nature methods

Article

https://doi.org/10.1038/s41592-026-03036-7

Highly accurate ab initio gene annotation with ANNEVO

Received: 18 April 2025

Accepted: 18 February 2026

Published online: 12 March 2026

Check for updates

Pengyu Zhang[1,2], Tun Xu[1,2], Songbo Wang[1,2], Xiaofei Yang[2,3,4], Peisen Sun[1,2], Peng Jia[1,2,5], Jiadong Lin[1,2], Bo Wang[1,2], Yizhe Zhang[6], Deyu Meng[2,7,8,9], Stephen J. Bush[1,2], Zemin Ning[10] & Kai Ye[1,2,5,11]

Accurate gene annotation is essential for deciphering the mapping from genomic sequences to their functional roles. However, current methods struggle to model complex gene transmission patterns, such as vertical inheritance and horizontal gene transfer. Here we introduce ANNEVO, a mixture of experts-based genomic language model that directly models distal sequence dependencies and joint evolutionary relationships from diverse genomes, enabling precise ab initio gene annotation. Through extensive benchmarking on 566 phylogenetically diverse species, we demonstrate that ANNEVO substantially outperforms existing ab initio methods and achieves performance comparable to state-of-the-art annotation pipelines. Furthermore, ANNEVO's independence from external evidence allows it to deliver more complete annotations than reference annotations for a broad range of species while correcting errors within them. These advancements will improve genome sequence interpretation and provide a framework capable of integrating evolutionary insights.

Accurate gene annotation—the identification of genes and their internal structures within genomes[1]—is fundamental for unlocking the biological insights encoded within genome sequences, yet it remains a major bottleneck in the postgenomic era. Recent advancements in sequencing technologies have fueled large-scale genomic projects such as the Earth BioGenome Project[2]. However, the mapping of genome sequences to their biological functions through gene annotation, analogous to AlphaFold's[3] deciphering of protein sequence mapping, has not kept pace. This has stalled the annotation of numerous eukaryotic species, impeding further research and application. The most direct approach to gene annotation is ab initio gene annotation, which infers gene structures solely from genomic sequence composition. This approach is the foundation of many early classical methods[4–8] such as Augustus and GeneMark, which employ hidden Markov models (HMMs) with static parameters to model gene structures. However, despite its conceptual elegance, ab initio annotation suffers from limited accuracy, prompting widespread reliance on automated pipelines[9–14] such as BRAKER3, which are often supplemented with repeat masking[15,16] and extensive extrinsic evidence such as known proteins and RNA sequencing (RNA-seq) data. However, this multi-faceted approach presents several challenges: species with limited wet-laboratory data suffer from poor annotations; over-reliance on wet-laboratory data can bias annotation by omitting genes with rare spatial-temporal expression patterns; and the intensive evidence data processing and comparisons raise computational demands. These limitations have directly resulted in nearly half of the species in the National Center for Biotechnology

[1]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [2]MOE Key Laboratory for Intelligent Networks and Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [3]Department of Dermatology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. [4]School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [5]Department of Gynecology and Obstetrics, Center for Mathematical Medicine, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. [6]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China. [7]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. [8]Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau. [9]Pazhou Laboratory (Huangpu), Guangzhou, China. [10]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [11]Genome Institute, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. e-mail: kaiye@xjtu.edu.cn

Information (NCBI) GenBank database lacking annotations[11], a challenge that will be further exacerbated by the Earth BioGenome Project's aim to sequence more than 1 million genomes within the next decade. This underscores an urgent need for rapid, highly accurate and evidence-free gene annotation methods.

These pressing needs have prompted a renewed focus on ab initio gene annotation. Recent deep learning-based methods such as Helixer[17,18] and Tiberius[19] have revitalized this field by framing gene annotation as a nucleotide-level sequence labeling task. They employ deep learning architectures such as long short-term memory[20] to predict the class probability of each nucleotide, which is then post-processed to define gene structures. By more effectively modeling the statistical patterns of genomic sequences compared to HMMs, these methods represent important progress in ab initio annotation. While these methods excel at general syntactic pattern recognition, they do not yet sufficiently account for the heterogeneity in genomic semantics across taxonomically divergent clades, which is fundamentally a product of complex evolutionary processes. Eukaryotic gene evolution is shaped by diverse forces, including vertical inheritance, horizontal transfer, gene fusion and gene loss. Genes inherited vertically from distant common ancestors undergo dynamic transformations via variation, introgression, gene fusion and clade-specific gains or losses[21–24]. In parallel, hybridization or horizontal gene transfer—the movement of genetic materials between species—require considering gene variations across diverse clades. As such, gene evolution through both vertical and horizontal inheritance necessitates tracking gene transmission beyond the confines of a single phylogenetic tree, highlighting the interconnectedness of different clades. This complexity underscores the need for a more comprehensive understanding of gene evolution, one achieved by simultaneously analyzing the relationships and heterogeneity among multiple evolutionary branches. Such an approach would model different clades and their relationships at a higher taxonomic level, leveraging insights from each to improve overall predictions.

Current gene annotation methods, however, fall short of this ideal. For instance, Augustus primarily relies on training gene models from closely related species, limiting their ability to trace gene evolution across broader taxonomic groups. Moreover, these methods often employ inflexible models, such as HMMs, that utilize constant parameters and fail to account for varying characteristics of genes across different clades. Even the advanced deep learning methods of Helixer and Tiberius do not consider the vertical and horizontal evolutionary trajectories of eukaryotic organisms in their modeling. Furthermore, a distinct but equally critical challenge lies in modeling long-range dependencies, particularly in long genes that are the products of gene fusion or the expansion of exon and/or intron sequences[25,26]. These long genes often exhibit complex features, including distal interactions within genomes and intricate patterns of paired splicing sequences. However, existing methods struggle to effectively use crucial information contained within these distal sequences.

To overcome these challenges, we introduce ANNEVO, a deep learning framework for accurate ab initio gene annotation by explicitly modeling the complex evolutionary relationships between diverse taxonomic groups and simultaneously capturing long-range sequence dependencies within individual genomes. We formulated the annotation task as sequence-based prediction of the longest coding transcript across phylogenetically diverse species. Through comprehensive benchmarking against baseline methods, we demonstrate that ANNEVO exhibits broad applicability across species in different clades, and even outperforms state-of-the-art evidence-guided pipelines. Furthermore, ANNEVO's predictive capability also extends to the refinement of existing reference annotations. For example, ANNEVO successfully corrected over 3% of the total BUSCO (benchmarking universal single-copy orthologs) set within the *Brassica oleracea* genome assembly (Ensembl database), each correction independently validated by RNA-seq evidence, showcasing its capability to refine and enhance the accuracy of existing annotations.

Results

Overview of the ANNEVO framework

ANNEVO comprises three main components: context extension, neural network and gene structure decoding. The context extension component strategically extends the flanking regions to mitigate model bias caused by insufficient contextual information at the edges of genome segments (Fig. 1a and Methods). This extension ensures that the flanking regions provide valid nucleotide information without negatively affecting model training by employing a targeted loss masking strategy (Methods). By leveraging preidentified erroneous regions (Supplementary Method 1), ANNEVO uses their nucleotide information to enhance predictions at other sites but prevent the propagation of incorrect annotations.

The neural network component performs end-to-end nucleotide-resolution prediction (Fig. 1b). This component follows a two-level hierarchical strategy to model gene evolution at two scales. To address the deep divergence between major evolutionary groups, we developed five separate neural network models. Each model is dedicated to one of the following major clades: Fungi, Embryophyta, Invertebrates, Vertebrate_other and Mammalia (following RefSeq's taxonomic groupings). When annotating a genome, the appropriate clade-specific model is selected, preventing negative interference from evolutionarily distant groups. To account for the diversity within each major clade, every clade-specific neural network is built on a mixture of experts (MoE) architecture, where internal experts learn to specialize on the distinct genomic features of various subclades. Each of these five neural networks consists of three core modules: (1) the distal information modeling module integrates information within subregions using a convolutional tower to form bins, similar to how the Hi-C technique represents distal sequence interactions. We used the Transformer technique[27] to compute relationships between these bins, effectively modeling distal interactions (Fig. 1d, Extended Data Fig. 1a and Methods). This approach enables ANNEVO to model sequences up to ~40 kilobases (kb) in length. (2) The joint evolutionary modeling module utilizes a broad array of species from higher taxonomic levels to model multiple subclades. To train specific networks for different subclades, it incorporates a controller that dynamically weighs the contributions of different subnetworks based on the sequence's characteristics. This controller analyzes the sequence's composition and decides how much each specialized subnetwork should contribute (Fig. 1d, Extended Data Fig. 1b and Methods). Consequently, during training, each specialized subnetwork becomes an expert in a specific subclade, while during prediction, the model automatically adjusts its feature representation based on the sequence composition. (3) The resolution restorer module employs multiple layers of transposed convolutions, interleaved with normalization and convolution operations, to progressively upsample the local representations back to the original input resolution. This enables nucleotide-level prediction and ensures precise localization of fine-grained sequence features in the output (Fig. 1d, Extended Data Fig. 1c and Methods).

The gene structure decoding component combines the predictions of each segment to identify potential gene regions. It then applies the Viterbi algorithm[28] to generate gene structures that conform to biological rules (Fig. 1c, Extended Data Fig. 2 and Methods). Because the model has learned evolutionary and contextual learning, the decoding step requires minimal parameter adjustment, reducing the risk of introducing biases.

ANNEVO enables gene annotation across divergent clades

To demonstrate ANNEVO's broad applicability across diverse species, we evaluated ANNEVO using 566 RefSeq species spanning phylogenetically diverse clades: Fungi, Embryophyta, Invertebrates,

https://doi.org/10.1038/s41592-026-03036-7

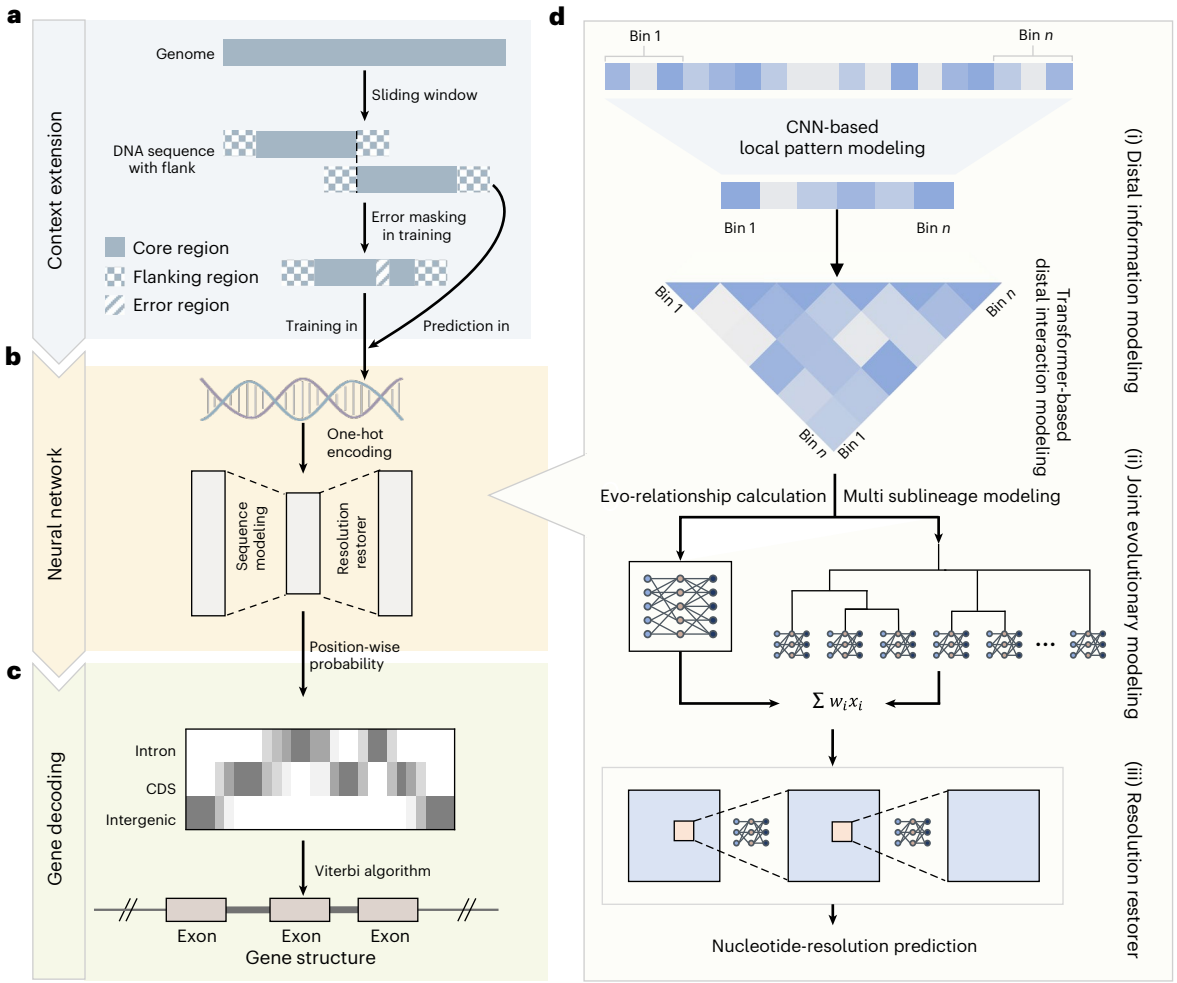


Fig. 1 | A visual overview of ANNEVO's architecture. a, Context extension component: this panel illustrates how ANNEVO tackles the challenge of insufficient context at the edges of sequence segments. The genome is divided into consecutive core regions using a sliding window. Each core region is extended by flanking sequences on both sides, providing additional context for the model. To ensure robustness during training, a soft masking strategy is applied to both the flanking regions and preidentified erroneous regions, preventing these regions from disproportionately influencing the training process. **b**, Neural network component: this module enables end-to-end position-wise predictions by modeling both long-range interactions within sequences and multiple sublineages across a diverse set of species. **c**, Gene structure decoding component: this module defines the gene structure states of eukaryotic species and reconstructs biologically valid gene structures by applying soft connection and decoding algorithms to the position-wise prediction of each segment. **d**, Overview of modeling modules in the neural network component. (i) The distal information modeling module combines convolutional layers, which capture local patterns within subregions, with Transformer encoder layers that compute interactions between subregions, enabling the model to effectively integrate distal information. (ii) The joint evolutionary modeling module models eight distinct evolutionary relation using species with broad evolutionary diversity. A relationship computation controller integrates these sublineages, allowing the module to collectively determine the overall predictions. The weight of the *i*th expert is denoted by W_i, and X_i represents the feature representation produced by the *i*th expert. The final feature representation is computed as the weighted sum of the feature representations from all experts. (iii) The resolution restorer module progressively upsamples the local feature representations back to the original nucleotide resolution, ensuring precise localization of fine-grained sequence features in the output.

Vertebrate_Mammalia and Vertebrate_other (classified under RefSeq's taxonomic groupings; Fig. 2a and Supplementary Figs. 1–5). For a rigorous baseline comparison, we implemented Augustus under optimized conditions where species-specific training was conducted for each test organism using evolutionarily proximate relatives (Supplementary Note 1). ANNEVO demonstrated consistent superiority across all five clades (Fig. 2b), achieving average improvements of 8.8–25.6% in nucleotide-level F1 score (Supplementary Note 2), 13.5–45.2% in gene-level F1 score and 11.6–37.8% in BUSCO scores (Supplementary Note 3) compared to Augustus, with each range representing the minimum-to-maximum mean performance gains across the five clades (Supplementary Tables 1–5). Notably, these gains persisted despite ANNEVO's use of a species-agnostic model—applying the same gene model to all species within a clade—whereas Augustus predictions relied on bespoke models tailored to individual species. This performance disparity underscores ANNEVO's capacity to distill cross-species evolutionary signals into a unified predictive framework. By integrating joint evolutionary constraints through its dynamic network architecture (Fig. 1d), ANNEVO resolves the accuracy limitations inherent to conventional homology-driven methods that require species-specific parameterization.

Beyond annotation accuracy, ANNEVO exhibited a fivefold speed improvement over Augustus under identical 48-thread parallelization configurations (Fig. 2c, Supplementary Table 6 and Supplementary Note 4), attributable to parallelization optimizations in its gene structure decoding component. This efficiency translates to practical benefits for large-scale genome projects, enabling ANNEVO to annotate the *Arabidopsis* genome in just 3.4 minutes and the human genome in 1.4 hours, making it an ideal solution for rapid and comprehensive genome annotation.

https://doi.org/10.1038/s41592-026-03036-7

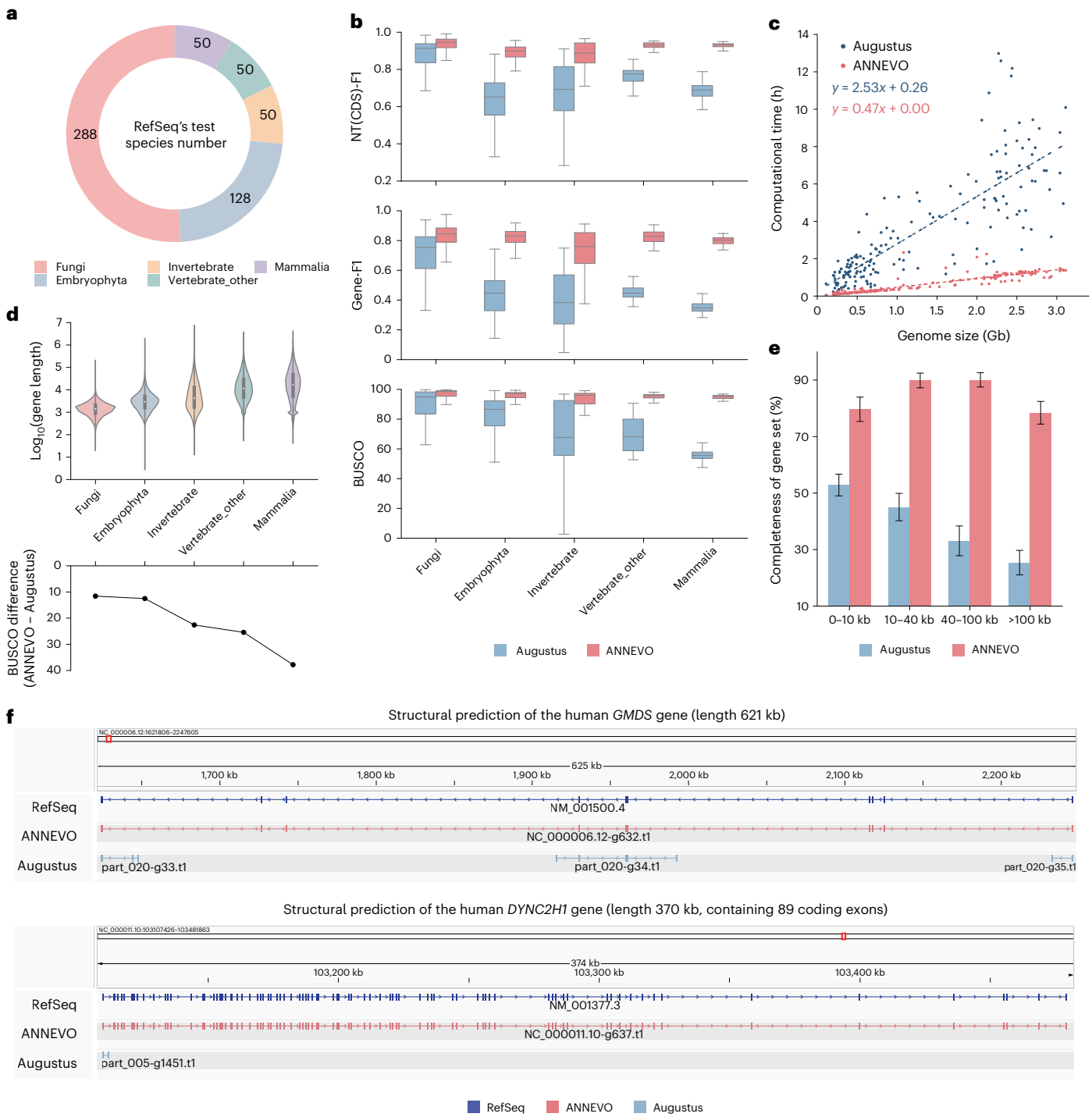


Fig. 2 | Performance evaluation. a, Numbers of species in RefSeq's test set. All candidate test species from Fungi and Embryophyta were evaluated. For Metazoan lineages (Invertebrate, Vertebrate_other and Mammalia), 50 species were randomly subsampled from candidate test sets to balance taxonomic diversity and computational feasibility, prioritizing model organisms and phylogenetically representative taxa. **b**, Performance comparison of ANNEVO and Augustus across different clades in the RefSeq's test set. Three evaluation metrics were calculated at nucleotide (NT), gene and highly conserved BUSCO gene levels. The boxplot displays the median, first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile). The box represents the interquartile range (IQR) (Q1–Q3). Whiskers extend to the most extreme data points within 1.5× IQR from the quartiles. **c**, Computational efficiency comparison. Augustus runtime reflects gene decoding time only, while ANNEVO includes both neural network inference and gene decoding. All benchmarks were conducted on GPU cluster and CPU cluster (Supplementary Note 17). **d**, Correlation between coding gene length distributions and BUSCO score differentials. Gene length defined as the distance between the start and stop codons of longest transcript (including introns). Violin plots show the distribution of gene lengths based on kernel density estimation. The central line represents the median. The bounds of each box correspond to the 25th and 75th percentiles (IQR). Whiskers extend to the most extreme data points within 1.5× IQR. The gene counts for the five lineages were 2,989,915; 4,211,458; 827,430; 1,097,831 and 1,035,058, respectively. BUSCO differentials represent ANNEVO's improvement over Augustus. **e**, Proportion of fully predicted genes by gene length groups. The bar chart compares the proportion of fully predicted genes by ANNEVO and Augustus across various gene length groups from Mammalia. Bar heights represent mean values, with error bars indicating standard deviations. **f**, Long gene modeling capability. ANNEVO accurately reconstructed the 621-kb human *GMDS* locus and 89 coding exons in the *DYNC2H1* locus, while Augustus fragmented it into multiple erroneous gene predictions with misannotated splice junctions.

ANNEVO achieves length-robust gene prediction

ANNEVO demonstrated differential performance gains across phylogenetic clades, with the most substantial improvements observed in Mammalia, likely due to their characteristically longer gene lengths (Fig. 2d). To rigorously assess this length-dependent performance, we stratified 1,035,026 protein-coding genes from all mammalian test species into four length groups: 0–10 kb, 10–40 kb, 40–100 kb and >100 kb, and quantified fully predicted genes per group (Supplementary Note 5). Quantitative analysis revealed a dramatic decline in Augustus's performance with increasing gene length. While Augustus achieved 52.9% complete prediction for short genes (0–10 kb), its performance plummeted to a mere 25.2% for ultra-long genes (>100 kb), indicating a severe length-dependent limitation (Fig. 2e and Supplementary Table 7). This limitation stems from its short-term memory.

In marked contrast, ANNEVO exhibited length-robustness, maintaining or even improving its prediction accuracy for longer genes (Fig. 2e and Supplementary Table 8). ANNEVO accurately reconstructed the complete structures of 13 genes longer than 400 kb (Fig. 2f and Supplementary File 1), including the longest case, the human *GMDS* gene spanning 621 kb. For the *DYNC2H1* gene, which features a highly complex splicing structure with 89 coding exons, ANNEVO also achieved perfectly accurate reconstruction (Fig. 2f). In contrast, Augustus fragmented these genes into multiple incomplete genes, introducing misannotated splice sites. This superior performance in predicting long genes is directly attributable to ANNEVO's Hi-C-inspired long-range dependency modeling, which effectively captures the complex interactions within these extended genome regions.

ANNEVO outperforms integrative evidence-based pipelines

Despite its evidence-free nature, we sought to assess ANNEVO's performance against state-of-the-art evidence-based pipelines, which represent current gold standard for gene annotation. Integrated annotation pipelines leveraging transcriptomic and homology data remain the community standard for species with accessible evidence. To rigorously evaluate ANNEVO's performance across methodological paradigms, we augmented Augustus with repeat masking provided by RefSeq, as well as RNA-seq and proteome evidence for comparison (Supplementary Note 1). We further included two state-of-the-art evidence-guided pipelines, GeneMark-ETP[12] and BRAKER3[11], which were provided with the same evidence as above (Supplementary Note 6), as well as two publicly available deep learning-based annotation tools, Helixer[17,18] (Supplementary Note 7) and Tiberius[19] (Supplementary Note 8), both of which rely solely on genomic sequence information, similar to ANNEVO. Among these, Augustus-Evidence, GeneMark-ETP, BRAKER3 and Helixer supports evaluation across all five clades, whereas Tiberius is restricted to two vertebrate-related clades. Benchmarks spanned 12 model species across Fungi, Embryophyta, Invertebrates, Mammalia and Vertebrate_other, and included a dedicated comparison against Tiberius across all the mammalian species tested.

We optimized evidence-dependent methods with extensive RNA-seq data from five tissues (Supplementary Table 9) and proteomes from evolutionarily proximate species, intentionally favoring these approaches. Even with this deliberate advantage given to evidence-based pipelines, ANNEVO still demonstrated comparable or even superior performance. For example, in *Sus scrofa*, ANNEVO surpassed all other tools across three key metrics, achieving 21.6% increase in full-gene completeness and a 20.1% improvement in BUSCO scores compared to BRAKER3, while maintaining lower false-positive rates (Fig. 3a–c). This demonstrates ANNEVO's ability to deliver superior annotations even when compared to highly optimized evidence-based pipelines.

When further evaluating the generalization capacity of various methods across phylogenetic clades, ANNEVO consistently outperformed all other tools across the 12 representative model species (Fig. 3d, Extended Data Figs. 3 and 4 and Supplementary Table 10). Given that Tiberius was exclusively developed for Mammalia and officially recommended only for vertebrate genomes, we conducted a focused comparison of average performance across six vertebrate species (Extended Data Fig. 5a and Supplementary Note 9). Tiberius showed the smallest performance gap compared to ANNEVO within the Mammalia clade, demonstrating its strong annotation capabilities for mammalian model species. However, Tiberius exhibited a pronounced performance decline in the Vertebrate_other clade, highlighting its taxonomic limitations (Extended Data Figs. 3 and 4). Given that Tiberius is most optimized for mammalian genomes, we extended the comparison to all mammalian species tested (Supplementary Note 9). Tiberius tends to produce more conservative predictions, exhibiting 2.3% higher gene precision than ANNEVO. However, this conservative strategy resulted in a 4.5% reduction in gene-recall compared to ANNEVO, and correspondingly lower BUSCO scores (Extended Data Fig. 5b,c and Supplementary Tables 11 and 12). These differences in prediction tendencies may stem from gene length modeling during the decoding process (Supplementary Note 10). The conservative decoding strategy adopted by Tiberius may lead to the omission of certain coding regions, resulting in notably lower NT(CDS)-F1 and BUSCO scores across all evaluated species (Extended Data Fig. 5b). Overall, as a generalizable framework applicable across clades with diverse gene length distributions, ANNEVO demonstrates superior overall performance.

Beyond accuracy, ANNEVO offers a substantial practical advantage in terms of computational efficiency and resource usage. As Tiberius does not support Fungi, Embryophyta and Invertebrate, it was excluded from evaluations on model species from these clades. All other methods were evaluated on all 12 representative model species. Runtime was assessed for all methods under both graphical processing unit (GPU)-enabled and central processing unit (CPU)-only environments (Supplementary Table 13). On GPU-enabled computing nodes, ANNEVO was approximately 5× faster than Tiberius, 24× faster than Helixer and 83× faster than BRAKER3 (BRAKER3's time excludes RNA-seq alignment; Extended Data Fig. 5d). Notably, even in CPU-only environments, ANNEVO maintains greater than 10× speed advantages over all competing methods, with a particularly striking more than 20× faster execution compared to Tiberius, highlighting its low dependence on GPU infrastructure. Furthermore, ANNEVO exhibits notably lower GPU memory requirements than other deep learning-based approaches. For mammalian-scale genomes under identical batch size configurations (set to 8), ANNEVO consumed only 3 GB of GPU memory, while Helixer required 8.6 GB and Tiberius exceeds 32 GB (Supplementary Note 8). ANNEVO exhibits low GPU memory requirements and scalability independent of genome size, making it suitable for efficient annotation across genomes of varying sizes.

ANNEVO improves reference annotation completeness and quality

A key advantage of ANNEVO's evidence-free approach is its ability to circumvent annotation errors caused by missing or incomplete external evidence data, thereby enhancing the completeness of reference annotations in RefSeq and Ensembl for numerous species. Across 793 species (566 from RefSeq and 227 from Ensembl), ANNEVO achieved higher BUSCO scores compared to reference annotations in 318 species (40.1% of total), demonstrating its ability to improve annotations across all five phylogenetic clades (Fig. 4a and Supplementary Tables 1–5 and 14–18). The improvements were particularly striking in Fungi, Embryophyta and Vertebrate_other, where ANNEVO outperformed Ensembl annotation in 60% (24 out of 40), 72% (60 out of 83) and 68% (34 out of 50) of species, respectively (Fig. 4b). These improvements demonstrate ANNEVO's capacity to complement existing annotation frameworks through its evidence-agnostic methodology.

To further validate whether these improvements are correct and reflect genuine biological signals rather than annotation artifacts constrained to highly conserved BUSCO genes, we conducted a detailed

https://doi.org/10.1038/s41592-026-03036-7

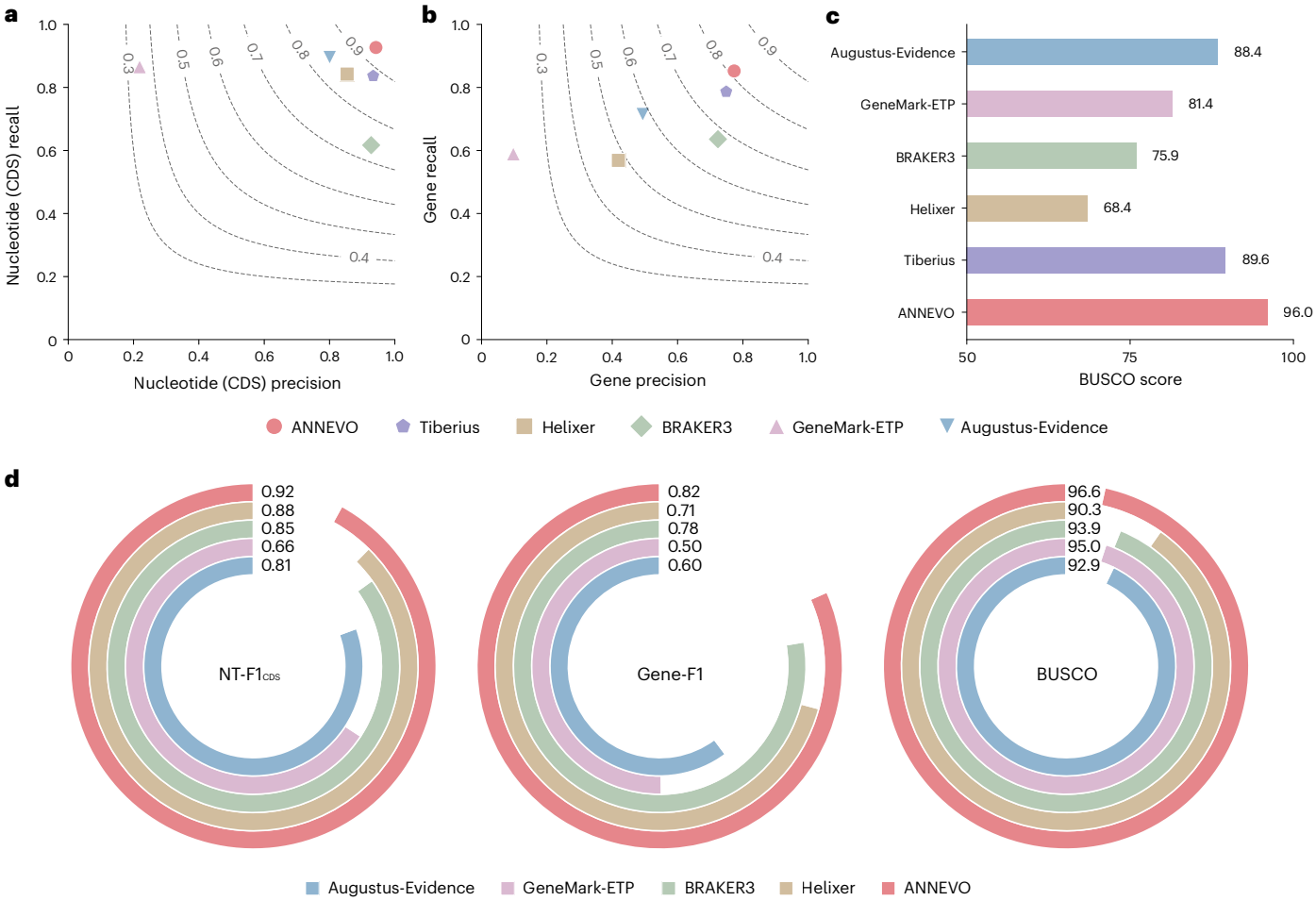


Fig. 3 | Benchmarking against both evidence-assisted annotation pipelines and deep learning methods on model species. a, Nucleotide-level performance on *Sus scrofa*. ANNEVO achieved optimal F1 scores (0.934), outperforming BRAKER3 (0.741) in integrative metrics of completeness and false-positive rate. **b**, Gene-level annotation accuracy on *Sus scrofa*. ANNEVO attained 85.2% gene completeness at comparable precision to BRAKER3, surpassing BRAKER3 by 21.6% absolute completeness. **c**, BUSCO comparison on *Sus scrofa*. ANNEVO recovered 96% complete BUSCOs, demonstrating superior ortholog detection capability. **d**, Average performance on 12 model organisms. Using only genomic sequences, ANNEVO outperforms BRAKER3, which incorporates multi-tissue RNA-seq and closely related species' proteins, while substantially exceeding the deep learning baseline Helixer (improvements: 4% nucleotide-level F1, 11% gene-level F1, 6.3% BUSCO completeness).

locus-specific analysis in *Brassica oleracea* (Ensembl database). ANNEVO identified 55 additional BUSCO genes with full RNA-seq support, representing over 3% of total BUSCO orthologs (Fig. 4c, Supplementary File 2 and Supplementary Note 11). A representative example revealed that Ensembl's annotation erroneously fused two distinct genes, a structural error unsupported by spliced transcript evidence (Fig. 4d). Beyond gene fusion errors, other typical types of corrected error included spurious gene losses, incomplete annotations caused by premature termination codons and misassigned start codons caused by incorrect splice sites (Extended Data Fig. 6). Crucially, this improvement extended beyond the conserved BUSCO gene set. Within proximal regions of these corrected BUSCO gene loci, ANNEVO provided 23 additional RNA-seq-supported corrections of erroneous Ensembl annotations, predominantly involving non-BUSCO genes with lower evolutionary conservation (Supplementary File 3). This demonstrates the broad applicability of ANNEVO's evidence-agnostic framework across genes with varying evolutionary pressures.

Model ablation and fine-tuning

Understanding the contributions of key design and training decisions in ANNEVO is critical for promoting model development in genome annotation tasks. To this end, we conducted ablation studies to evaluate the contribution of ANNEVO's three core modules (Supplementary Note 12). First, we found that the joint evolutionary modeling module is the most important for achieving best performance. When the MoE was removed from ANNEVO while keeping parameter sizes unchanged, the three metrics dropped by 2.3%, 7.8% and 6.3%, respectively (Extended Data Fig. 7a and Supplementary Table 19). This indicates that ANNEVO's use of MoE to model cross-clade relationships plays a key role in enhancing gene annotation accuracy and generalizability. Next, we evaluated how the resolution restorer module contributes to performance. We removed this module and instead directly mapped the encoded features to a one-to-one nucleotide-level prediction. Despite maintaining the same parameter sizes, this modification also led to a decline in ANNEVO's performance, suggesting that the resolution restorer is essential for precise base-resolution inference.

We further tested the distal information modeling module by replacing it with architectures from other sequence models to assess its irreplaceable role within ANNEVO. SpliceAI[29], a widely used model for splice site prediction and highly related to gene annotation, was adopted as a substitute for ANNEVO's long-range modeling component. However, this replacement resulted in a 2.5-fold increase in GPU requirement and an 8.8-fold (Extended Data Fig. 7b) reduction in training speed, extending the projected training time for Mammalia model

to an impractical 56 days. While computational overhead of such architecture may be acceptable for tasks with limited data (for example, SpliceAI is restricted to human transcriptomic regions), it becomes a critical bottleneck when scaling to multi-species and whole-genome annotation tasks. This highlights the value of ANNEVO's design principle in its distal information modeling module, which effectively balances long-range modeling capacity with computational efficiency.

While ANNEVO is designed as a generalizable model applicable across the entire evolutionary tree, certain use cases may prioritize achieving optimal performance within narrowly defined phylogenetic groups over broad generalizability. In such cases, targeted fine-tuning strategies focused on specific subclades or species of interest warrant exploration. To investigate this, we conducted targeted fine-tuning experiments on two representative species: *Homo sapiens*, a widely studied model organism, and *Echinops telfairi*, which exhibited the largest discrepancy in BUSCO completeness between ANNEVO and RefSeq among all mammalian species. For each target species, we fine-tuned ANNEVO using closely related species within the same subclade and evaluated the impact of different fine-tuning strategies (Supplementary Note 13). As a baseline, we also fine-tuned a non-MoE version of ANNEVO with an equivalent number of parameters. While full-parameter fine-tuning of ANNEVO did not yield performance improvements, likely due to overfitting risks and instability introduced by the MoE architecture[30], fine-tuning only the task-specific module led to modest performance gains (Extended Data Fig. 7c and Supplementary Table 20), and substantially outperformed the fine-tuned non-MoE model, highlighting the value of modular fine-tuning strategies in subclade-specific applications.

Discussion

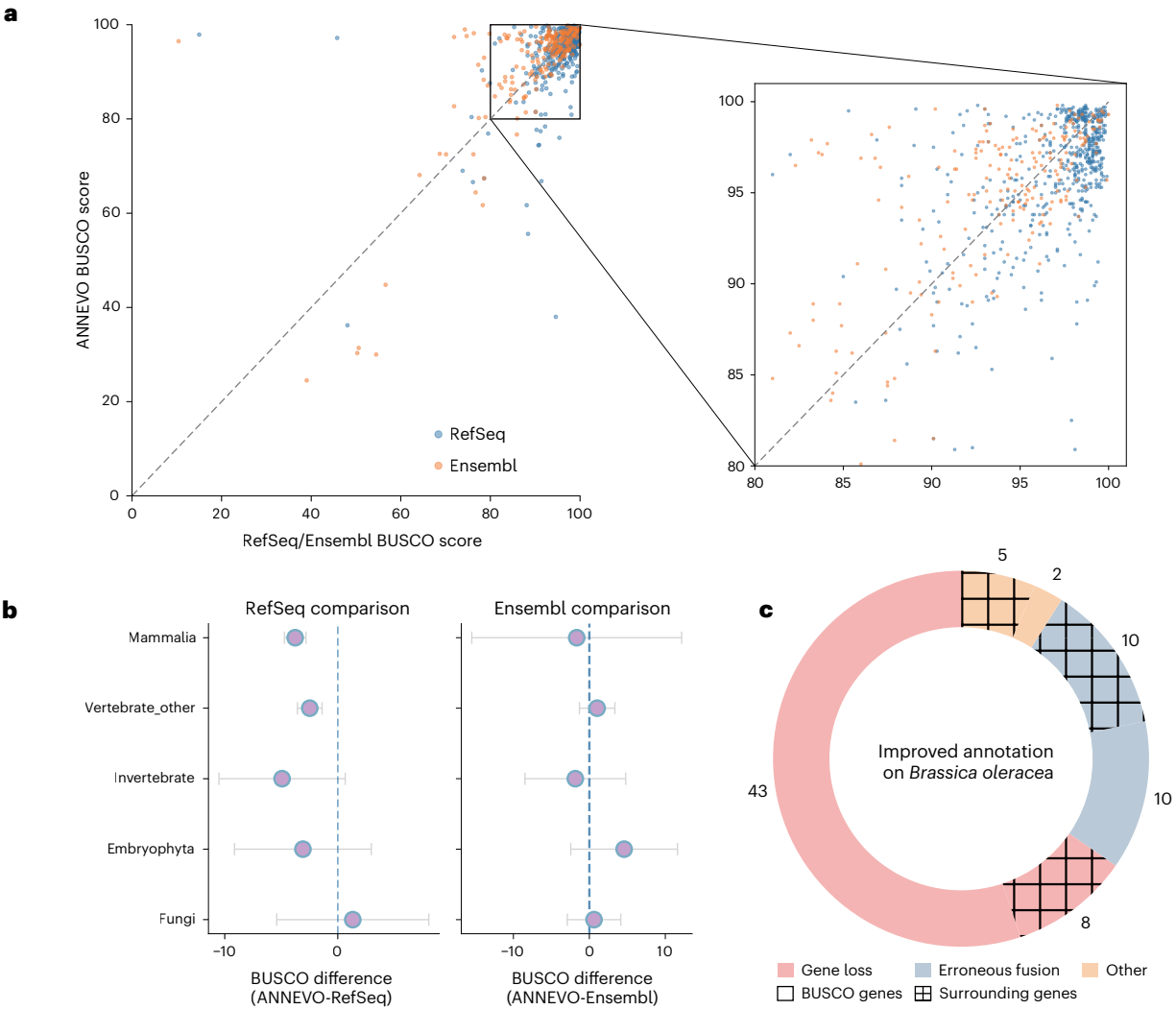
ANNEVO represents a important advancement in ab initio gene annotation, providing high-quality gene predictions directly from genomic sequences, thereby eliminating the dependence on often-limited extrinsic evidence. This constitutes the most significant distinction between ANNEVO and Augustus's derivative family of tools. This independence is particularly crucial given the substantial data paucity in many species. Among the 566 RefSeq species in our test set, a staggering 46% (261 out of 566) lack any transcriptomic data, while 72% (410 out of 566) have fewer than five transcriptomic data available in the NCBI Sequence Read Archive (Supplementary Table 21). Even in species with abundant transcriptomic resources, spatiotemporal expression variability can lead to incomplete gene annotations. This highlights the critical need for evidence-free annotation tools, especially for newly sequenced genomes generated by large-scale genomic initiatives. Even for genomes with existing annotations, ANNEVO serves as a potent validation framework to refine reference database entries, as demonstrated by its capacity to correct erroneous gene models in Ensembl/RefSeq annotations through intrinsic sequence pattern recognition. This dual functionality—as both a primary annotation engine for emerging genomes and a refinement tool for established databases—underscores its broad utility in modern genome annotation pipelines.

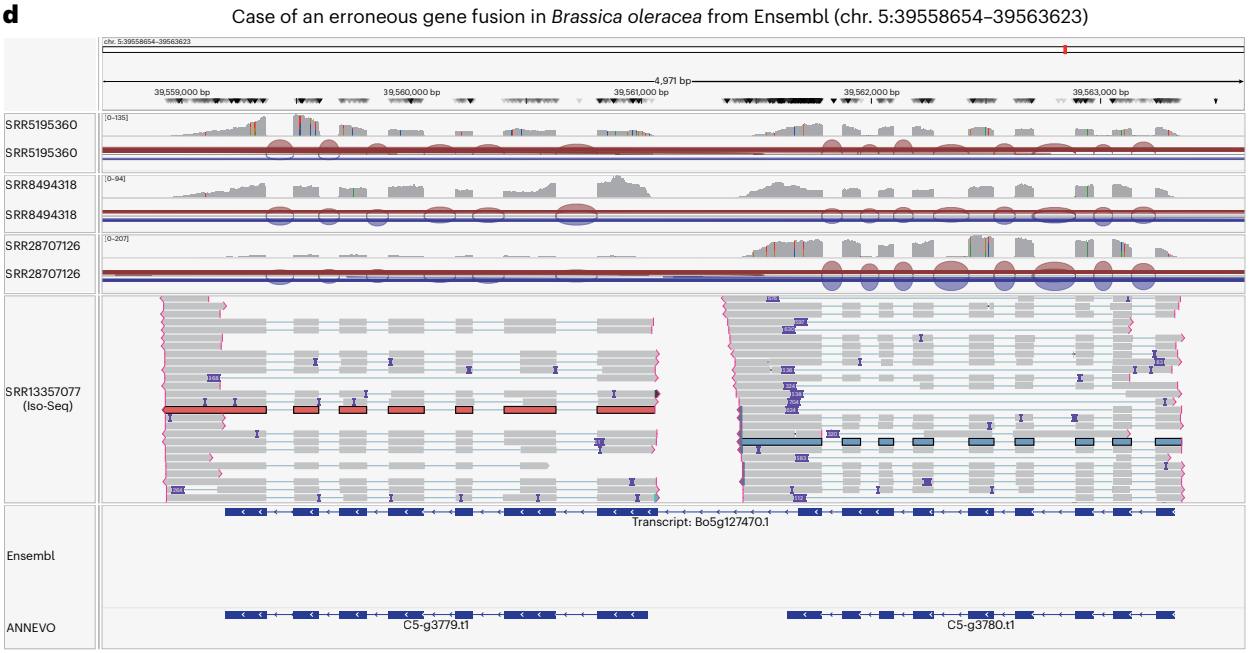
A critical consideration in ANNEVO lies in computational efficiency. Recent advances in genomic large language models[31–33] (LLMs) have demonstrated notable success across various genome analysis tasks. However, the extreme computational demands of such models remain a major barrier to real-world deployment. Even alternative architectures such as SpliceAI, which forgo the computationally intensive Transformer in favor of a convolutional design, have not achieved improvements in training speed or computational cost. ANNEVO addresses this challenge through a deliberate trade-off between modeling capacity and practical deployment. In particular, the efficiency stems from its distal information modeling module and decoding optimizations, which constrain search space and enhance parallelism. Unlike approaches that rely heavily on GPU acceleration, ANNEVO's high computational efficiency results from principled architectural and algorithmic design choices. This substantially reduces its dependency on GPUs and enables broader applicability. Such efficiency is especially critical in genome-scale applications. Unlike protein-centric tasks, which typically operate at the resolution of individual sequences where minute-level processing times are acceptable, genome annotation requires architectures capable of maintaining high throughput across contig-scale inputs. As such, it is essential to avoid reintroducing computational bottlenecks, which could otherwise undermine the scalability of genome annotation and impede the progress of large-scale genomic initiatives.

At present, ANNEVO is limited to annotating only the longest transcript of protein-coding gene, in contrast to conventional pipelines that support multi-isoform annotation. To ensure consistency in benchmarking, we adopted the longest transcript from reference annotations as the evaluation target. Notably, allowing a prediction to match any annotated isoform (rather than only the longest) would further improve ANNEVO's reported performance (Extended Data Fig. 7d and Supplementary Table 22). Another important aspect of evaluation concerns the nature of ANNEVO's false positives (Supplementary Note 14). A detailed analysis revealed that 60.6% of these cases arise from fragmented messenger RNA predictions, while 18.3% exhibit exon-level overlap with annotated noncoding transcripts (Extended Data Fig. 7e). These findings suggest that ANNEVO's false positives still show biological relevance, supporting the potential of extending ANNEVO to more comprehensive genome annotation tasks beyond protein-coding genes in the future. Expanding its capabilities to include noncoding RNAs therefore represents a promising future direction. While traditional sequence modeling approaches often struggle to capture essential RNA secondary structures[34], representation learning strategies, such as seq2img techniques[35,36], show considerable potential in modeling structural patterns such as RNA hairpins. A further challenge relates to underrepresented regions in the training data, notably untranslated regions and rare splicing patterns. The high variability of untranslated regions, caused by alternative transcription start sites[37] and alternative polyadenylation[38], poses significant challenges for benchmarking annotation accuracy in these regions. This phenomenon is illustrated by examples in Supplementary Files 2 and 3 and is further supported by previous Iso-Seq analysis[39]. Rare splicing events, such as AT-AC introns, also present unique challenges. Their low frequency in the genome renders them difficult to accurately predict for all current ab initio annotation tools, including SpliceAI, a state-of-the-art model specifically designed for splice site prediction in the human genome (Supplementary Note 15). Future work incorporating fragmented and

Fig. 4 | Comparative evaluation with reference databases and annotation improvements. a, BUSCO completeness comparison between ANNEVO and reference annotations across 793 species. ANNEVO achieved >90% BUSCO completeness in the most species and outperformed reference annotations in 40.1% (318 out of 793) of cases. **b**, Distribution of BUSCO score differences between ANNEVO and reference annotations. The sample sizes for each clade and the detailed values are provided in Supplementary Tables 1–5 and 14–18. ANNEVO demonstrated superior performance in at least some species within each major phylogenetic clade, with particularly significant improvements in Embryophyta (mean BUSCO: 95.6% versus 91%; $n = 83$, $P = 1.8 \times 10^{-6}$, two-sided Mann–Whitney *U*-test). Bar heights represent mean values, with error bars indicating standard deviations. **c**, Correction of annotation errors by ANNEVO in *Brassica oleracea*. ANNEVO corrected 55 annotation errors (Supplementary File 2) on BUSCO genes and 23 annotation errors (Supplementary File 3) on surrounding genes in Ensembl. These corrections addressed distinct error types, all validated by independent RNA-seq evidence. **d**, Representative example of gene fusion correction. ANNEVO resolved an Ensembl annotation error where two distinct genes were erroneously merged at a splice junction unsupported by RNA-seq reads and Iso-Seq. This correction properly separated the BUSCO gene (left) from its neighboring gene (right), as visualized in the IGV.

https://doi.org/10.1038/s41592-026-03036-7





multi-modal evidence may help alleviate these issues and further position ANNEVO as a flexible annotation platform capable of modeling hierarchical interdependencies across diverse genomic features.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03036-7.

References

1. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).
2. Lewin, H. A. et al. The Earth BioGenome Project 2020: starting the clock. *Proc. Natl Acad. Sci. USA* **119**, e2115635118 (2022).
3. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
5. Korf, I. Gene finding in novel genomes. *BMC Bioinf.* **5**, 59 (2004).
6. Lukashin, A. V. & Borodovsky, M. GeneMark. HMM: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
7. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
8. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
9. Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M. & Kitts, P. Eukaryotic genome annotation pipeline. *The NCBI Handbook* Vol. 2 (National Center for Biotechnology Information, 2013).
10. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* **12**, 1–14 (2011).
11. 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).
12. 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).
13. Brůna, T. et al. Galba: genome annotation with miniprot and AUGUSTUS. *BMC Bioinf.* **24**, 327 (2023).
14. Aken, B. L. et al. The Ensembl gene annotation system. *Database* **2016**, baw093 (2016).
15. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
16. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4.10. 11–14.10. 14 (2004).
17. Holst, F. et al. Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *Nat. Methods* https://doi.org/10.1038/s41592-025-02939-1 (2025).
18. Stiehler, F. et al. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* **36**, 5291–5298 (2021).
19. Gabriel, L., Becker, F., Hoff, K. J. & Stanke, M. Tiberius: end-to-end deep learning with an HMM for gene prediction. *Bioinformatics* **40**, btae685 (2024).
20. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
21. Sætre, G. P. & Saether, S. A. Ecology and genetics of speciation in Ficedula flycatchers. *Molecular Ecology* **19**, 1091–1106 (2010).
22. Parkin, I. A. et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol.* **15**, R77 (2014).
23. Chen, L., DeVries, A. L. & Cheng, C.-H. C. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl Acad. Sci. USA* **94**, 3811–3816 (1997).
24. Vosseberg, J. et al. The emerging view on the origin and early evolution of eukaryotic cells. *Nature* **633**, 295–305 (2024).
25. Zhou, Y. et al. Gene fusion as an important mechanism to generate new genes in the genus *Oryza*. *Genome Biol.* **23**, 130 (2022).
26. Bang, M.-L. et al. The complete gene sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ. Res.* **89**, 1065–1072 (2001).
27. Vaswani, A. et al. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)* https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (2017).
28. Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **13**, 260–269 (1967).
29. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
30. Zoph, B. et al. St-MoE: designing stable and transferable sparse expert models. Preprint at https://arxiv.org/abs/2202.08906 (2022).
31. Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
32. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. Preprint at *bioRxiv* https://doi.org/10.1101/2025.02.18.638918 (2025).
33. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
34. Harrison, P. W. et al. Ensembl 2024. *Nucleic Acids Res.* **52**, D891–D899 (2024).
35. Wang, S. et al. De novo and somatic structural variant discovery with SVision-pro. *Nat. Biotechnol.* **43**, 181–185 (2025).
36. Lin, J. et al. SVision: a deep learning approach to resolve complex structural variants. *Nat. Methods* **19**, 1230–1233 (2022).
37. de Klerk, E. & t Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* **31**, 128–139 (2015).
38. Xia, Z. et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′- UTR landscape across seven tumour types. *Nat. Commun.* **5**, 5274 (2014).
39. Zhang, R. X. et al. A high-resolution single-molecule sequencing-based *Arabidopsis* transcriptome using novel methods of Iso-seq analysis. *Genome Biol.* **23**, 149 (2022).

Methods

Data sources and processing

Data download and taxonomic levels. All data were acquired using the latest genome assemblies and annotations as of July 2024. RefSeq data were downloaded from NCBI's FTP site (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/), containing 1828 species. The version details for each species are provided in Supplementary Tables 23–27. Ensembl data were downloaded from various sources, containing 428 species: animals from Ensembl release 112 (https://ftp.ensembl.org/pub/release-112/), fungi from Ensembl Fungi release 59 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/fungi/release-59/) and plants from Ensembl Plants release 59 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-59/). Species with critical errors preventing data parsing were excluded. For instance, in RefSeq's *Actinidia eriantha* (version GCF_019202715.1_MaoHua_MHT), sequence with GenBank ID NC_034914.1 has a defined range of 1–156,964, but includes a gene region spanning 156,727–157,788, which exceeds this range. Following RefSeq's taxonomic levels, Ensembl species were grouped into five clades: Fungi, Embryophyta, Mammalia, Vertebrate_other (nonmammalian vertebrates) and Invertebrate. Separate models were then trained for each clade (Supplementary Note 16).

Database selection. To ensure optimal training and evaluation, species within each clade were assessed using their BUSCO scores[40]. The database with the higher average BUSCO score for species within a given clade was selected as the primary source. Based on the comparison analysis (Supplementary Fig. 6a,b and Supplementary Tables 28–32), RefSeq was chosen as the primary source for training and evaluation across all five clades, and Ensembl was used for supplementary evaluation of ANNEVO's completeness.

Species usage. Given the uneven distribution of genome annotation quality across the evolutionary tree, selecting species solely based on high BUSCO scores could result in the exclusion of entire sublineages with generally lower scores. This would introduce bias in model training and benchmarking. To balance evolutionary diversity with annotation quality, we first divided each of the five major clades into several subclades (Supplementary Table 33). Within each subclade, species were selected based on the top percentile of BUSCO scores, ensuring that each subclade was represented by at least one species and that selected species had relatively high annotation quality.

To manage training resource constraints, we capped the total genome size per clade at approximately 50–60 gigabases (Gb) (for fungi, which have small genomes, the total genome size was limited to 12 Gb). As a consequence, the top percentile used for species selection varied across clades (Supplementary Table 33). For example, within Mammalia, we selected the top 5.5% (this percentile cutoff was applied solely to control the total genomic scale during training) of species per subclade, resulting in a total genome size of 61 Gb. Selected species were then randomly assigned to either the training or validation set, maintaining a genome size ratio of approximately 5/1 between training and validation. This strategy ensured cross-species training and evaluation to avoid data leakage. Furthermore, by randomly partitioning at a relatively high taxonomic level, we ensured that the validation set species were not closely related to those in the training set, thereby strengthening our cross-species study paradigm. Species not used in the training process were designated as candidates for the test set. For Fungi and Embryophyta, all candidate species were included in the final test set due to their relatively small genome sizes. For the remaining three animal clades, due to large genome sizes and computational limitations, we randomly selected 50 species per clade from the candidate test pool. This selection deliberately included model organisms not used during training, allowing for high-confidence comparisons with other deep learning models and annotation pipelines. Detailed species usage (training, validation or testing) is provided in Supplementary Tables 23–27.

ANNEVO methodology

Context extension. The genome was systematically segmented into consecutive core regions, each spanning 30,720 bp, utilizing a sliding window approach. To provide sufficient contextual information for each nucleotide within these core regions, they were extended by 5,120 bp both upstream and downstream, forming the flanking sequences. Any flanking regions shorter than 5,120 bp were padded with zero vectors. ANNEVO utilizes only the sequence information of these flanking regions, without incorporating their annotations for model parameter updates (see section 'Loss masking during model training' for details).

Sequence encoding and label definition. DNA sequences were represented using one-hot encoding, with ambiguous bases (for example, *N*) encoded as the vector [0.25, 0.25, 0.25, 0.25]. Each site was labeled based on the annotation of the longest transcript, with categories defined as intergenic, CDS0, CDS1, CDS2 or intron. The three coding DNA sequence (CDS) labels represent coding sequences in different reading frames (phases 0, 1 and 2, respectively), which is critical for modeling codon structure and ensuring reading-frame consistency.

Sequence modeling with deep learning. ANNEVO is an end-to-end deep learning model comprising three core modules: the distal information modeling module, the joint evolutionary modeling module and the resolution restorer module (Fig. 1d and Extended Data Fig. 1). Given the extreme class imbalance in genomic sequences (for example, the ratio of intergenic to CDS regions can reach tens of thousands to one), we employed focal loss[41] as the primary objective for classification and supplemented it with dice loss[42] to enhance the optimization of minority classes such as CDS and introns. Notably, in the application of focal loss, we adjusted only the focusing parameter that emphasizes harder-to-classify examples, while keeping the class-weight parameter unchanged. This design choice was made to account for the potential inconsistency in class distributions when retraining ANNEVO on alternative datasets or fine-tuning it for specific species. Explicit class reweighting would require careful recalibration across datasets with varying class ratios, potentially undermining the model's generalizability and reusability. Model training was performed on a GPU node equipped with four Tesla V100S. The average GPU memory consumption was approximately 2 GB per sample, and the training process took 6–7 days for the largest dataset (mammalian genomes). Further details regarding the model architecture, loss function, training pipeline and hyperparameter settings can be found in Supplementary Method 2.

(1) Distal information modeling module: this module captures both local patterns and long-range interactions within the sequence. Similar to Hi-C technology, which observes long-range interactions by measuring contact between regions rather than single nucleotides, ANNEVO first learns local sequence patterns through a convolution tower with residual connections[43]. It then leverages positional encoding and encoder layers within a Transformer framework[27] to model long-range interactions between local patterns, enabling effective modeling of sequences up to ~40 kb in length. This approach substantially decreases computational demands and minimizes the model's parameter size. The output of this module is fed into the joint evolutionary modeling module.
(2) Joint evolutionary modeling module: this module draws inspiration from the MoE framework[44], utilizing diverse expert networks to acquire domain-specific knowledge and incorporating a gating mechanism to dynamically allocate weights to each network based on the input data's features. It comprises eight sublineage networks (analogous to the experts in the MoE) and a relationship computation controller (similar to the gate in the MoE). The relationship computation controller

calculates the association between the input sequence and each subnetwork, with each subnetwork tasked with learning specific evolutionary relationships. The final representation is obtained by combining the outputs of the relationship computation controller with the feature representations of the subnetworks through a weighted sum, effectively reflecting the contribution of each sublineage. This strategy allows ANNEVO to model at higher taxonomic levels while simultaneously obtaining detailed information about sublineages and their interconnections. The output of this module is channeled into a resolution restorer module.

(3) Resolution restorer module: this module serves as the inverse process of the initial local feature extractor. It utilizes a hierarchical stack of transposed convolutional layers to progressively upsample the coarse-grained local representations. This stepwise refinement restores the spatial resolution of the features to match the original nucleotide-level input, thereby enabling accurate position-wise predictions required for gene structure annotation. The output provides per-nucleotide probability distributions over five gene structure categories: three reading-frame-specific CDS types, introns and intergenic regions. These probability values are subsequently used as emission probabilities in downstream gene structure decoding (see section 'Gene structure decoding' for details).

Loss masking during model training. During the model training, ANNEVO outputs probabilities only for core regions, and losses associated with the flanking regions are masked to prevent them from affecting the update of model weights. Within the core regions, we selectively mask certain erroneous sites. These sites primarily involve gene structures that clearly violate biological rules or annotations with ambiguous information. For example, some adjacent exons are so close that the inferred intron length is only 2 bp, which contradicts basic splicing rules. In other cases, gene loci are annotated but lack internal structural details, likely due to artifacts introduced by other annotation tools or manual curation. A detailed list of such error types is provided in Supplementary Methods 1. To mitigate their impact, sites identified as annotation errors are assigned a weight of zero. During training, the loss at each site is scaled by its corresponding weight, enabling the model to minimize the adverse effects of potentially erroneous annotations.

Gene structure decoding. The gene structure decoding employs a probabilistic approach to generate biologically valid gene architectures. Initially, potential gene regions are delineated by establishing broad genomic boundaries for each candidate gene. These boundaries serve not as precise discriminative cutoffs, but rather as segmentation markers to enable segment-based parallel processing. This design significantly enhances computational efficiency compared to conventional HMM-based methods that can only be parallelized at the contig-level. The decoding framework leverages ANNEVO's unique capability of single nucleotide-resolution modeling through evolutionary and contextual information. This approach eliminates the need for extensive gene structure states and manual parameter controls as implemented in Augustus. Instead, it operates with a minimal set of biologically defined gene structure states: one intergenic state, start codon state, end codon state, six CDS states and multiple intron state groups. Notably, the six CDS states are designed to represent the three possible reading frames (phases) of coding sequences. This design ensures complete prevention of premature stop codons within a reading frame, even in the presence of codons that span splice junctions (Extended Data Fig. 2). Each CDS state is associated with a set of intron states corresponding to three splicing modes. On completion of an intron state, the decoding process transitions to the next CDS phase state: for example, an intron entered from the CDS0 state will exit into the CDS1 state, rather than looping back to CDS0. Since emission probabilities are extensively informed by the model's evolutionary and contextual learning, it requires no manual parameter tuning for emission probabilities. This design principle effectively prevents the introduction of human bias and preserves the integrity of probabilistic pathways specific to individual genes. Finally, the Viterbi algorithm[28] is employed to calculate the most probable state path within each potential gene region, resulting in the final gene structure. Details on potential gene region detection, state definitions, transition conditions and the use of model predictions can be found in the Supplementary Method 3.

Evaluation metrics. Consistent with previous studies[4,10], this work utilizes nucleotide-level F1 scores (Supplementary Note 2), gene-level F1 scores and BUSCO scores[40] (Supplementary Note 3) to evaluate gene annotation quality. Nucleotide-level F1 scores provide a comprehensive evaluation across all genomic sites and reflect the potential to identify new biological regions, yet it struggles to capture gene structural information. In contrast, BUSCO scores incorporate gene structure but are restricted to a highly conserved core set of genes, which often represent only a minor fraction of the total gene repertoire (for example, BUSCO genes in the Embryophyta database represent only ~6% of *Arabidopsis thaliana* genes). Moreover, BUSCO assessments do not reflect false-positive predictions. The gene-level F1 score serves as a complementary metric that provides genome-wide coverage of gene structural information while accounting for false positives. However, given the incomplete reference annotation of alternative splicing isoforms in most species, correctly predicted isoforms by gene prediction tools may be erroneously classified as false predictions. To mitigate this issue, we referred to the BUSCO protein-mode protocol and implement more rigorous alignment standards to assess gene structure accuracy across total gene repertoire at the protein level. Specifically, the complete set of protein sequences was obtained separately from the predicted annotations and the reference annotations, and a protein blast (blastp) was performed on these two groups of protein sequences. A gene is considered complete at the gene level if it meets all the following criteria:

(1) Bit score and *e*-value thresholds. Only the highest matching segment were considered for each protein, and the highest matching segment between the predicted gene's protein sequence and the reference gene's protein sequence must meet predefined bit score and *e*-value thresholds to confirm similarity, set at 50 and 1×10^{-5}, respectively.

(2) Genomic coordinate overlap. A prediction is considered complete only if the predicted gene and the reference gene are located in the same genomic region.

(3) Protein sequence coverage. Fragmented-prediction or over-prediction may result in a fragment that meets the first two conditions, but the gene structure is obviously only partially correct (either too short or too long). Therefore, the highest matching segment must cover more than 70% of both the predicted and reference protein sequences for the prediction to be considered complete.

For each gene in the reference annotation, only its transcript with longest coding region was considered. The number of genes meeting the above criteria represents the true positives (TP), the total number of genes in the reference annotation represents the positives (*P*) and the total number of predicted genes represents the predicted positives (PP). Consequently, gene F1 can be calculated with the following formula:

$$\text{recall} = \frac{\text{TP}}{P} \quad (1)$$

$$\text{precision} = \frac{\text{TP}}{\text{PP}} \quad (2)$$

$$\text{gene} - \text{F1} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (3)$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sources of RNA-seq are listed in Supplementary Table 9. The genome and annotation of RefSeq are listed in Supplementary Tables 23–27 and are available from the NCBI's FTP site at https://ftp.ncbi.nlm.nih.gov/genomes/refseq/. The genome and annotation of Ensembl are available via the Ensemble page for release 112 at https://ftp.ensembl.org/pub/release-112/, Ensembl Fungi release 59 at https://ftp.ensemblgenomes.ebi.ac.uk/pub/fungi/release-59/ and Ensembl Plants release 59 at https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-59/.

Code availability

ANNEVO is available via GitHub at https://github.com/xjtu-omics/ANNEVO. The repository is free for non-commercial use by academic, government and nonprofit/not-for-profit institutions.

References

40. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. in *Gene Prediction: Methods Protocols* (ed Kollmar, M.) 227–245 (Springer, 2019).
41. Lin, T. Y., Goyal, P., Girshick, R., He, K. M. & Dollár, P. Focal loss for dense object detection. In *Proc. IEEE International Conference on Computer Vision* (eds Ikeuchi, K. et al.) 2999–3007 (IEEE, 2017).
42. Li, X. Y. et al. Dice loss for data-imbalanced NLP tasks. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D. et al.) 465–476 (Association for Computational Linguistics, 2020).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Bajcsy, R. et al.) 770–778 (IEEE, 2016).
44. Fedus, W., Zoph, B. & Shazeer, N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **23**, 1–39 (2022).

Acknowledgements

We thank the Vertebrate Genomes Project and Darwin Tree of Life for helpful comments. K.Y. is supported by National Key R&D Program of China (grant no. 2022YFC3400300) and National Natural Science Foundation of China (grant nos. 32125009 and 32430017). S.W. is supported by National Natural Science Foundation of China (grant no. 323B2015). X.Y. is supported by the National Natural Science Foundation of China (grant nos. 32422019 and 62172325), the Natural Science Foundation of Shaanxi Province (grant no. 2024JC-JCQN-28) and the Fundamental Research Funds for the Central Universities (grant no. xzy012024088). P.J. is supported by the National Natural Science Foundation of China (grant no. 32400509). J.L. is supported by the National Natural Science Foundation of China (grant no. 62302386). D.M. is supported by the National Natural Science Foundation of China (grant no. 12426105) and the Major Key Project of Pengcheng Laboratory (grant no. PCL2024A06).

Author contributions

K.Y. designed and supervised the research. P.Z. developed the ANNEVO algorithm and performed the performance evaluation. T.X. and S.W. contributed to the data processing. X.Y. and B.W. contributed to the training and prediction of Augustus. P.J., P.S. and Y.Z. contributed to the RNA-seq analysis. J.L. contributed to the supplemental analysis of ANNEVO. D.M. contributed to the model ablation and fine-tuning. Z.N. contributed to the evaluation metrics. P.Z., S.J.B., Z.N. and K.Y. wrote the paper with input from all other authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

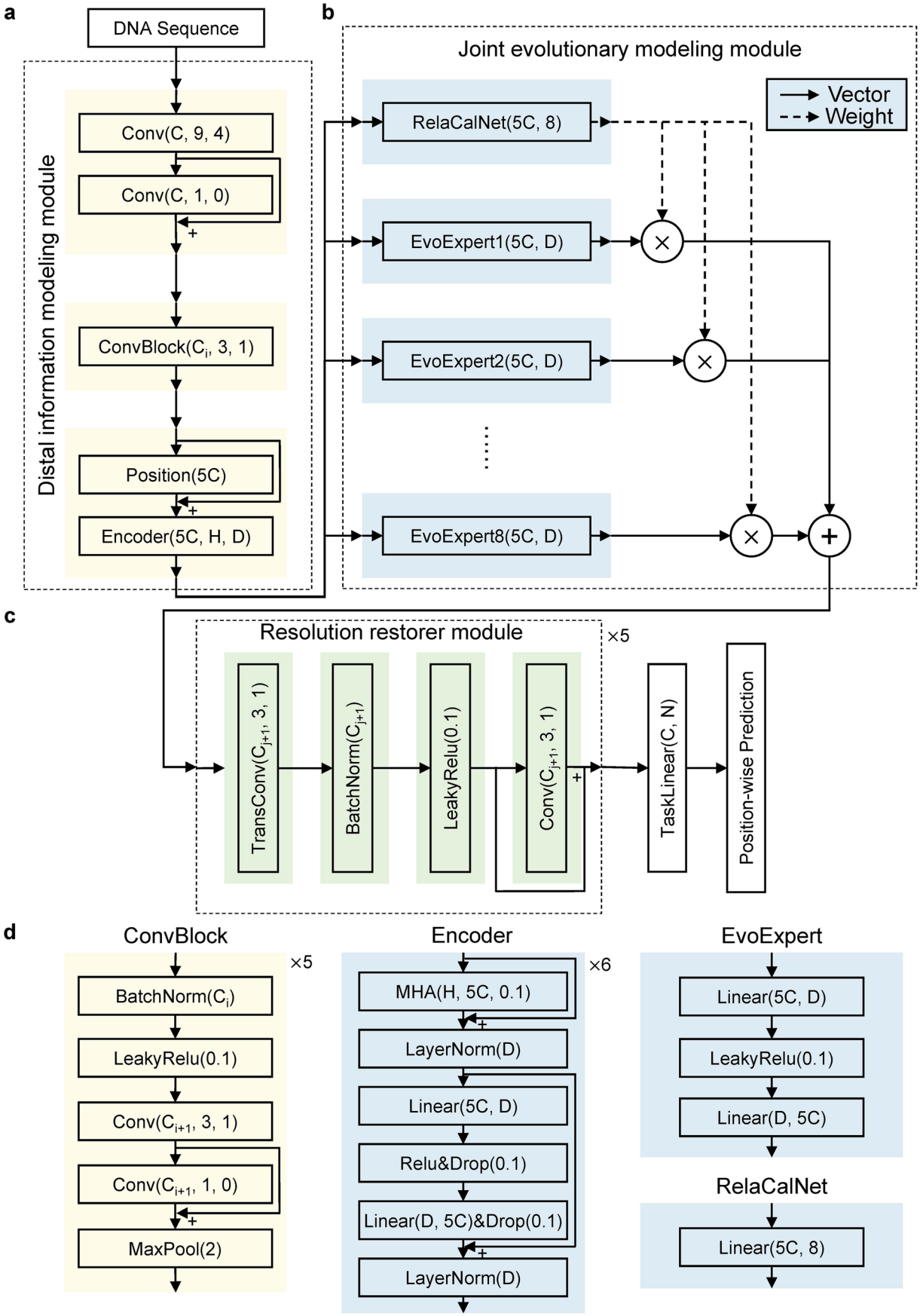
Extended data is available for this paper at https://doi.org/10.1038/s41592-026-03036-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-026-03036-7.

Correspondence and requests for materials should be addressed to Kai Ye.

Peer review information *Nature Methods* thanks Michael Hiller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-026-03036-7

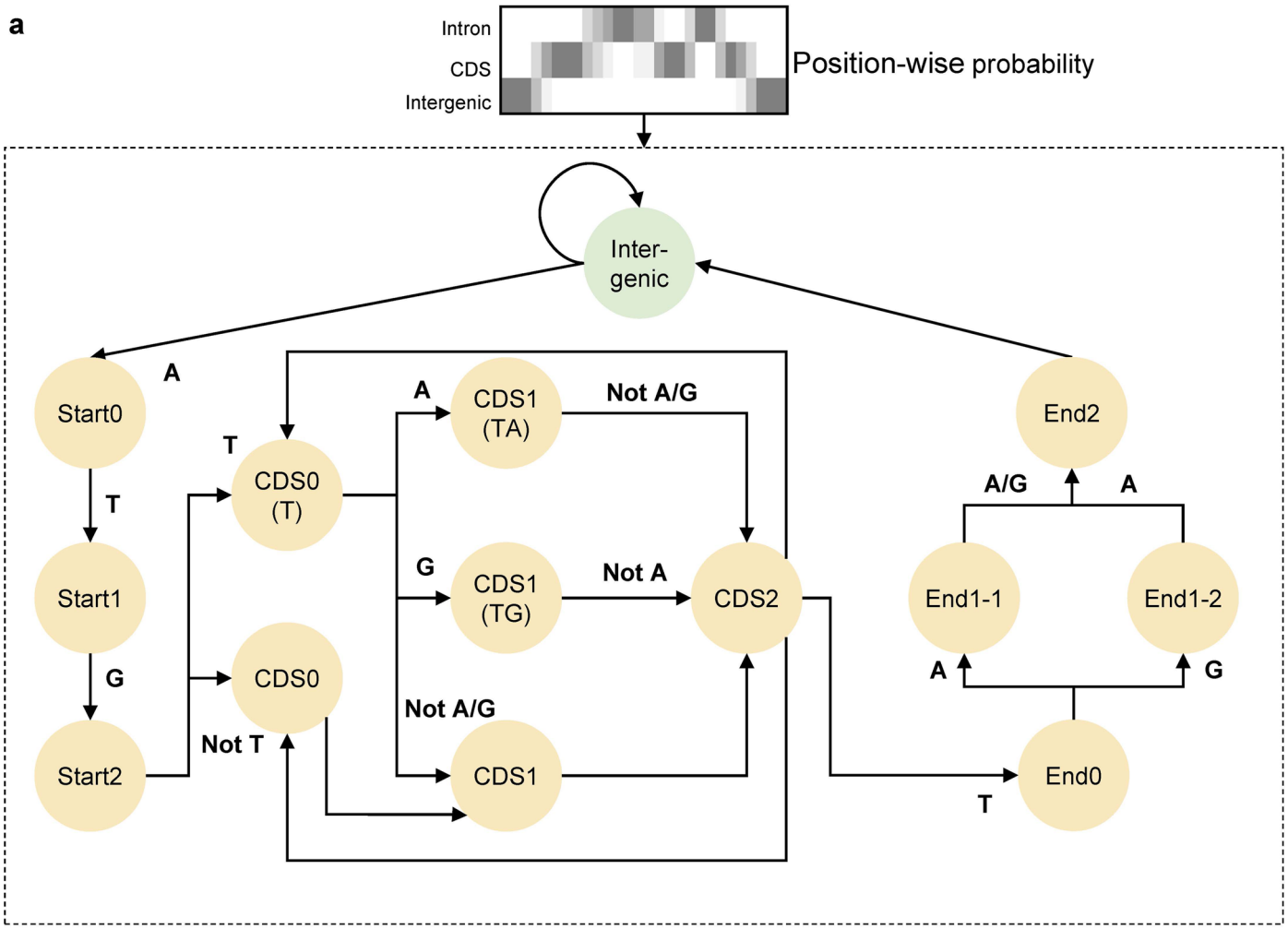


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41592-026-03036-7

Extended Data Fig. 1 | Detailed model architecture of ANNEVO's neural network component. a, Distal Information Modeling Module. This module extracts local sequence patterns using five consecutive ConvBlocks and learns long-range dependencies through positional encoding and Transformer encoder layers. The parameters are as follows: C = 64, H = 8, D = 768. **b**, Joint Evolutionary Modeling Module. The module consists of eight sub-lineage networks and a relationship computation controller. The sub-lineage networks capture diverse evolutionary relationships, while the relationship computation network models the affinity between the input sequence and each EvoExpert. The parameters are as follows: C = 64, D = 768. **c**, Resolution Restorer Module. The Resolution Restorer Module serves as the inverse process to the ConvBlocks, designed to reconstruct the feature vector back to nucleotide resolution. Its primary purpose is to transform the 320-channel feature vector (which has been expanded from the original 64 channels by the ConvBlocks) back to a representation that aligns with the original nucleotide-level resolution, effectively inverting the channel expansion and restoring the spatial detail for gene annotation. **d**, Detailed Architecture of Network Blocks. The ConvBlocks progressively increase the number of channels, with the convolutional layer channels expanding from C to 5 C. After passing through five ConvBlocks, the features reach 5 C = 320 channels. Each ConvBlock compresses information from two adjacent positions, embedding information from every 32 nucleotides into the same dimension. Encoder use the classical six-layer Transformer encoder structure to minimize parameter tuning. In ANNEVO, the number of attention heads is set to H = 8, and the hidden layer dimension is D = 768. EvoExpert employs two simple linear layers designed to preserve the distinct characteristics of different sub-clades. These layers map the feature vector's dimension from 320 to 768, and then back to 320. Relationship calculation network uses a single linear layer to map the feature vector from dimension 320 to 8, corresponding to the number of expert networks.

https://doi.org/10.1038/s41592-026-03036-7

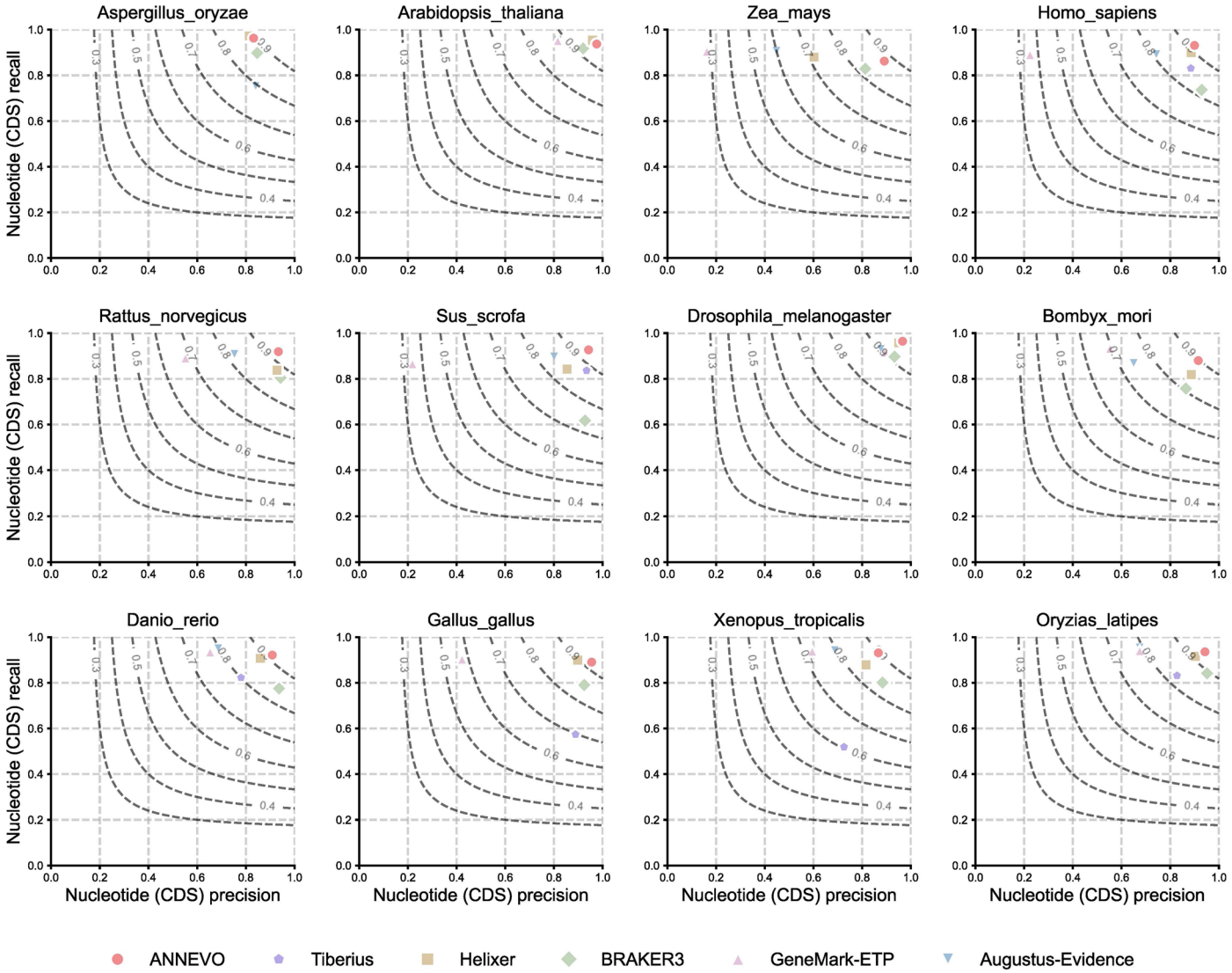


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41592-026-03036-7

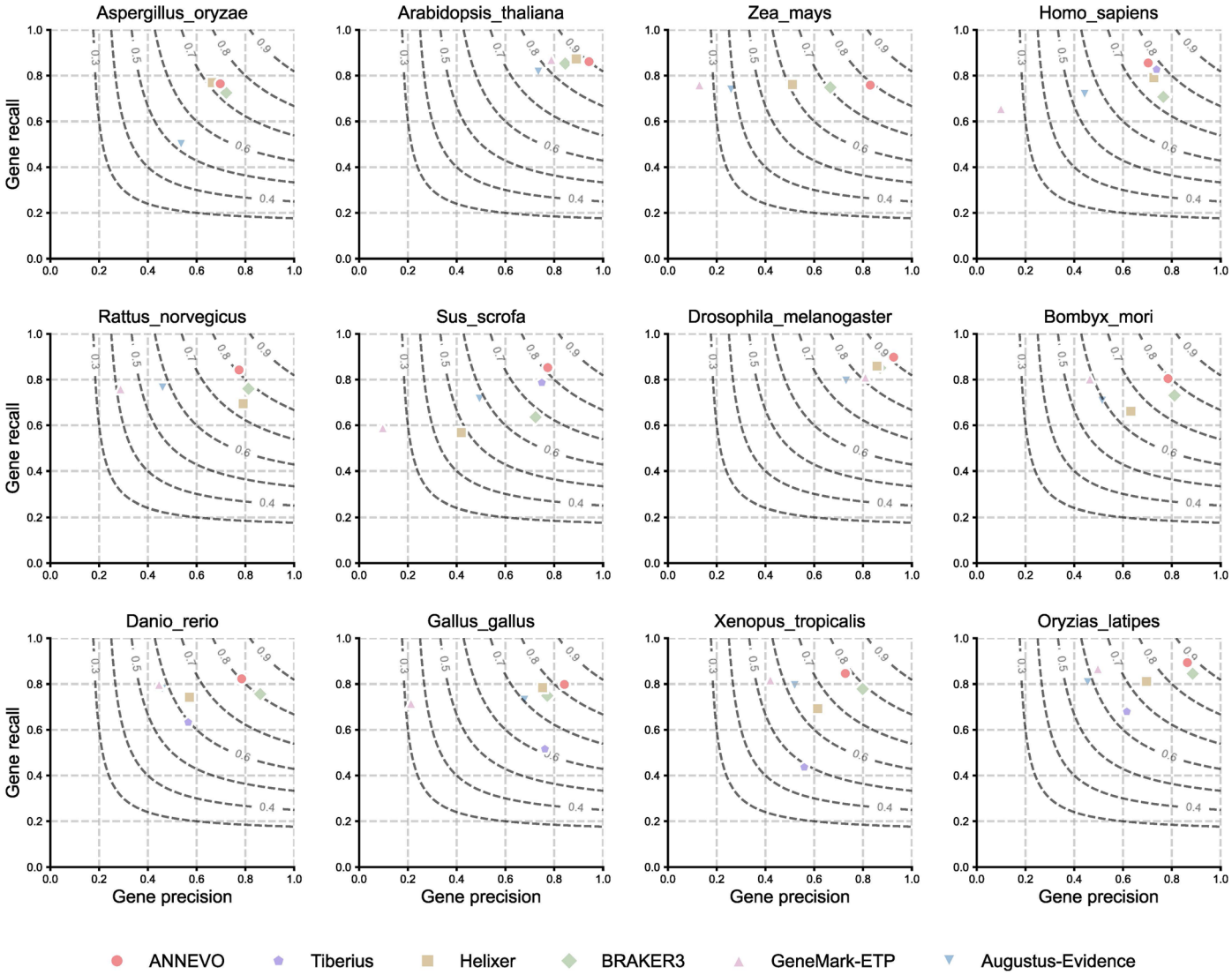
Extended Data Fig. 2 | Overview of the gene structure decoding component in ANNEVO. a, Predefined primary gene structure states. Gene structure states are defined based on the typical gene structure of eukaryotes. Each arrow in this diagram represents a possible state transition, with adjacent nucleotides specifying the required composition for the transition. Gene structure decoding utilizes the Viterbi algorithm, leveraging the prediction probabilities provided by the deep learning model to determine the most likely sequence of states. **b**, Intron state groups. The intron state account for three primary splicing patterns: GC-AG, GT-AG, and AT-AC. These splicing patterns are incorporated in the decoding process and considered during gene structure predictions. Importantly, an exit from an intron state to a CDS state does not return to the original CDS phase; instead, it transitions to the next CDS phase. For example, if the model enters an intron state from CDS0, it will exit to CDS1.

https://doi.org/10.1038/s41592-026-03036-7



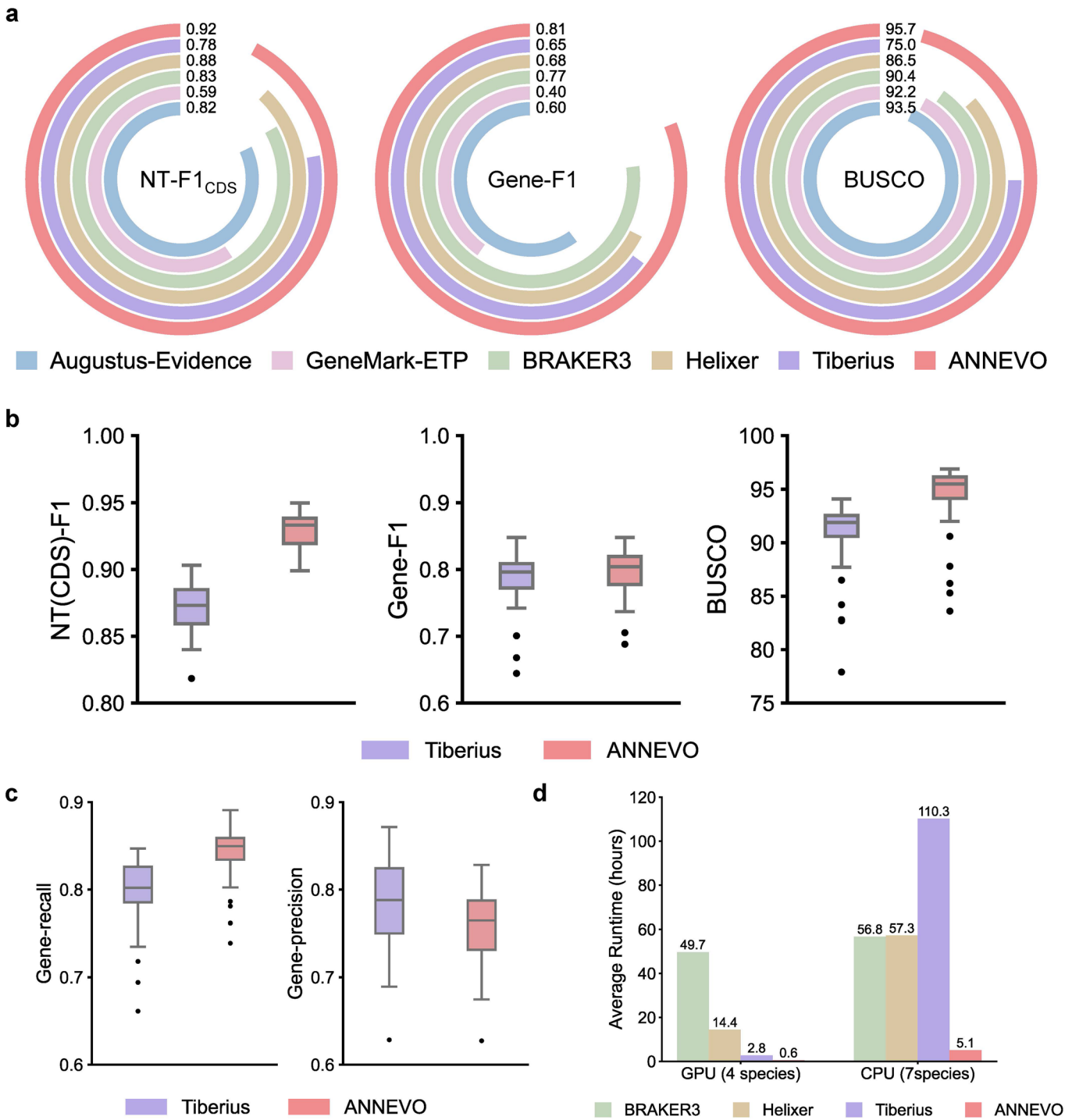
Extended Data Fig. 3 | Nucleotide-level performance comparison with evidence-assisted annotation pipelines across 12 model species. ANNEVO achieved the highest mean F1 score (0.92), driven by its superior recall (0.922), indicating more complete coding region identification. BRAKER3 exhibited slightly lower precision (0.906 vs. 0.919 for ANNEVO) and its completeness was substantially lower (recall: 0.805 vs. 0.922 for ANNEVO). The 7% absolute improvement in F1 by ANNEVO reflects its optimized balance between precision and recall, surpassing evidence-dependent methods despite its relying solely on genomic sequence inputs.

https://doi.org/10.1038/s41592-026-03036-7



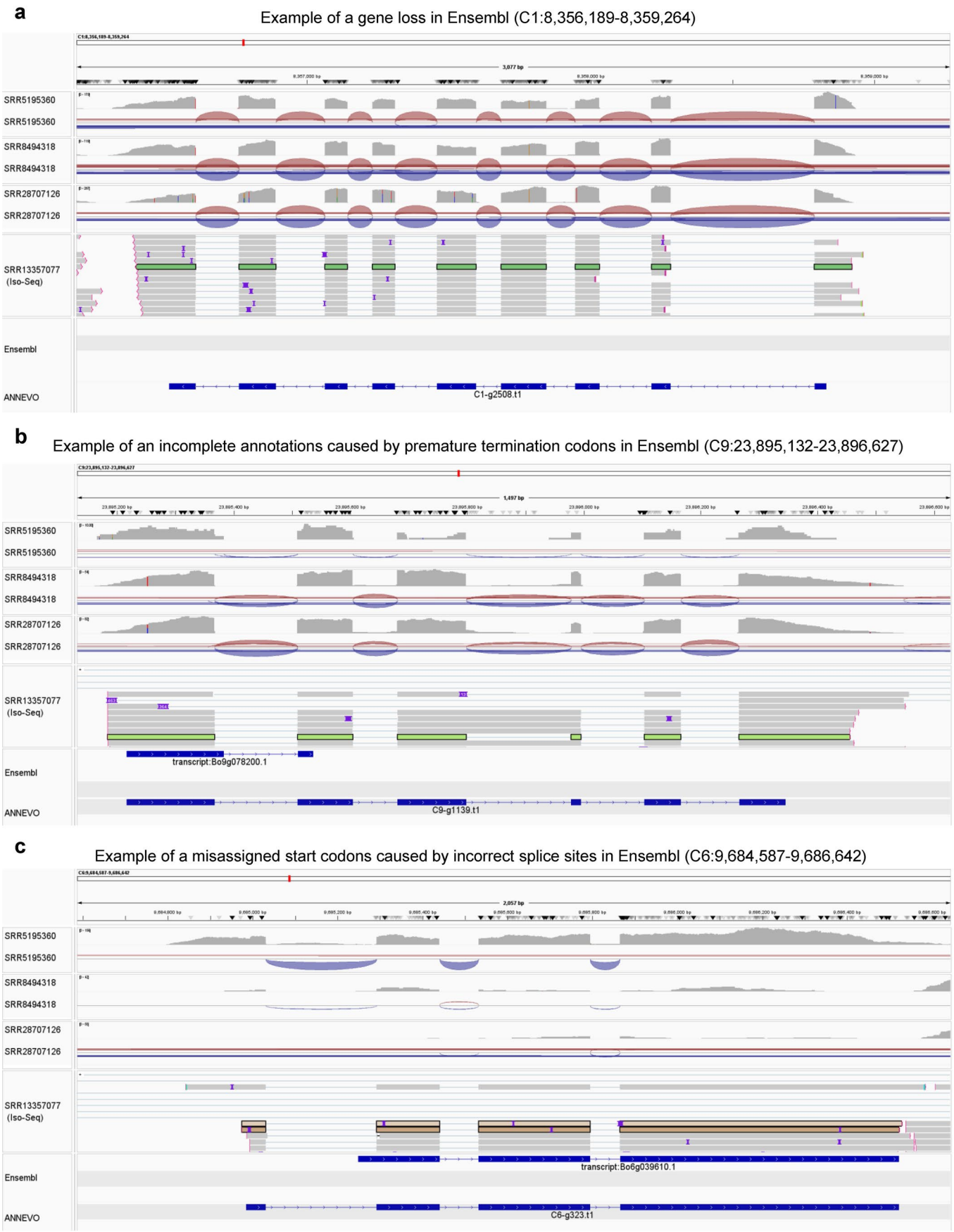
Extended Data Fig. 4 | Gene-level performance comparison with evidence-assisted annotation pipelines across 12 model species. ANNEVO achieved optimal performance in most species. ANNEVO demonstrates complete gene structure recovery (highest recall) that aligned with nucleotide-level performance. Across all 12 model species, ANNEVO achieves a 4% absolute improvement in F1 score over BRAKER3 and an 11% absolute improvement over the deep learning baseline Helixer.

https://doi.org/10.1038/s41592-026-03036-7



Extended Data Fig. 5 | Benchmarking against evidence-assisted annotation pipelines and deep learning methods. a, Average performance across six vertebrate model species. This panel presents a comparative analysis restricted to vertebrates, aligning with Tiberius's stated scope. ANNEVO consistently and substantially outperforms all other methods in this comparison set. Tiberius shows a notably lower average performance, primarily due to a sharp decline in its performance on Vertebrate_other clade (as detailed in Extended Data Figs. 3,4). **b**, Performance comparison between ANNEVO and Tiberius on all (43) test mammalian species. ANNEVO demonstrates superior performance across most test mammalian species, with higher NT(CDS)-F1, gene-F1 and BUSCO scores than Tiberius by an average of 5.9%, 1.0%, and 3.5%, respectively. The boxplot elements are defined as described in the Fig. 2b legend. **c**, Comparison of prediction tendencies between ANNEVO and Tiberius on all (43) test mammalian species. ANNEVO exhibits a tendency to recover more gene models, while Tiberius demonstrates a more conservative prediction behavior. The boxplot elements are defined as described in the Fig. 2b legend. **d**, Comparison of runtime across different deep learning-based gene annotation methods under GPU and CPU-only environments. BRAKER3 was used as the baseline for comparison. ANNEVO is substantially faster than all other methods in any settings. Note that due to the extreme resource demands of Tiberius, it could not be executed on a GPU with 32 GB of memory. Therefore, GPU-based evaluations were conducted on four vertebrate model species, while CPU-only evaluations were performed including mammalian model species.

https://doi.org/10.1038/s41592-026-03036-7

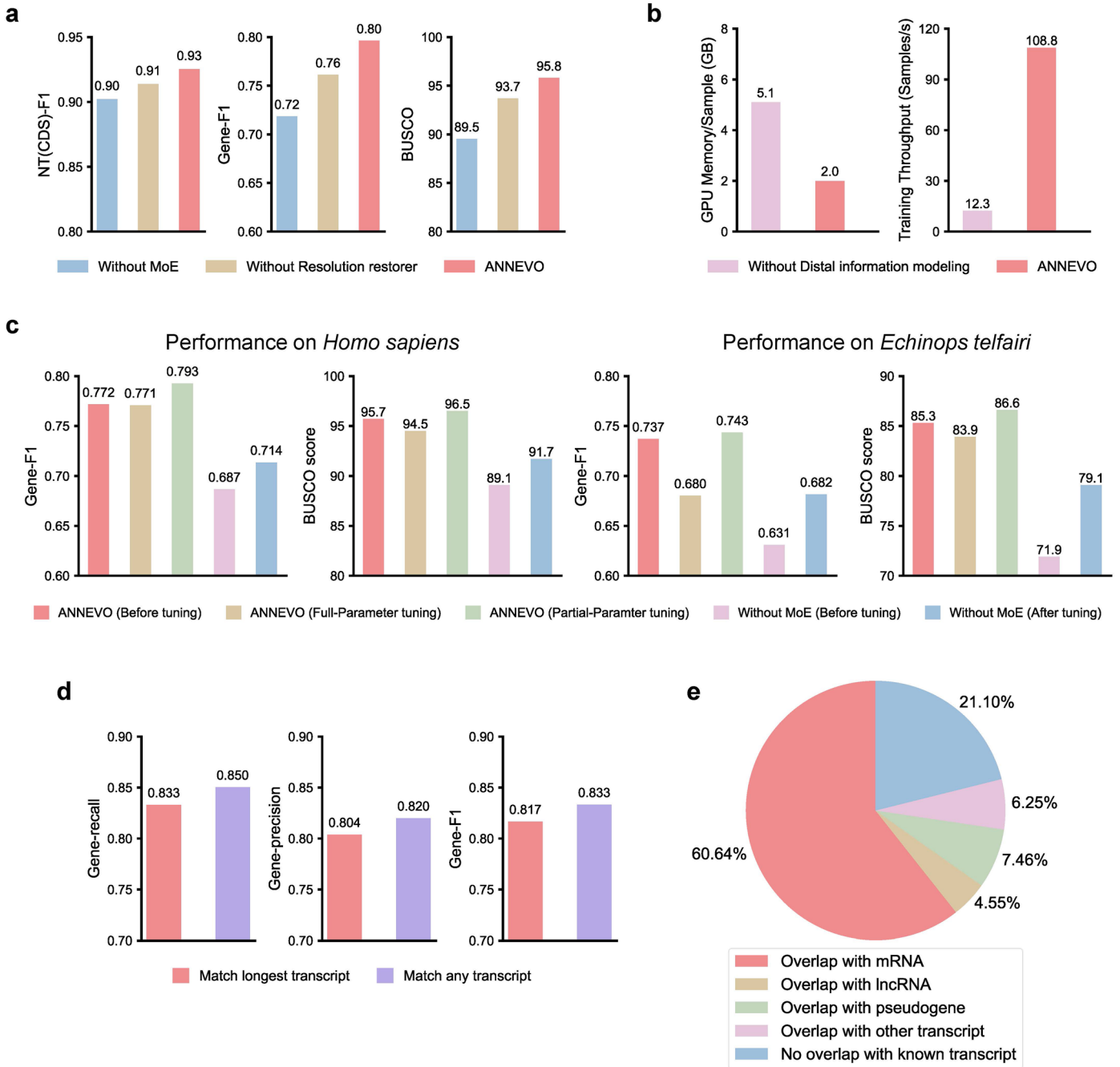


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41592-026-03036-7

Extended Data Fig. 6 | Examples of ANNEVO Correcting Erroneous Gene Annotations in Ensembl. a, Correction of a gene loss. Ensembl failed to annotate a conserved BUSCO gene at this region. ANNEVO recovered this gene, fully supported by RNA-seq evidence. **b**, Correction of a premature stop codon. Due to an erroneous splice site, the Ensembl annotation introduced a premature stop codon, resulting in the loss of four downstream exons. ANNEVO recovered this gene, fully supported by RNA-seq evidence. **c**, Correction of an incorrect start codon. Ensembl missed an upstream exon, leading to an incorrect initiation site. ANNEVO recovered this gene, fully supported by RNA-seq evidence.

https://doi.org/10.1038/s41592-026-03036-7



Extended Data Fig. 7 | Analysis of ANNEVO's performance. a, Ablation analysis of ANNEVO's performance. Performance decreases when either the MoE module or the resolution restorer module is removed from ANNEVO, with the MoE module contributing the most prominently to overall accuracy. **b**, Ablation analysis of ANNEVO's efficiency. Eliminating ANNEVO's distal information modeling module, even when substituting it with a pure CNN-based architecture, results in a 2.5-fold increase in training cost and an 8.8-fold decrease in training speed. **c**, Effects of different fine-tuning strategies on performance. We used the no-MoE version as the baseline. Partial fine tuning of specific modules led to consistent performance improvements across both model organisms and species with previously lower BUSCO. **d**, Performance comparison of ANNEVO when matching only the longest transcript versus any transcripts. The results indicate a modest performance improvement when matching against any transcript isoform, suggesting that some ANNEVO predictions correspond to alternative splice variants. Nevertheless, the overall predictions remain largely aligned with the longest transcript structures. **e**, Analysis of false-positive predictions by ANNEVO. The majority of false positives were identified as fragmented mRNAs or known non-coding transcripts, indicating that these predictions might still be biologically relevant rather than merely spurious.

nature portfolio

Corresponding author(s): Kai Ye

Last updated by author(s): Dec 10, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used wget (v1.14) to download the genome and annotation file. We used parallel-fastq-dump (v0.6) to download RNA-seq data.
Data analysis	Augustus (v3.5.0) is used as the baseline for performance comparison. ete3 (v3.1.3) is used to draw evolutionary trees. SeqKit (v2.8.2) is used to divide sequences. GNU Parallel (v20160622) is used to perform parallel predictions of Augustus. Fastp (v0.23.4) is used to quality control of RNA-seq reads. HISAT2 (v2.2.1) and SAMtools (v1.12) is used to align and sort RNA-seq reads. miniprot (v0.13) is used to align the protein sequence to genomes. gffread (v0.12.7) is used to extract protein sequences. BUSCO (v5.3.2) is used to evaluate annotation completeness. Another GFF Analysis Toolkit (v1.4.2) is used to convert the GTF format into a GFF file. GeneMark-ETP (latest, https://github.com/gatech-genemark/GeneMark-ETP) is the gene annotation pipeline compared in this study. BRAKER3 (v3.0.8) is the gene annotation pipeline compared in this study. Helixer (downloaded on 2025-03) is used for performance comparison of deep learning methods. Tiberius (downloaded on 2025-06) is a deep learning-based gene annotation method compared in this study. ANNEVO (Latest version) is available at GitHub (https://github.com/xjtu-omics/ANNEVO).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The sources of RNA-seq are listed in Supplementary Table 9. The genome and annotation of RefSeq are listed in Supplementary Table 23-27 and were downloaded from NCBI's FTP site (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/). The genome and annotation of Ensembl were downloaded from Ensembl release 112 (https://ftp.ensembl.org/pub/release-112/), Ensembl Fungi release 59 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/fungi/release-59/), and Ensembl Plants release 59 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-59/).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We downloaded all genome assemblies and annotations of Fungi, Embryophyta, Vertebrate_other Invertebrate and Mammalia available from RefSeq and Ensembl. The number of tested species reached 793, involving various evolutionary clades.
Data exclusions	Species with critical errors preventing data parsing were excluded. For instance, in RefSeq's Actinidia eriantha (version: GCF_019202715.1_MaoHua_MHT), sequence with ID NC_034914.1 has a defined range of 1–156,964, but includes a gene region spanning 156,727–157,788, which exceeds this range.
Replication	Replication was not relevant to our study. This study used deterministic algorithms without statistical analysis, and this study aims to demonstrate ANNEVO and its application in ab initio gene annotation.
Randomization	Randomization was not relevant to our study. ANNEVO is a deterministic method. All analysis in this study was done with existing data sources.
Blinding	Blinding was not relevant to our study. We used publicly available data, no data acquisition or statistical analysis was involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Not applicable
Novel plant genotypes	Not applicable
Authentication	Not applicable