catalogue of sequenced species is matched by equally rich and accurate descriptions of their biology.

Published online: 17 February 2026

References

1. Clinton, W. J. Remarks on the completion of the first survey of the human genome. *The American Presidency Project* https://www.presidency.ucsb.edu/node/227458 (2000).
2. Amaral, P. et al. The status of the human gene catalogue. *Nature* **622**, 41–47 (2023).
This paper describes the efforts to characterize every transcript isoform for every human gene and the challenges involved in determining which isoforms are functional.
3. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
4. Robinson, G. E. et al. Creating a buzz about insect genomes. *Science* **331**, 1386 (2011).
5. Cheng, S. et al. 10KP: a phylodiverse genome sequencing plan. *GigaScience* https://doi.org/10.1093/gigascience/giy013 (2018).
6. Lewin, H. A. et al. Earth biogenome project: sequencing life for the future of life. *Proc. Natl Acad. Sci. USA* **115**, 4325–4333 (2018).
This paper introduces a massive effort to sequence every eukaryotic species on the planet.
7. Blaxter, M. et al. The Earth Biogenome Project Phase II: illuminating the eukaryotic tree of life. *Front. Sci.* https://doi.org/10.3389/fsci.2025.1514835 (2025).
8. Koepfli, K.-P. & Paten, B. the Genome 10K Community of Scientists & O'Brien, S. J. The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* **3**, 57–111 (2015).
9. Borodovsky, M. & McIninch, J. GENMARK: parallel gene recognition for both DNA strands. *Computers Chem.* **17**, 123–133 (1993).
10. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
11. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
12. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
13. Majoros, W., Pertea, M. & Salzberg, S. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
14. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
This publication describes the second major upgrade to BLAST, which remains one of the most highly cited computational methods in the field.
15. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
16. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
This landmark paper describes the invention of RNA-seq technology and applies it to mice.
17. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
18. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
19. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e524 (2019).
20. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).
21. Chao, K.-H., Mao, A., Salzberg, S. L. & Pertea, M. Splam: a deep-learning-based splice site predictor that improves spliced alignments. *Genome Biol.* **25**, 243 (2024).
22. Scalzitti, N. et al. Spliceator: multi-species splice site prediction using convolutional neural networks. *BMC Bioinform.* **22**, 561 (2021).
23. Wang, R., Wang, Z., Wang, J. & Li, S. SpliceFinder: ab initio prediction of splice sites using convolutional neural network. *BMC Bioinform.* **20**, 652 (2019).
24. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
25. Dehghannasiri, R., Olivieri, J. E., Damljanovic, A. & Salzman, J. Specific splice junction detection in single cells with SICILIAN. *Genome Biol.* **22**, 219 (2021).
26. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice — improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
27. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
28. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
This paper introduces ribosome profiling and its use to study translated regions.
29. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
This paper describes AlphaFold2, the first program ever to predict the 3D structure of proteins with accuracy comparable with that achieved by crystallography experiments.
30. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
31. Sommer, M. J. et al. Structure-guided isoform identification for the human transcriptome. *eLife* **11**, e82556 (2022).
This paper describes how to use AlphaFold2 or ColabFold to determine the functional and non-functional transcript isoforms of a gene.
32. Adiconis, X. et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
33. Singh, A. et al. A scalable and cost-efficient rRNA depletion approach to enrich RNAs for molecular biology investigations. *RNA* **30**, 728–738 (2024).
34. Hafner, M. et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**, 1697–1712 (2011).
35. Faridani, O. R. et al. Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
36. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).

37. Mudge, J. M. et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
38. Varabyou, A. et al. CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *Genome Biol.* **24**, 249 (2023).
39. Goldfarb, T. et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res.* **53**, D243–D257 (2025).
40. Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A. & Richardson, J. E. Mouse genome database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–d806 (2019).
41. Öztürk-Çolak, A. et al. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics* https://doi.org/10.1093/genetics/iyad211 (2024).
42. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant. J.* **89**, 789–804 (2017).
43. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
44. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **49**, D212–D220 (2021).
45. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
46. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
The paper describes StringTie, a maximum-flow-based algorithm for assembling transcripts into gene structures and for quantifying expression.
49. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
50. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput. Biol.* **18**, e1009730 (2022).
51. Shao, M. & Kingsford, C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* **35**, 1167–1169 (2017).
52. Zhang, Q., Shi, Q. & Shao, M. Accurate assembly of multi-end RNA-seq data with Scallop2. *Nat. Comput. Sci.* **2**, 148–152 (2022).
53. Swarbreck, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2008).
54. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
55. Burdick, J. T. et al. Nanopore-based direct sequencing of RNA transcripts with 10 different modified nucleotides reveals gaps in existing technology. *G3 Genes Genomes Genet.* **13**, jkad200 (2023).
56. Chen, Y. et al. A systematic benchmark of nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. *Nat. Methods* https://doi.org/10.1038/s41592-025-02623-4 (2025).
57. Pardo-Palacios, F. J. et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat. Methods* **21**, 1349–1363 (2024).
58. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
59. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
This publication describes minimap2, one of the fastest and most accurate tools for aligning both short and long DNA sequencing reads to a genome.
60. Tung, L. H., Shao, M. & Kingsford, C. Quantifying the benefit offered by transcript assembly with Scallop-LR on single-molecule long reads. *Genome Biol.* **20**, 287 (2019).
61. Prjibelski, A. D. et al. Accurate isoform discovery with IsoQuant using long reads. *Nat. Biotechnol.* **41**, 915–918 (2023).
62. Gao, Y. et al. ESPRESSO: robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci. Adv.* **9**, eabq5072 (2023).
63. Chen, Y. et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat. Methods* **20**, 1187–1195 (2023).
64. Han, S. W., Jewell, S., Thomas-Tikhonenko, A. & Barash, Y. Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads. *Genome Res.* **34**, 1624–1635 (2024).
65. Nip, K. M. et al. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nat. Commun.* **14**, 2940 (2023).
66. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
67. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
68. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
69. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
70. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
71. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
72. Regev, A. et al. The human cell atlas. *eLife* https://doi.org/10.7554/eLife.27041 (2017).
73. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
74. Nip, K. M. et al. RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Res.* **30**, 1191–1200 (2020).
75. Liu, J., Liu, X., Ren, X. & Li, G. scRNAss: a single-cell RNA-seq assembler via imputing dropouts and combing junctions. *Bioinformatics* **35**, 4264–4271 (2019).
76. Shi, Q., Zhang, Q. & Shao, M. Transcriptome assembly at single-cell resolution with Beaver. *Bioinformatics* **41**, i323–i331 (2025).
77. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
78. Hagemann-Jensen, M. et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
79. Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
80. Healey, H. M., Bassham, S. & Cresko, W. A. Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics* https://doi.org/10.1093/genetics/iyac017 (2022).
81. Tian, L. et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol.* **22**, 310 (2021).
82. Dondi, A. et al. Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer. *Nat. Commun.* **14**, 7780 (2023).
83. Li, Z. et al. An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing. *Cell Genom.* https://doi.org/10.1016/j.xgen.2024.100641 (2024).
84. Westoby, J., Artemov, P., Hemberg, M. & Ferguson-Smith, A. Obstacles to detecting isoforms using full-length scRNA-seq data. *Genome Biol.* **21**, 74 (2020).
85. Deng, E. et al. Systematic evaluation of single-cell RNA-seq analyses performance based on long-read sequencing platforms. *J. Adv. Res.* **71**, 141–153 (2025).
86. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
87. Valen, E. et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255–265 (2009).
88. Noguchi, S. et al. FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
89. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
90. Ozsolak, F. et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018–1029 (2010).
91. Beck, A. H. et al. 3′-End sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* **5**, e8768 (2010).
92. Shepard, P. J. et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).
93. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep.* **1**, 753–763 (2012).
94. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
95. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).
96. Gruber, A. J. et al. A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**, 1145–1159 (2016).
97. You, L. et al. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* **43**, D59–D67 (2015).
98. Herrmann, C. J. et al. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.* **48**, D174–D179 (2020).
99. Moon, Y., Herrmann, C. J., Mironov, A. & Zavolan, M. PolyASite v3. 0: a multi-species atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res.* **53**, D197–D204 (2025).
100. Crappé, J. et al. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* **43**, e29–e29 (2015).
101. Chun, S. Y., Rodriguez, C. M., Todd, P. K. & Mills, R. E. SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinform.* **17**, 482 (2016).
102. Fields, A. P. et al. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell* **60**, 816–827 (2015).
103. Erhard, F. et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* **15**, 363–366 (2018).
104. Swirski, M. I. et al. Translon: a single term for translated regions. *Nat. Methods* **22**, 2002–2006 (2025).
105. 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).
This paper describes PhyloCSF, a powerful method for determining patterns of sequence conservation that indicate whether a genomic region encodes a protein.

106. Menschaert, G. et al. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteom.* **12**, 1780–1790 (2013).
107. Chong, C. et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
108. Ouspenskaia, T. et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* **40**, 209–217 (2022).
109. Mudge, J. M. et al. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).
110. Ji, H. & Salzberg, S. L. Upstream open reading frames may contain hundreds of novel human exons. *PLoS Comput. Biol.* **20**, e1012543 (2024).
111. Calviello, L., Hirsekorn, A. & Ohler, U. Quantification of translation uncovers the functions of the alternative transcriptome. *Nat. Struct. Mol. Biol.* **27**, 717–725 (2020).
112. Pertea, M., Lin, X. Y. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29**, 1185–1190 (2001).
113. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
114. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
115. Omer, S., Harlow, T. J. & Gogarten, J. P. Does sequence conservation provide evidence for biological function? *Trends Microbiol.* **25**, 11–18 (2017).
116. Platt, A., Ross, H. C., Hankin, S. & Reece, R. J. The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase. *Proc. Natl Acad. Sci. USA* **97**, 3154–3159 (2000).
117. Balmer, P. et al. A curated catalog of canine and equine keratin genes. *PLoS One* **12**, e0180359 (2017).
118. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* **19**, 575–592 (2018).
119. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
120. Zhang, R. et al. A CRISPR screen defines a signal peptide processing pathway required by flaviviruses. *Nature* **535**, 164–168 (2016).
121. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
122. Berns, K. et al. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437 (2004).
123. Moffat, J. et al. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
124. Paddison, P. J. et al. A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428**, 427–431 (2004).
125. Westbrook, T. F. et al. A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121**, 837–848 (2005).
126. Boutros, M. et al. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832–835 (2004).
127. Xu, Y. et al. CRISPR screens in *Drosophila* cells identify Vsg as a Tc toxin receptor. *Nature* **610**, 349–355 (2022).
128. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01773-0 (2024).
129. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
130. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
131. Berman, H. M. et al. The Protein Data Dank. *Nucleic Acids Res.* **28**, 235–242 (2000).
132. Orengo, C. A. et al. CATH — a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109 (1997).
133. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
134. Pratt, O. S. et al. AlphaFold 2, but not AlphaFold 3, predicts confident but unrealistic β-solenoid structures for repeat proteins. *Comput. Struct. Biotechnol. J.* **27**, 467–477 (2025).
135. Aubel, M., Eicholt, L. & Bornberg-Bauer, E. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research* **12**, 347 (2023).
136. Hashimoto, K. et al. CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. *Genome Res.* **25**, 1812–1824 (2015).
137. Zhou, K.-R. et al. ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.* **45**, D43–D50 (2017).
138. Ganot, P., Bortolin, M.-L. & Kiss, T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell* **89**, 799–809 (1997).
139. Kiss-László, Z., Henry, Y., Bachellerie, J.-P., Caizergues-Ferrer, M. & Kiss, T. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell* **85**, 1077–1088 (1996).
140. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
141. Elbashir, S. M. et al. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
142. Pillai, R. S. et al. Inhibition of translational initiation by Let-7 microRNA in human cells. *Science* **309**, 1573–1576 (2005).
143. Chendrimada, T. P. et al. MicroRNA silencing through RISC recruitment of eIF6. *Nature* **447**, 823–828 (2007).
144. 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).
145. Quek, X. C. et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **43**, D168–D173 (2015).
146. Volders, P.-J. et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* **41**, D246–D251 (2013).
147. Ma, L. et al. LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.* **47**, D128–D134 (2019).
148. Hon, C.-C. et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).
149. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
150. Li, J. et al. Long noncoding RNA *XIST*: Mechanisms for X chromosome inactivation, roles in sex-biased diseases, and therapeutic opportunities. *Genes Dis.* **9**, 1478–1492 (2022).
151. 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).
152. Tripathi, V. et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39**, 925–938 (2010).
153. Clemson, C. M. et al. An architectural role for a nuclear noncoding RNA: *NEAT1* RNA is essential for the structure of paraspeckles. *Mol. Cell* **33**, 717–726 (2009).
154. Berghoff, E. G. et al. *Evf2* (*Dlx6as*) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. *Development* **140**, 4407–4416 (2013).
155. Samson, J., Cronin, S. & Dean, K. BC200 (BCYRN1) — the shortest, long, non-coding RNA associated with cancer. *Non-coding RNA Res.* **3**, 131–143 (2018).
156. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
157. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
158. Cheetham, S. W., Faulkner, G. J. & Dinger, M. E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* **21**, 191–201 (2020).
159. Sisu, C. et al. Comparative analysis of pseudogenes across three phyla. *Proc. Natl Acad. Sci. USA* **111**, 13361–13366 (2014).
This paper summarizes the processes that create and degrade pseudogenes across the human, nematode and fruit fly genomes.
160. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000).
161. Wagner, A. The fate of duplicated genes: loss or new function? *Bioessays* **20**, 785–788 (1998).
162. Suzuki, I. K. et al. Human-specific *NOTCH2NL* genes expand cortical neurogenesis through delta/notch regulation. *Cell* **173**, 1370–1384 e1316 (2018).
163. Fiddes, I. T. et al. Human-specific *NOTCH2NL* genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369 e1322 (2018).
164. Pei, B. et al. The GENCODE pseudogene resource. *Genome Biol.* **13**, R51 (2012).
165. Gabriel, L. et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
This paper describes BRAKER3, an automated annotation pipeline that includes both ab initio prediction and RNA-seq evidence.
166. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 491 (2011).
167. Souvorov, A. et al. Gnomon–NCBI eukaryotic gene prediction tool. *National Center for Biotechnology Information* https://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf (2010).
168. Aken, B. L. et al. The Ensembl gene annotation system. *Database* https://doi.org/10.1093/database/baw093 (2016).
169. Banerjee, S. et al. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinform.* **22**, 205 (2021).
170. Brůna, T. et al. Galba: genome annotation with miniprot and AUGUSTUS. *BMC Bioinform.* **24**, 327 (2023).
171. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform.* **19**, 189 (2018).
172. Zimin, A. V., Puiu, D., Pertea, M., Yorke, J. A. & Salzberg, S. L. Efficient evidence-based genome annotation with EviAnn. *Nat. Methods* (in the press).
This paper describes EviAnn, an automated annotation pipeline that combines RNA-seq evidence, protein-to-DNA alignment and transcripts from related species.
173. 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).
174. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **6**, 31 (2005).

175. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btad014 (2023).
176. Gotoh, O. Spaln3: improvement in speed and accuracy of genome mapping and spliced alignment of protein query sequences. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btae517 (2024).
177. 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).
178. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
179. Gabriel, L., Becker, F., Hoff, K. J. & Stanke, M. Tiberius: end-to-end deep learning with an HMM for gene prediction. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btae685 (2024).
180. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
This publication describes Liftoff, the first standalone tool that could map annotation from one genome directly onto another.
181. Shumate, A. & Salzberg, S. LiftoffTools: a toolkit for comparing gene annotations mapped between genome assemblies. *F1000Research* **11**, 1230 (2022).
182. Chao, K.-H. et al. Combining DNA and protein alignments to improve genome annotation with LiftOn. *Genome Res.* **35**, 311–325 (2025).
183. Fiddes, I. T. et al. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018).
184. Shumate, A. et al. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**, 129 (2020).
185. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
This landmark paper describes the first ever complete human genome sequence, whose annotation includes 140 new protein-coding genes.
186. Alonge, M., Shumate, A., Puiu, D., Zimin, A. V. & Salzberg, S. L. Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. *Genetics* **216**, 599–608 (2020).
187. The International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
188. Morales, J. et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
This paper describes the MANE database, a very high-quality annotation containing a single splice isoform for nearly every human protein-coding gene.
189. Stelzer, G. et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54**, 1.30.31–31.30.33 (2016).
190. Lauledekind, S. J. F. et al. The rat genome database: genetic, genomic, and phenotypic data across multiple species. *Curr. Protoc.* **3**, e804 (2023).
191. Yoshimura, J. et al. Recompleting the *Caenorhabditis elegans* genome. *Genome Res.* **29**, 1009–1022 (2019).
192. Sakai, H. et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* **54**, e6 (2013).
193. Dyer, S. C. et al. Ensembl 2025. *Nucleic Acids Res.* **53**, D948–D957 (2025).
194. Salzberg, S. L. Genome re-annotation: a wiki solution? *Genome Biol.* **8**, 102 (2007).
195. Sternberg, P. W. et al. WormBase 2024: status and transitioning to alliance infrastructure. *Genetics* https://doi.org/10.1093/genetics/iyae050 (2024).
196. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89–e89 (2016).
197. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
198. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546–567 (1996).
199. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
200. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
201. Curwen, V. et al. The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004).
202. Cantarel, B. L. et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
203. Hubbard, T. et al. The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
204. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
205. The FlyBase Consortium. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31**, 172–175 (2003).
206. Adams, M. D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
207. Allen, J. E. & Salzberg, S. L. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21**, 3596–3603 (2005).

Acknowledgements
This work was supported in part by NIH grants R01-HG006677 (S.L.S.) and R35-GM156470 (M.P.).

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Additional information
Peer review information *Nature Reviews Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Related links
10,000 Plant Genomes Project: https://db.cngb.org/10kp/
5,000 arthropod genomes project: https://i5k.github.io/
AUGUSTUS: https://github.com/Gaius-Augustus/Augustus
BRAKER3: https://github.com/Gaius-Augustus/BRAKER
CAT: https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit
circBase: https://www.circbase.org/
Earth Biogenome Project: https://www.earthbiogenome.org/
ENCODE: https://www.encodeproject.org/
EviAnn: https://github.com/alekseyzimin/EviAnn_release
FANTOM: https://fantom.gsc.riken.jp/
FINDER: https://github.com/sagnikbanerjee15/Finder
GALBA: https://github.com/Gaius-Augustus/GALBA
GeMoMa: https://www.jstacs.de/index.php/GeMoMa
GeneCards: https://www.genecards.org/
GeneMark: https://genemark.bme.gatech.edu/
GlimmerHMM: https://ccb.jhu.edu/software/glimmerhmm/
HISAT2: https://daehwankimlab.github.io/hisat2/
Liftoff: https://github.com/agshumate/Liftoff
LiftOn: https://ccb.jhu.edu/lifton
LNCipedia: https://lncipedia.org/
minimap2: https://github.com/lh3/minimap2
miniprot: https://github.com/lh3/miniprot
miRBase: https://www.mirbase.org/
Scallop: https://github.com/Shao-Group/scallop2
SNAP: https://github.com/KorfLab/SNAP
Spaln3: https://github.com/ogotoh/spaln
STAR: https://github.com/alexdobin/STAR
StringTie: https://ccb.jhu.edu/software/stringtie/
Tiberius: https://github.com/Gaius-Augustus/Tiberius
Trinity: https://github.com/trinityrnaseq/trinityrnaseq/wiki
Vertebrate Genomes Project: https://vertebrategenomesproject.org/

© Springer Nature Limited 2026