nature genetics

Article

https://doi.org/10.1038/s41588-024-02053-6

Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation

Received: 28 August 2023

Accepted: 4 December 2024

Published online: 8 January 2025

Check for updates

Johannes Linder[1] ✉, Divyanshi Srivastava[1], Han Yuan[1], Vikram Agarwal[2] & David R. Kelley[1] ✉

Sequence-based machine-learning models trained on genomics data improve genetic variant interpretation by providing functional predictions describing their impact on the *cis*-regulatory code. However, current tools do not predict RNA-seq expression profiles because of modeling challenges. Here, we introduce Borzoi, a model that learns to predict cell-type-specific and tissue-specific RNA-seq coverage from DNA sequence. Using statistics derived from Borzoi's predicted coverage, we isolate and accurately score DNA variant effects across multiple layers of regulation, including transcription, splicing and polyadenylation. Evaluated on quantitative trait loci, Borzoi is competitive with and often outperforms state-of-the-art models trained on individual regulatory functions. By applying attribution methods to the derived statistics, we extract *cis*-regulatory motifs driving RNA expression and post-transcriptional regulation in normal tissues. The wide availability of RNA-seq data across species, conditions and assays profiling specific aspects of regulation emphasizes the potential of this approach to decipher the mapping from DNA sequence to regulatory function.

A long-standing goal in genetics is to accurately predict the effect of modifying each of the three billion nucleotides in the human genome with respect to gene-regulatory activity, ranging from chromatin accessibility and transcriptional activation to splicing and polyadenylation. Such predictions would dramatically improve researchers' ability to interpret pathogenic mutations and prioritize functional variants at loci implicated in genome-wide association studies (GWAS), or even improve GWAS itself through functionally informed discovery and fine mapping[1–3].

Machine-learning models trained to predict function from DNA sequences have been successful at characterizing regulatory syntax and interpreting genetic variant effects. Thus far, such models have focused on assays in which measured activity is proportional to local sequencing read counts. For example, transcription factor (TF) chromatin immunoprecipitation with sequencing (ChIP–seq) or DNase I hypersensitivity site sequencing (DNase-seq) and assay for transposase-accessible chromatin with sequencing (ATAC–seq) reads indicate a TF binding event or accessible DNA at the site where the reads align. This allows for accurate predictions using relatively short surrounding regions of sequence, typically 500–2,000 bp[4–10].

By contrast, the most popular sequencing assay, RNA sequencing (RNA-seq), does not have this property; RNA-seq reads aligned across a transcript will depend on a much larger region of sequence containing the gene's exons and relevant *cis*-regulatory elements. A read aligned to the 3′ end of a gene may be hundreds of thousands of nucleotides away from its promoter and enhancers that influence the magnitude of signal from the assay. Furthermore, RNA-seq coverage patterns integrate multiple layers of gene regulation; namely, transcription, splicing, termination or polyadenylation and RNA stability. These properties make the prediction of RNA-seq coverage from sequence challenging.

[1]Calico Life Sciences LLC, South San Francisco, CA, USA. [2]mRNA Center of Excellence, Sanofi Pasteur Inc., Cambridge, MA, USA.
✉e-mail: jlinder@calicolabs.com; drk@calicolabs.com

Previous models have only attempted to work with RNA-seq after summarizing gene expression in a single statistic. By processing a large region centered on the transcription start site (TSS), several models can predict normalized gene counts[11–13]. This approach depends on accurate TSS annotation and ignores isoform complexity. Other models predict cap analysis of gene expression (CAGE), which measures expression at the 5′ end of capped RNA (representing the TSS) and does not capture coverage at individual exons. Similarly, sequence-based models of post-transcriptional regulation rely on genome annotations and transformed measurements extracted from RNA-seq to isolate each regulatory mechanism (for example, percent spliced-in for splicing)[14–23]. However, such metrics inevitably struggle to describe complex splicing outcomes, unannotated de novo events or the intricate and sometimes competitive relationship between transcription, splicing and (intronic) polyadenylation[24–26].

Modeling RNA-seq coverage directly would have several benefits. First, RNA-seq is far richer than previously modeled assays. Although modeling multiple regulatory layers simultaneously is more challenging, it contains great promise; cross-talk between layers is common and their simultaneous consideration may improve models for each regulatory process. Thus far, models (for example, those trained on ChIP or ATAC) have mainly focused on one regulatory layer. Second, there are large amounts of RNA-seq data available, describing a wide variety of cell and tissue states across many species. Models trained on data from multiple species have been shown to improve performance[9], but chromatin profiling and the CAGE gene expression assays have been performed on far fewer species than RNA-seq.

Given that mammalian genes often span hundreds of thousands of nucleotides, effective RNA-seq modeling requires working with very large sequences and algorithms that propagate information across large distances. Recent work on the Enformer model using self-attention has demonstrated a path toward achieving this goal[13]. Therefore, we set out to model RNA-seq and additional epigenetic assays' coverage across diverse samples as a function of the underlying DNA sequence, without prior knowledge of gene annotation. We developed a model, named Borzoi, that effectively learns several layers of gene regulation. By applying attribution methods to predicted coverage patterns of individual RNA-seq experiments present in the training data, Borzoi derives the primary cell-type-specific or state-specific TF motifs and a genome-wide map of nucleotide influence on gene structure and expression. Our model improved performance relative to Enformer on downstream tasks to identify distal enhancers and predict genetic variant effects on gene expression, and it introduced new capabilities to predict variant effects on splicing and polyadenylation that match or exceed the state of the art. We anticipate that this toolkit will accelerate progress to determine mechanisms by which the many unsolved human genetic associations affect traits.

Results

RNA-seq model design

RNA-seq is a base-resolution readout of transcribed and usually processed RNAs. Thus, modeling RNA-seq coverage at base resolution would be ideal. However, the long span of mammalian genes means that we must also work with very long sequences to cover all exons and relevant regulatory elements. Computational limitations create a trade-off between these two considerations. We lean toward using longer sequences at the expense of some resolution, choosing 524 kb sequences for which we predict coverage in 32 bp bins. Training examples are extracted in tiled 524 kb windows spanning the human and mouse genome, thus containing genes at variable locations per window.

Our neural network model, called Borzoi, is illustrated in Fig. 1a. We use the core Enformer architecture, which includes a tower of convolution and subsampling blocks followed by a series of self-attention blocks operating at 128 bp resolution[27,28]. Self-attention is a critical operation, allowing every pair of positions to exchange information. From this point, we make use of a U-net architecture to increase the resolution back to 32 bp[29,30]. For each sequence length expansion (and resolution increase), we upsample the position vectors from the attention blocks and combine them with the corresponding feature map of equal size produced by the initial convolution tower (see Methods). To transition from embeddings representing 128 bp to those representing 32 bp, we perform this block twice, upsampling by a factor of two each time.

We chose to work with uniformly processed RNA-seq from ENCODE, providing 866 human and 279 mouse datasets measured across diverse biosamples, including cell lines, adult human tissues and developing mice[31,32]. We also included two to three replicates for each Genotype-Tissue Expression (GTEx) tissue processed by the recount3 project[33–35]. To help the model identify salient regulatory elements, we paired these data with the thousands of training datasets from the Enformer model, including CAGE, DNase-seq, ATAC–seq and ChIP–seq tracks (Methods). To assess model performance variance and enable ensembling, we trained four randomly initialized replicate models. We evaluated performance on a set of randomly held-out sequences from the human genome and orthologous mouse regions.

Borzoi accurately predicts RNA-seq and other assays

Despite the challenges involved with modeling RNA-seq coverage from only underlying DNA sequence, Borzoi predicts exon–intron coverage patterns with striking concordance for even long genes with many exons, as exemplified in Fig. 1b by the 190 kb gene *INSR*. Test set predictions matched RNA-seq coverage with a mean Pearson's *R* value of 0.74 across human samples when using one model replicate. Pearson's *R* increased to 0.75 when averaging the predictions across the full ensemble (Fig. 1c). Performance is difficult to compare directly to Enformer owing to differences in data processing (Methods). Nevertheless, test accuracies on overlapping datasets are broadly similar (Extended Data Fig. 1a–e) with two exceptions: the average Pearson's *R* is lower than Enformer for DNase and higher for CAGE.

To study predictions at the gene level, we aggregate and $\log_2$-normalize coverage in exon-overlapping bins. When comparing predicted to measured gene-level coverage values, we observe a mean Pearson's *R* of 0.87 across held-out genes (0.86 per model replicate) (Fig. 1d and Supplementary Fig. 1a–d). After quantile-normalizing the predictions across experiments and subtracting each gene's mean expression (so that the value represents the residual expression beyond the mean), we observe a mean Pearson's *R* of 0.58 (0.55 per replicate) (Fig. 1e), indicating that the model explains a significant amount of variation observed between tracks (such as tissue-specific and cell-type-specific differences). Finally, we note that Borzoi accurately predicts variation within the transcript structure; evaluated on the top 20% of test set genes with the highest variance in coverage across the span of exons and introns, the average Pearson's *R* value between predicted and measured RNA coverage (at the bin level) was 0.88 across all genes and samples (Supplementary Fig. 1e).

In the Supplementary Information, we show that the model relies on well-known regulatory features to make predictions and that the model's attention matrices comprehensively capture gene structure (Extended Data Fig. 2 and Supplementary Fig. 2).

Inference of tissue-specific expression and isoform usage

Gene expression is a multi-faceted process governed by numerous regulatory steps, including transcription initiation, splicing and polyadenylation, and these steps may exhibit tissue-specific effects. To study Borzoi's ability to make tissue-specific predictions, we focused on a set of five GTEx tissues: whole blood, liver, brain, muscle and esophagus. We first noted that Borzoi could accurately predict tissue-specific gene expression coverage on held-out test genes

https://doi.org/10.1038/s41588-024-02053-6

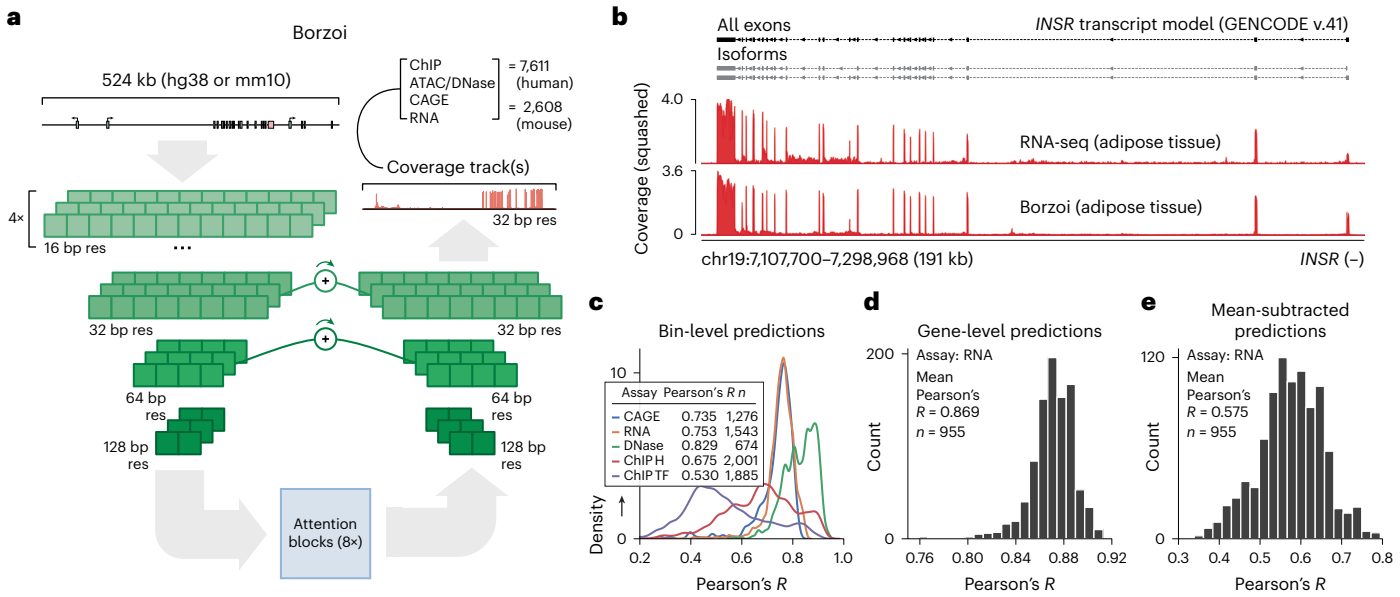


Fig. 1 | Borzoi: a neural network for predicting RNA-seq coverage from sequence. a, The Borzoi neural network architecture consists of a number of convolution and downsampling layers followed by a stack of self-attention layers with relative positional encodings operating at 128 bp resolution, similar to the Enformer architecture. The output is then repeatedly upsampled and put through additional convolution layers with matched U-net connections to predict at 32 bp resolution. Connections with '+' symbols represent a combination of the outputs of a previous layer with the inputs of a new layer through residual convolution. **b**, RNA-seq coverage prediction for the held-out test gene *INSR* (GTEx 'adipose tissue'), obtained by averaging the predictions of four model replicates. The 'squashed' scale refers to the transformed scale applied to the training data (Methods). **c**, Bin-level Pearson correlation on held-out test data across coverage tracks when predicting CAGE, RNA-seq, DNase-seq or ChIP–seq (n = number of coverage tracks). Predictions were averaged across four model replicates. **d**, Gene-level Pearson correlation when comparing the predicted to measured sum of RNA coverage across exons (n = number of sequencing experiments). **e**, Gene-level Pearson correlation after quantile-normalizing the RNA coverage tracks and subtracting the average gene expression across tracks (n = number of sequencing experiments).

(for example, the blood-specific gene *ADGRE1* visualized in Fig. 2a; see also Supplementary Fig. 3a,b). We compared the predicted and measured fold change in gene-level coverage of one tissue relative to the average coverage of the four other tissues, observing a Spearman's R range from 0.52 to 0.75 when using the ensemble of four model replicates (Fig. 2b).

Genes often have alternative TSSs, which are differentially used across tissues[36–38]. For example, *SGK1* harbors an upstream TSS that is highly expressed in brain but not blood (Fig. 2c; see Extended Data Fig. 3a for additional examples). We computed TSS usage ratios for the 5′-most and 3′-most TSSs from our ensembled predictions (Methods) and found correlations with experimental measurements (Spearman's R = 0.85; Supplementary Fig. 3c), FANTOM5 TSS usage proportions (Supplementary Fig. 3d) and tissue-specific TSS usage ratio fold changes (Spearman's R = 0.29–0.50 on held-out genes; Fig. 2d and Supplementary Fig. 3e).

The 3′ untranslated region (UTR) harbors regulatory regions called polyadenylation signals (PASs), which can generate multiple isoforms with distinct 3′ ends through alternative polyadenylation (APA)[39–41]. For example, *RWDD1* exhibits biased usage of the distal-most PAS in brain[42] (Fig. 2e; see Extended Data Fig. 3b for additional examples). Predicted tissue-pooled distal-to-proximal polyadenylation coverage ratios of held-out genes were highly correlated with measurements from GTEx (Spearman's R = 0.81; Supplementary Fig. 3f) and PolyADB v.3 (refs. 43,44) (Supplementary Fig. 3g). Predicted tissue-specific coverage ratio fold changes showed moderate correlation with measured fold changes between GTEx tissues (Spearman's R = 0.23–0.41; Fig. 2f and Supplementary Fig. 3h).

In the Supplementary Information, we show that although Borzoi competitively identifies splice junctions from matched negatives, the model has not learned to predict alternative splicing across tissues well (Extended Data Fig. 4; see Discussion).

Borzoi identifies regulatory motifs driving RNA expression

Borzoi enables direct characterization of tissue-specific *cis*-regulatory TF motifs by applying attribution methods to the predicted RNA-seq coverage statistics[45–50]. Focusing on the five GTEx tissues analyzed in the previous section, we selected 1,000 genes for each tissue with maximal transcript per million (TPM) fold change relative to other tissues and computed tissue-specific aggregated exon coverage gradients per gene. These saliency scores describe the contribution of each nucleotide to the predicted expression. As an example, gradients at the position of maximal liver-specific saliency for gene *CFHR2* highlight motif hits for CEBPA/B and HNF4A/G (Fig. 3a). We found that the gradient scores were broadly similar across replicates and closely matched in-silico saturation mutagenesis (ISM) (Supplementary Fig. 4).

Next, for each set of 1,000 tissue-specific genes, we selected the corresponding gradients and subtracted the average gradient of all other tissues, obtaining residual tissue-specific scores. We ran TF-MoDISco, a de novo motif clustering tool[51], for all five tissue gene sets and aligned motif clusters to their most likely database match using the Tomtom MEME suite and HOCOMOCO (v.11)[52,53]. A selection of top-scoring motifs are shown alongside their saliency distributions across genes in Fig. 3b (see also Supplementary Fig. 5a,b). We detect well-known regulators for each tissue, such as SPI1/B and IRF4/8 for blood, HNF4A/G and HNF1A for liver, SOX9 and REST for brain and MYOD1 and MEF2D for muscle. Motifs shared between tissues generally tend to regulate distinct loci (Fig. 3b, inset). We similarly recapitulate known regulatory motifs for esophagus and K562 (Supplementary Fig. 5c–e).

Finally, we aggregated the difference in gradient saliency for each pair of tissues among seqlets matching each TF, obtaining a scalar score that describes the importance of a particular TF in one tissue relative to another. These scores were highly correlated with observed TPM fold changes for the corresponding TFs (Fig. 3c and Supplementary

https://doi.org/10.1038/s41588-024-02053-6

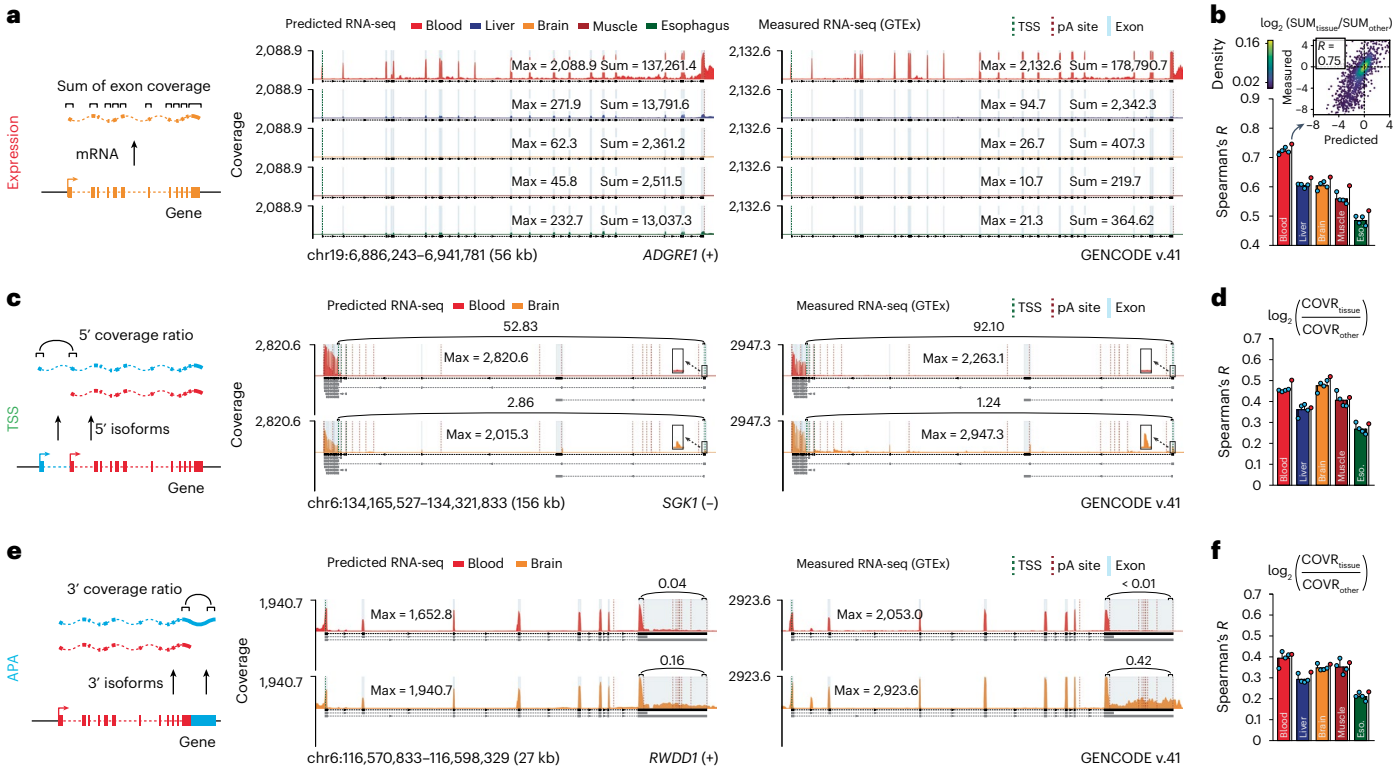


Fig. 2 | Predicting tissue-specific patterns of RNA-seq coverage in normal tissues. a, Example of tissue-specific gene expression predictions using Borzoi in five GTEx tissues for the blood-specific gene *ADGRE1*. The predicted and measured coverage of each RNA-seq experiment is aggregated in the bins that overlap exons (blue shaded regions; 'max' and 'sum' indicate maximum and total coverage). Exon annotations are shown below each coverage track (GENCODE v.41). **b**, Comparison of predicted and measured fold change between the aggregated coverage in a given tissue and the average coverage of the four other tissues for held-out test genes (n = 1,940). Blue and red dots represent replicate and ensemble model performance, respectively. Bar height represents average correlation. Inset, predictions for blood (color bar indicates Gaussian kernel density estimate). **c**, Example of alternative TSS isoform predictions for gene *SGK1*. TSS usage is estimated as a coverage ratio between bins overlapping each alternative start site (the ratio is annotated above each track). **d**, Comparison of predicted and measured TSS coverage ratio fold change, calculated between the coverage ratios (COVR) of a given tissue and the average coverage ratio of the remaining four tissues (n = 337 held-out genes with at least two TSSs). **e**, Example of 3′ UTR APA isoform predictions for gene *RWDD1*. Distal site usage is estimated as the coverage ratio of bins overlapping the distal-most and proximal-most polyadenylation sites. **f**, Comparison of predicted and measured fold change between APA coverage ratios of a given tissue and the remaining four tissues (n = 994 held-out genes with at least two sites).

Fig. 5f,g). For example, Spearman's R reached 0.77 when comparing TF saliency in blood and muscle. Note that a repressor element such as REST should be off-diagonal in comparison to brain, so we do not expect a perfect correlation.

Improved context use for gene expression prediction

We next assessed Borzoi's ability to identify and prioritize distal enhancer–gene interactions, which is critical to cell and tissue-specific regulation[54–57]. For each target gene, we computed input gradients of the aggregated exon coverage prediction in K562 RNA-seq samples, highlighting regulatory elements that drive the gene's expression prediction. Statistics derived from the gradient saliencies, averaged across the model ensemble, were compared to measurements from high-throughput CRISPR screens[58–62]. Compared to Enformer[13], Borzoi can score sites that are up to twice as far away from the gene, 262 kb, and we make use of exon annotations rather than TSS annotations, which are generally more robust to alternative isoforms. Fig. 4a,b displays the gradient attributions for genes *HBE1* and *MYC*, in which Borzoi correctly identifies both proximal (distance to TSS, <20,000 bp) and distal (distance to TSS, >200,000) enhancers, although false positives are also present.

When comparing Borzoi, Enformer and a distance-to-TSS baseline on their ability to classify measured positive from negative enhancer–gene interactions in data from previous works[60–65], we find that Borzoi has superior average precision (AUPRC) and area under the receiver operating characteristic curve (AUROC) at all distances (Fig. 4c and Extended Data Fig. 5a). Similar results are obtained on the data from Gasperini et al. (2019)[58] (Fig. 4d and Extended Data Fig. 5b). In line with recent work[66], we find a general decreasing trend in average predicted percent expression change with TSS distance for both positive and negative examples (Supplementary Fig. 6a). We study coverage patterns across the transcript in more detail in Supplementary Fig. 6b–e. Through ablation experiments, we find that including training data such as DNase-seq and ATAC–seq in addition to RNA-seq improves performance (Supplementary Fig. 7a–c).

To further demonstrate the model's reliance on a broader genomic context for its predictions, we analyzed expression data of seven distinct promoters that had been integrated into thousands of genomic positions by the TRIP assay[67,68]. We predicted activity scores from multiple classes of coverage tracks, including DNase, histone modifications, CAGE and RNA-seq (Supplementary Fig. 8a,b and Methods). In general, the scores derived from DNase tracks were most concordant with the measured expression levels (Fig. 4e and Supplementary Fig. 8c; 20-fold cross-validation, Spearman's R = 0.58 for promoter *ARHGEF9*). These predictions were better correlated with expression than LMNB1 DamID-seq, which measures nuclear lamina interactions and constitutes a strong baseline.

Borzoi prioritizes genetic variants that influence expression

Accurately predicting the influence of genetic variants on gene expression is crucial for understanding the regulatory mechanisms of genetic

https://doi.org/10.1038/s41588-024-02053-6

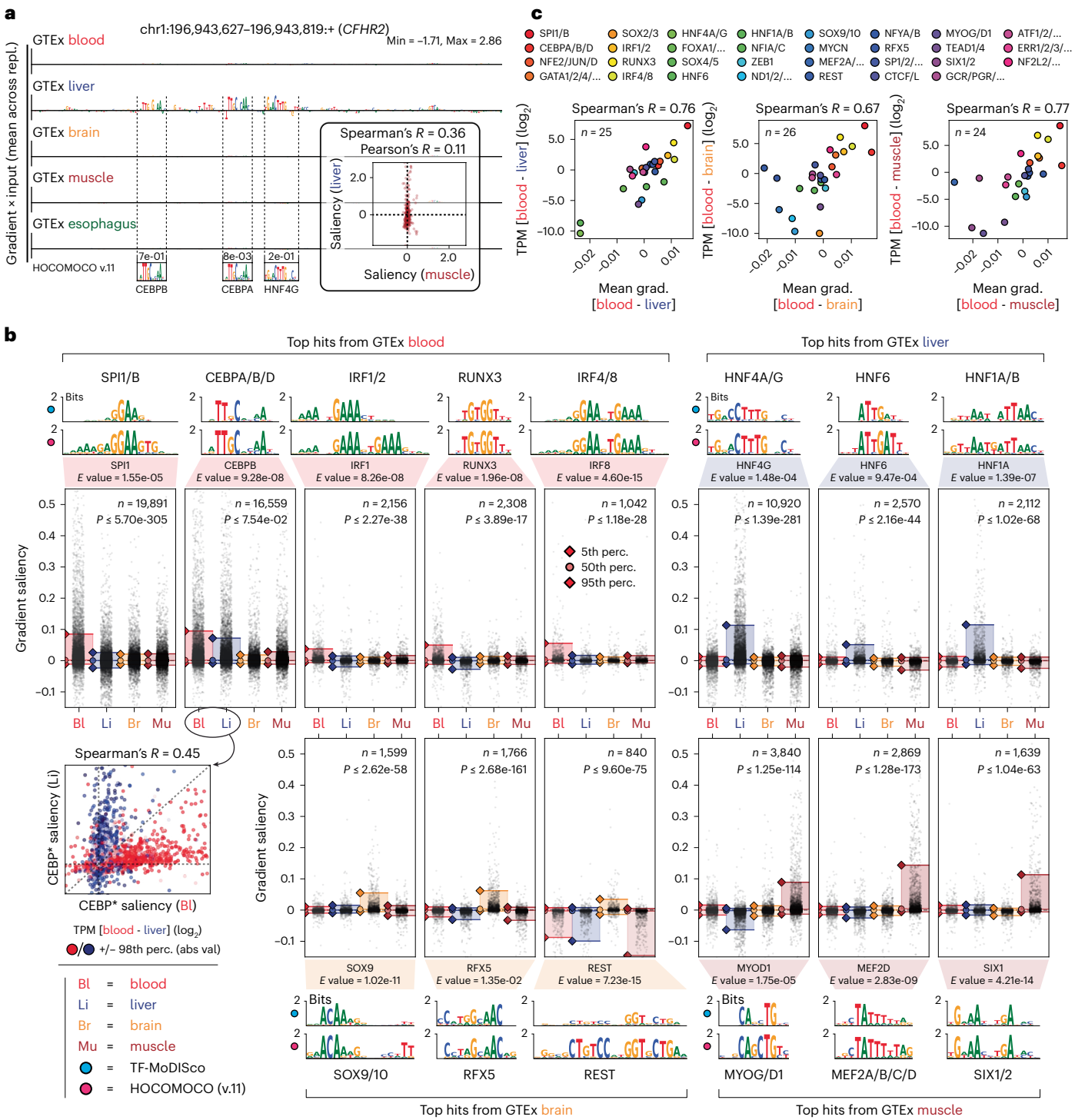


Fig. 3 | Identifying transcriptional *cis*-regulatory motifs through tissue-specific attribution. a, Gradient attributions at the mode of maximum saliency for five GTEx tissues for liver-specific gene *CFHR2* (mean ensemble saliency). All sequence logos have identically scaled *y* axes (min and max are displayed in the top right corner). Probable motif hits and their position weight matrices (PWMs) from HOCOMOCO (v.11) are shown. Annotated Tomtom *E* values represent the significance of the motif match. Inset, comparison of nucleotide-level saliencies for liver and muscle coverage tracks. **b**, A selection of motif clusters identified by MoDISco from gradient saliencies corresponding to four GTEx tissues. Shown are the MoDISco PWMs, the best-matching PWMs from HOCOMOCO and the distributions of tissue-specific gradient saliencies for seqlets belonging to a given cluster (n = number of seqlets). *P* values are computed using a two-sided Wilcoxon test between the gradient saliencies of the tissue with the largest and second largest 95th percentile of values. *P* values ranged from 0.075 for CEBPA/B/D (not significant) to 5.7×10^{-305} for SPI1/B. The *E* values represent the significance of motif matches as computed by Tomtom. Bottom left, comparison of seqlet saliencies for putative CEBPA/B/D between whole blood and liver. Each dot is colored by the measured difference in log(TPM) for the target gene. **c**, Comparison between the average difference in gradient saliency of seqlets belonging to motif clusters for pairs of GTEx tissues and the difference in measured log(TPM) for the corresponding TF genes. The median TPMs of genes belonging to the same TF subfamily (HOCOMOCO) were averaged.

https://doi.org/10.1038/s41588-024-02053-6

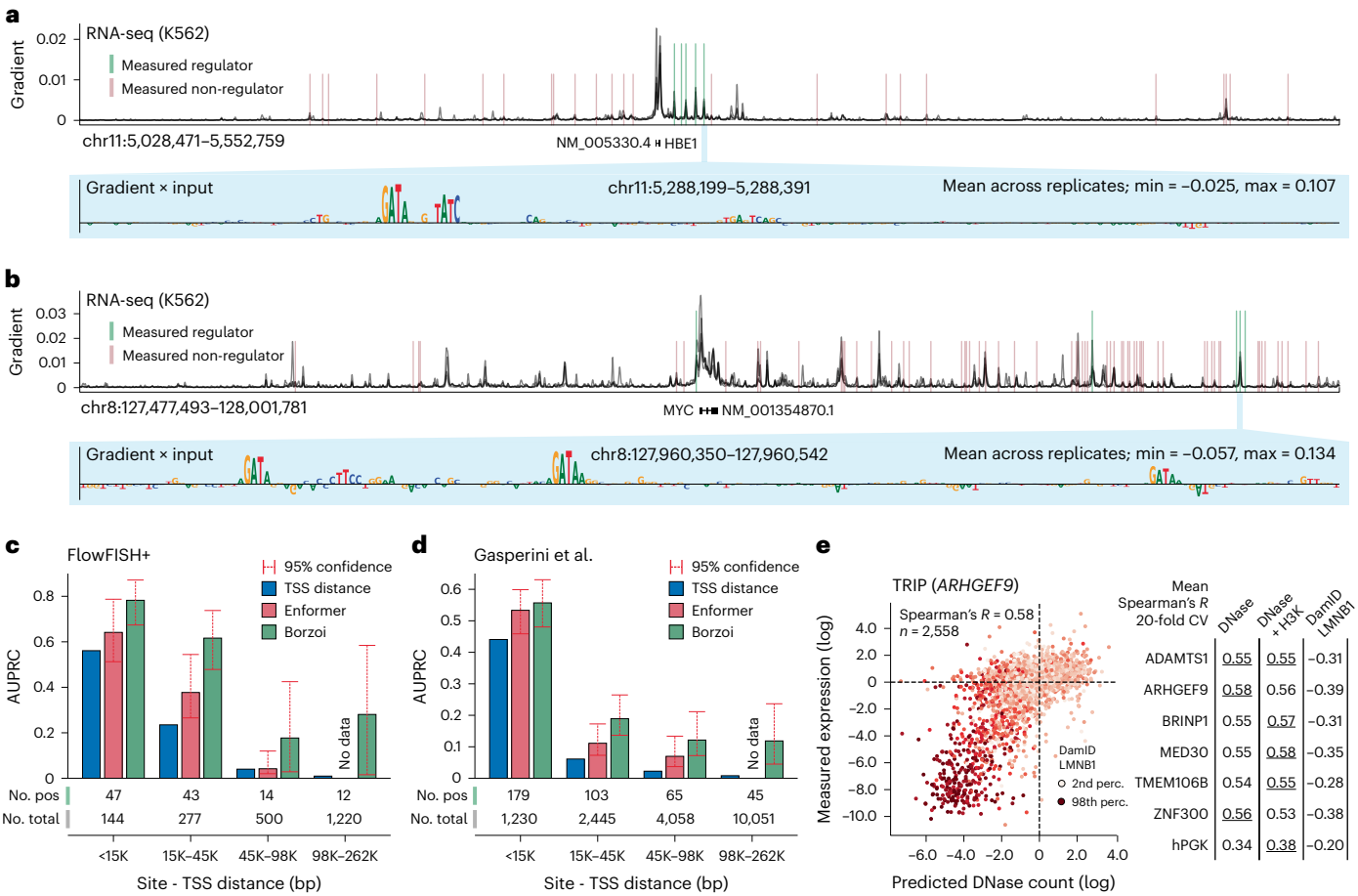


Fig. 4 | Predicting the impact of context and distal regulatory elements on gene expression. a, Exon-aggregated gradient saliency for *HBE1* across the 524 kb input (curves for four model replicates). CRE regions that are measured to regulate (green) or not regulate (red) *HBE1* are annotated. Input-gated gradients for a 192 bp window centered on the most distal enhancer are shown at the bottom (min and max are displayed in the top right corner). **b**, Exon-aggregated gradients for *MYC*. **c**, Average precision (AUPRC) when using a statistic computed from the Borzoi or Enformer gradients within a local window around each CRE locus to classify whether it regulates the target gene (measurements from a previous publication[65]). The number of positives and total number of examples are displayed below each distance bin. The total number of examples is (<15K) $n = 144$, (15K–45K) $n = 277$, (45K–98K) $n = 500$ and (98K–262K) $n = 1,220$. 95% confidence intervals were estimated from 1,000-fold bootstrapping. **d**, AUPRCs when using Borzoi or Enformer gradients to classify regulating and non-regulating CREs in data from Gasperini et al. (2019)[58]. The total number of examples is (<15K) $n = 1,230$, (15K–45K) $n = 2,445$, (45K–98K) $n = 4,058$ and (98K–262K) $n = 10,051$. 95% confidence intervals were estimated from 1,000-fold bootstrapping. **e**, Left, predicted vs measured expression levels of TRIP reporter constructs based on Borzoi DNase coverage in K562 (promoter, *ARHGEF9*). Color corresponds to DamID LMNB1 measurements. Right, average Spearman's R (20-fold cross-validation (CV)) when predicting TRIP expression based on different scores.

associations in human populations. Here, we evaluated Borzoi's ability to distinguish fine-mapped GTEx expression quantitative trait loci (eQTLs) from a set of matched negatives, controlling for TSS distance[1]. As an example, Fig. 5a shows RNA-seq coverage predictions for the gene *SHTN1* in GTEx whole blood, for both the reference sequence and an altered sequence substituting the alternative allele of single nucleotide polymorphism (SNP) rs1905542. We also show the measured coverage in GTEx individuals harboring each allele. Borzoi correctly predicts the upregulation of SHTN1 expression owing to the creation of a CEBP binding motif[69–72] (see Supplementary Fig. 9a and Extended Data Fig. 6a,b for additional examples).

Borzoi predicts coverage across a large sequence region from which a variant effect score must be distilled. For RNA-seq tracks, we compute either the log fold-change sum or L2 norm of differential coverage across exons (Methods). Using Borzoi's ensemble with an L2 score was superior to Enformer and its original sum aggregation at discriminating eQTLs (mean AUROC = 0.794 vs 0.747 across tissues; Fig. 5b,c). Borzoi still outperformed Enformer when using a single model (AUROC = 0.788) or when switching to the original sum statistic (AUROC = 0.772). Borzoi also exhibits greater Spearman correlation than Enformer when comparing effect size predictions to fine-mapped eQTL coefficients (mean $R = 0.334$ across tissues vs $R = 0.227$; Fig. 5d and Supplementary Fig. 9b,c). Borzoi outperforms Enformer with even a single model (mean $R = 0.292$). In ablation experiments, we found that training on DNase-seq and ATAC–seq data in addition to RNA-seq, as well as mouse data, substantially improved predictions (Supplementary Fig. 9d). We further evaluated the model's ability to prioritize true eGenes among other genes surrounding an eQTL (Supplementary Fig. 9e). The model performed, at best, marginally better than a TSS distance baseline.

To further test the utility of Borzoi-derived variant scores, we investigated the degree to which the model can distinguish common variation, which is generally benign, from a matched set of singletons (rare variants observed in a single individual), which are relatively enriched for pathogenicity, in the GnomAD database[73,74]. For comparison, we considered CADD (v.1.6) scores[75,76]. Restricted to ENCODE candidate *cis*-regulatory elements, Borzoi and CADD exhibited equal discriminative power (mean AUROC = 0.55; Fig. 5e and Supplementary Fig. 9f). Combining their scores resulted in the highest accuracy (mean AUROC = 0.57).

https://doi.org/10.1038/s41588-024-02053-6

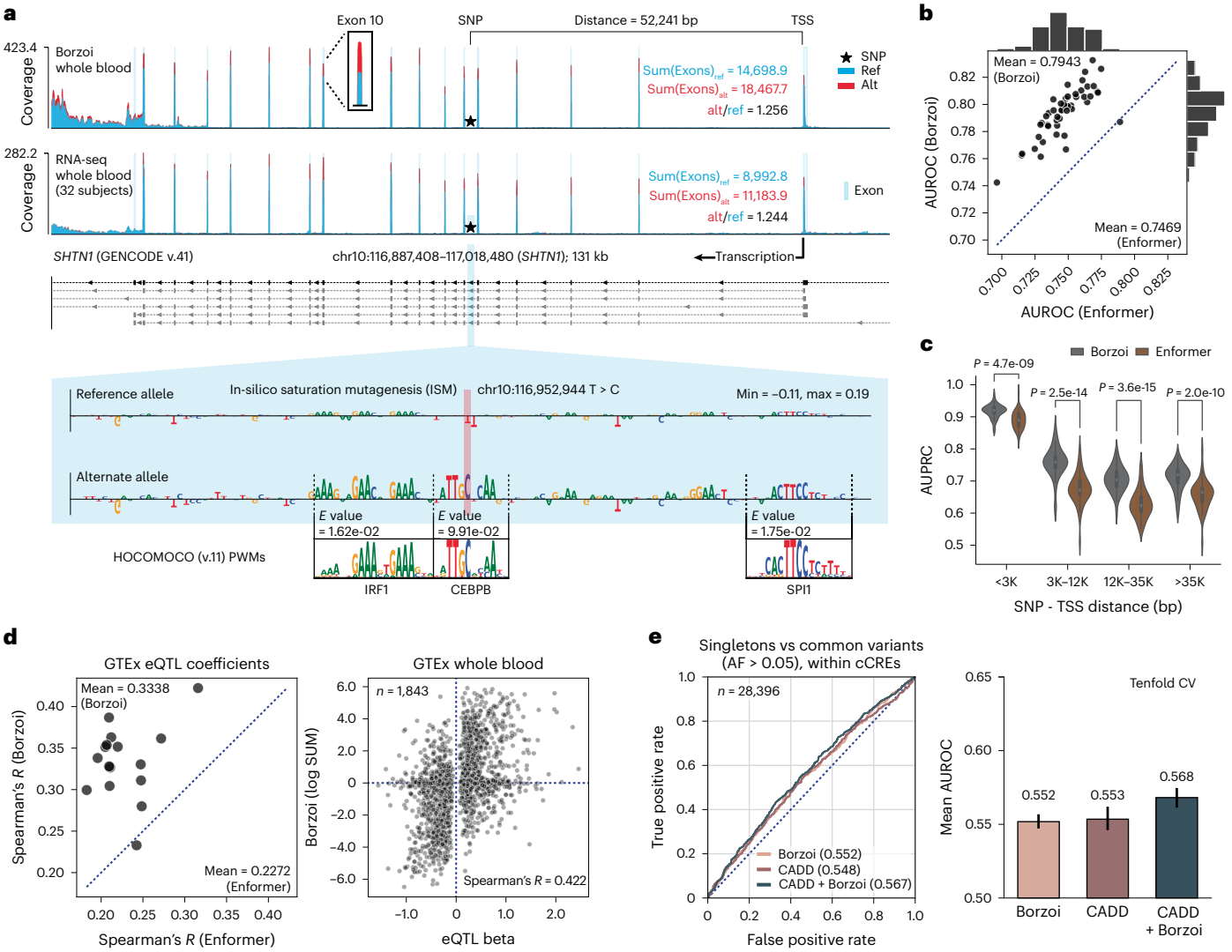


Fig. 5 | Borzoi predictions of variant effects align with eQTL results and negative selection. a, Example eQTL rs1905542. Shown are the predicted RNA-seq whole blood coverage tracks for the reference (blue) and alternate (red) alleles, as well as the measured, aggregated RNA-seq coverage in whole blood for 32 homozygous carriers of the reference allele and 32 heterozygous or homozygous carriers of the alternate allele. Exon-overlapping bins are shaded light blue. Exon-aggregated coverage for each allele and their ratio are annotated. ISM maps are shown at the bottom with equally scaled *y* axes, along with probable motif hits and Tomtom motif *E* values. **b**, AUROC per GTEx tissue when using Borzoi or Enformer to classify fine-mapped eQTLs from distance-matched negatives. Each model's mean AUROC is annotated. **c**, Comparison of tissue-specific GTEx eQTL classification performance as a function of distance to the TSS. Each violin plot shows the median AUPRC, interquartile range and 1.5× interquartile range as whiskers. *P* values are computed using a two-sided Wilcoxon test ($n = 49$ tissues). **d**, Left, comparison of Spearman's *R* between predicted and observed GTEx eQTL effect sizes, using either Borzoi or Enformer with the differential log sum coverage statistic ('SUM'; Methods). Each model's mean Spearman's *R* value is annotated. Right, predicted vs observed eQTL effect sizes in whole blood for Borzoi. **e**, Left, ROC obtained when classifying singleton variants from common variation (AF > 0.05) from gnomAD. Right, Mean AUROC with error bars indicating the 95% confidence interval, estimated from 1,000-fold bootstrapping (tenfold cross-validation). All variants were sampled from ENCODE candidate *cis*-regulatory elements (cCREs). AUROC scores are displayed in the legend.

In the Supplementary Information, we show that Borzoi exhibits competitive performance compared to Enformer when predicting non-coding regulatory mutations in promoters and enhancers as measured by massively parallel reporter assays (MPRAs) (Supplementary Fig. 10).

Functional polyadenylation variant interpretation

Another important class of disease variants alters 3′ mRNA processing[77]. We first probed Borzoi's predicted coverage in 3′ UTRs with attribution methods to understand which sequence features affect the predicted shape (Fig. 6a). Motifs for well-known polyadenylation regulators (for example, CFIm, CPSF, CstF) emerge from the attribution scores of the predicted distal polyadenylation ratio (Fig. 6b). Although we generally do not find determinants of mRNA half-life in the 3′ UTR attributions, we do observe a correlation between codon-aggregated gradient saliencies of gene exon coverage and MPRA measurements from a previous publication[78] (Pearson' $R = 0.59$) (Supplementary Fig. 11a). We also note that window-shuffled ISM is a more reliable attribution method in 3′ UTRs because of buffering effects (Supplementary Fig. 11b).

We next investigated Borzoi's ability to distinguish between fine-mapped 3′ QTLs from the eQTL catalog[79,80] (polyadenylation QTLs (paQTLs); $n = 1,058$) and a set of expression-matched negatives, controlling for PAS distance. We calculated variant effect scores as the maximal absolute change in predicted coverage ratio between any 3′ cleavage junction from tissue-pooled GTEx tracks. We focused on tissue-pooled predictions because the limited number of QTLs

https://doi.org/10.1038/s41588-024-02053-6

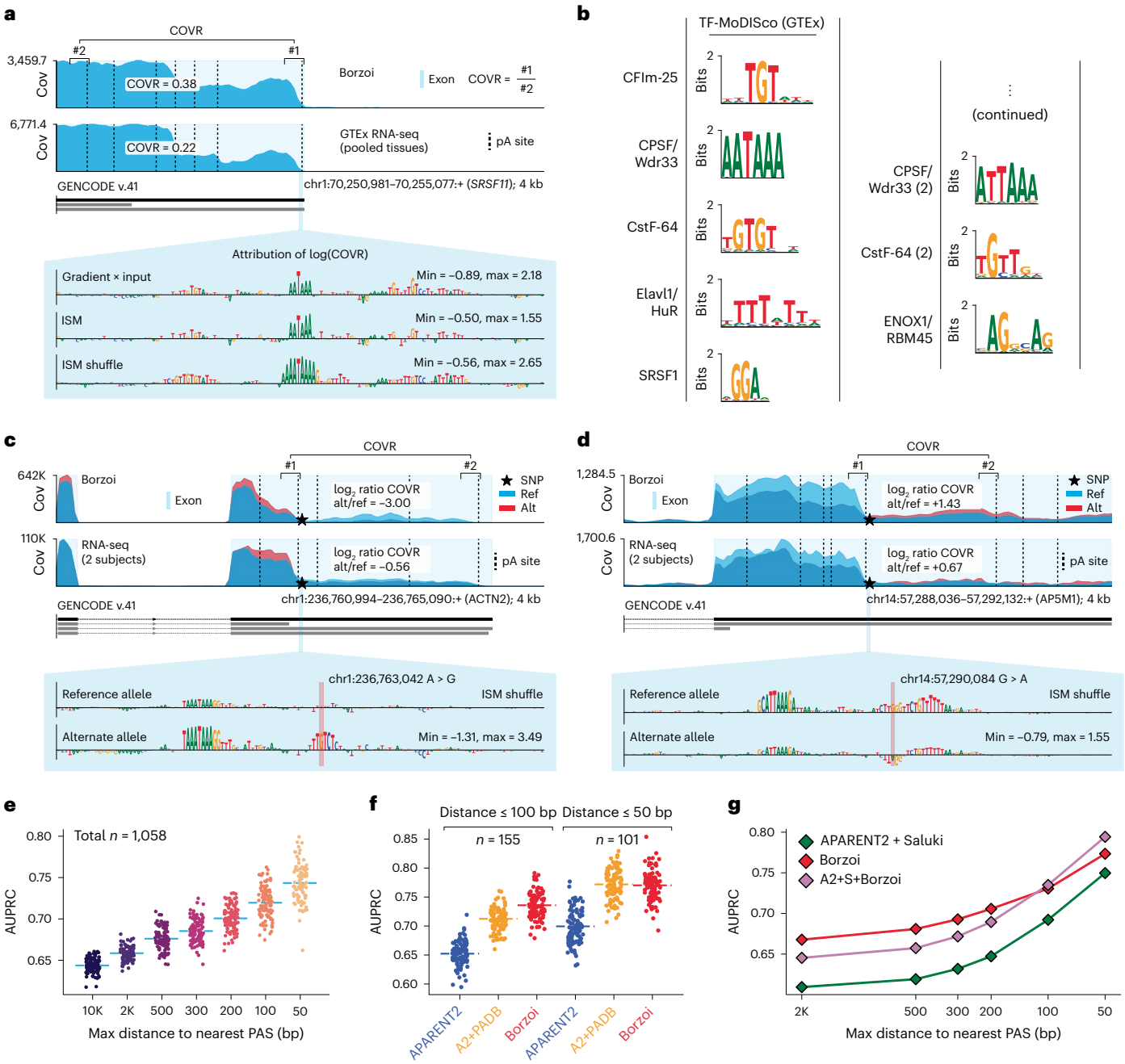


Fig. 6 | Predicting APA and 3′ polyadenylation QTLs. a, Predicted and measured RNA-seq coverage across the distal PAS of *SRSF11* (GTEx pooled tissue). Calculation of polyadenylation-centric coverage ratios (COVR) is illustrated in the figure. Attribution scores based on gradient saliency, ISM and ISM shuffle are shown at the bottom (min and max displayed in the right corner). **b**, MoDISco PWMs of well-known APA regulators, obtained from pooled GTEx coverage ratio gradients calculated for the Gasperini gene set[58]. **c**, Predicted RNA-seq coverage (GTEx pooled) for variant rs114880747, along with measured coverage in two individuals with the reference allele and two heterozygous individuals (three tissues). The log ratio between the variant and reference COVR statistics is annotated in the plot. Attribution scores (bottom; plotted with equal *y* scale) suggest gain of a CstF motif. **d**, Predicted and measured coverage in individuals without and with variant rs80168986 (two individuals, three tissues each). Attribution scores (bottom) suggest gain of an HNRNPA1 motif. **e**, AUPRC when classifying fine-mapped GTEx paQTLs based on predicted RNA-seq coverage ratio statistics (tissue pooled), plotted as a function of decreasing distance threshold to the nearest 3′ UTR PAS. Each dot represents a permutation test ($n = 100$; dashed line, mean; Methods). **f**, paQTL classification AUPRC comparing variant predictions of Borzoi, APARENT2 and APARENT2+PolyADB. Each dot represents a permutation test ($n = 100$; dashed line, mean). **g**, Mean paQTL classification AUPRC of 100 permutations, plotted as a function of decreasing distance threshold to the nearest PAS. 'A2+S+Borzoi' represents an ensemble of all models.

prohibited a tissue-specific analysis. Coverage predictions for two paQTLs are shown in Fig. 6c–d (and Supplementary Fig. 11c,d). Compared to RNA-seq tracks of GTEx individuals harboring the alternative allele, Borzoi correctly predicts the change in site usage caused by each variant. Extended Data Fig. 7a shows more examples.

The variant effect scores derived from the predicted RNA-seq tracks discriminated paQTLs from the matched negatives with a monotonic increase in accuracy at closer distances to the nearest PAS (Fig. 6e; AUPRC = 0.64–0.74). Compared to variant scores predicted by the APARENT2 model[22], Borzoi was consistently more accurate (Fig. 6f).

https://doi.org/10.1038/s41588-024-02053-6

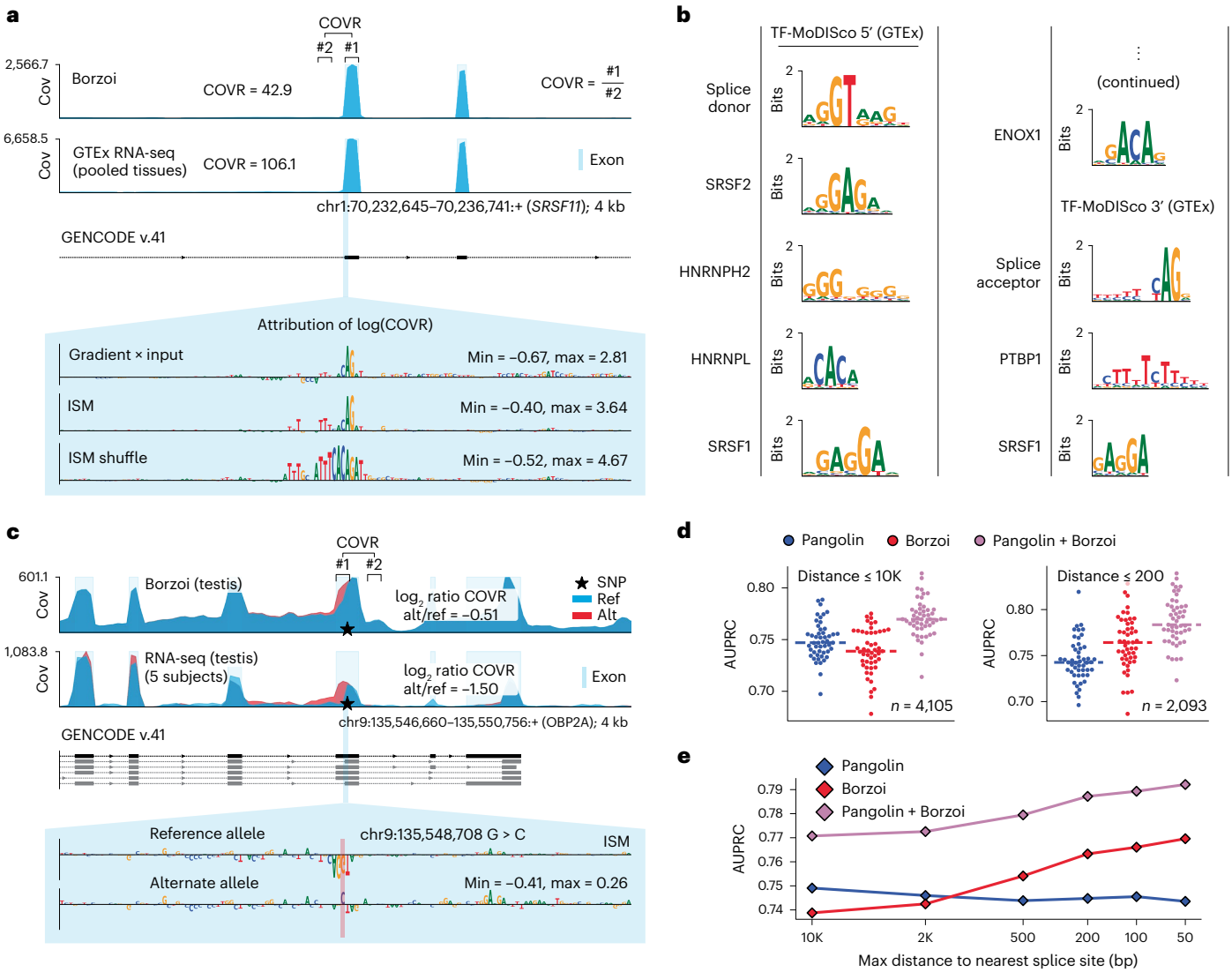


Fig. 7 | Classifying sQTLs and intronic paQTLs from RNA-seq coverage predictions. a, Predicted and measured RNA-seq coverage across an exon in the *SRSF11* gene (GTEx pooled tissue). Calculation of exon-to-intron coverage statistics (COVR) is illustrated in the figure. Attribution scores based on gradient saliency, ISM and ISM shuffle are shown below (min and max displayed in the right corner). **b**, PWMs of putative splicing regulators, obtained by running MoDISco on pooled GTEx coverage ratio gradients. **c**, Predicted RNA-seq coverage (GTEx tissue testis) for variant rs55695858, along with measured coverage in testis for five individuals with the reference allele and five heterozygous individuals (the sQTL is significant in testis). The log ratio between the variant and reference COVR statistics is annotated in the plot. Attribution scores are shown below (*y* axes plotted with equal scale). **d**, Comparison between the variant effect predictions of Borzoi, Pangolin and an ensemble of both models at the task of classifying fine-mapped splicing QTLs from GTEx, at different distance thresholds from an annotated splice junction. Each dot represents the AUPRC metric of each model for a given GTEx tissue (median AUPRC drawn as a dashed line). **e**, Average AUPRC for Pangolin, Borzoi and their ensemble as a function of decreasing distance threshold to the nearest splice junction.

However, the performance gap decreased when scaling APARENT2's predictions by the reference isoform percent from PolyADB, suggesting that context is an important determinant. We further compared to a 3′ UTR-wide ensemble of APARENT2 and Saluki[23] (Methods). Borzoi performs better at longer distances (dAUPRC > 0.050 at 2,000 bp) with a more comparable performance closer to the PAS (dAUPRC = 0.025 at 50 bp) (Fig. 6g). At closer distances, the average rank of all model predictions (Borzoi, APARENT2 and Saluki) surpasses either model's individual performance.

Functional splicing variant interpretation

Repeating the analyses of the previous section for RNA splicing, we defined a splice-centric attribution score based on the predicted exon-to-intron coverage ratio spanning a splice junction (Fig. 7a). When running MoDISco on gradients from tissue-pooled exon-to-intron coverage ratios for genes from the Gasperini set[58], we found known splice-regulatory motifs (Fig. 7b). Buffering effects were less problematic when interpreting repeat-like splicing motifs with ISM (Supplementary Fig. 12a).

We curated fine-mapped splicing QTLs (sQTLs) from the eQTL catalog and constructed expression-matched and splice distance-matched negatives ($n = 4{,}105$)[80]. This relatively large set of variants allowed for a tissue-specific analysis. Variant effect scores were calculated from the predictions as the maximum absolute difference in relative coverage across bins within the gene span. RNA-seq coverage predictions for an example sQTL (rs55695858) are shown in Fig. 7c (see Supplementary Fig. 12b and Extended Data Fig. 8a,b for more examples), along with measured coverage for five GTEx individuals with or without the alternative allele. The variant weakens an alternative 3′ splice site, which upregulates extension of the corresponding exon. When comparing

Borzoi to Pangolin[16] for the task of classifying the causal sQTLs from matched negatives, Pangolin has a slight advantage (Fig. 7d–e and Supplementary Fig. 12c; dAUPRC = 0.01, evaluated on all SNPs within distances of ≤10,000 bp from an annotated splice site). Most far-away SNPs are de novo splice–gain mutations and are relatively easy for Pangolin to classify based on the local predicted effect at the variant allele, whereas Borzoi's splice–gain predictions appear less well-calibrated. By contrast, Borzoi is better at distances closer to the junction (Fig. 7d–e; dAUPRC = 0.02, evaluated on variants ≤200 bp from an annotated junction). Importantly, the average rank prediction of both models is superior to either model alone (dAUPRC > 0.02).

Intronic polyadenylation variant interpretation

Candidate polyadenylation sites frequently occur in introns, resulting in competition between the PAS and the enveloping splice junctions. In this case, the intron is either spliced out or retained and polyadenylated[40,81]. Curious as to whether Borzoi has learned about this competition between distinct regulatory functions, we filtered the paQTLs from the eQTL catalog for SNPs that were closer to intronic polyadenylation sites than 3′ UTR sites and constructed new expression-controlled negatives that were matched for intronic polyadenylation distance. Borzoi predicts fine-mapped causal intronic paQTLs well, with an average AUPRC of 0.725 (Extended Data Fig. 9a,b and Supplementary Fig. 13a).

Discussion

In this paper, we propose a new sequence-based machine-learning model, Borzoi, that learns to predict sequencing coverage from a vast set of RNA-seq experiments. Borzoi enables variant scoring and interpretation through multiple layers of regulation, including transcription, splicing and polyadenylation, and demonstrates competitive performance to state-of-the-art models in classifying fine-mapped QTLs. When averaging predictions across an ensemble of model replicates, Borzoi's performance improved further. By applying sequence attribution methods to statistics derived from the predicted coverage tracks, Borzoi provides tissue-specific interpretations of enhancers driving RNA expression and post-transcriptional regulation within the transcript. Through a number of ablation studies, we discovered that training on DNase-seq and ATAC–seq data in addition to RNA-seq consistently improved test set accuracies compared to training on RNA-seq alone and delivered better concordance with eQTL measurements and enhancer–gene linking data. This observation suggests that recent multiome datasets, which measure both accessibility and expression in single cells, would be valuable as joint training data. Variant prediction quality was only marginally affected by whether or not the variant occurs in genomic sequences seen during training, meaning that genetics researchers can ignore this factor when using the model.

Challenges to modeling RNA-seq coverage remain, and Borzoi is far from perfect in predicting these data. For example, although differential 5′ (TSS) and 3′ (APA) isoforms of held-out genes were predicted accurately across tissues, most tissue-specific splicing events were not captured well by the model, which rather tended to predict the average RNA-seq shape. Furthermore, we did not find sequence elements related to mRNA half-life in Borzoi's sequence attributions[23,82]. Disentangling these layers of regulation is particularly difficult in the presence of sequencing bias. For example, reads aligning with greater density at the 3′ end of transcripts[83,84] and other confounders (for example, GC bias) caused false positives as we attempted to classify alternatively used splice sites based on predicted coverage. We also emphasize the importance of choosing appropriate attribution methods to interpret the model. Although input gradients and ISM produced high-quality attributions for splice junctions and enhancer–promoter regions, we found that window-shuffled ISM worked better for 3′ UTRs owing to buffering effects.

For researchers intending to use Borzoi in their genetic variant analyses, we recommend using the gene-centric variant effect scores derived in this paper to prioritize variants with respect to a particular target gene. These scores include (1) predicted exon-aggregated coverage log fold change of the target gene (for abundance differences), (2) predicted maximum difference in coverage log ratio between any 3′ cleavage site (for polyadenylation differences) and (3) predicted maximum normalized difference in any coverage bin within the gene body (for splicing differences). If target genes are unknown a priori, we recommend using a gene-agnostic statistic, such as the one based on total L2 norm, to quantify potential changes in coverage patterns across the entire output window.

In future work, we envision several directions for improvement. We believe that adding training data from additional assays based on RNA-seq will further improve model quality; for example, crosslinking and immunoprecipitation sequencing to measure RNA-binding proteins[85,86], ribosomal profiling to measure translation[87,88] and time-series measuring mRNA half-lives[89,90]. Similarly, we anticipate that training on experiments in which regulatory proteins have been perturbed will improve model performance in general and enable causal inference by tying particular regulators to sequence motifs[91,92]. Data quantity is a critical factor in successful machine learning and we believe that adding RNA-seq from more mammals is a viable path to increasing training data and model quality[93]. Relatedly, training on individual human genomes with matched RNA-seq data from population sequencing efforts like GTEx[33] may help further improve variant effect predictions[94,95]. Finally, we are eager to incorporate new efficient attention modules to boost the receptive field to megabase scale and predict at finer resolution[96].

In summary, we developed a neural network model for predicting RNA coverage from sequence and demonstrated its performance on multiple variant interpretation tasks. Direct modeling of RNA-seq opens the door to studying a wide range of experimental assays, increasing our ability to understand the impact of genetic variation on gene-regulatory processes.

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/s41588-024-02053-6.

References

1. Wang, Q. S. et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
2. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
3. Fowler, D. M. et al. An atlas of variant effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).
4. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
5. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
6. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
7. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).

8. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
9. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).
10. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).
11. Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
12. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
13. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
14. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
15. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
16. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).
17. Leung, M. K. K., Delong, A. & Frey, B. J. Inference of the human polyadenylation code. *Bioinformatics* **34**, 2889–2898 (2018).
18. Vainberg Slutskin, I., Weinberger, A. & Segal, E. Sequence determinants of polyadenylation-mediated regulation. *Genome Res.* **29**, 1635–1647 (2019).
19. Arefeen, A., Xiao, X. & Jiang, T. DeepPASTA: deep neural network based polyadenylation site analysis. *Bioinformatics* **35**, 4577–4585 (2019).
20. Li, Z. et al. DeeReCT-APA: prediction of alternative polyadenylation site usage through deep learning. *Genomics Proteomics Bioinformatics* **20**, 483–495 (2022).
21. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106.e23 (2019).
22. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol.* **23**, 232 (2022).
23. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol.* **23**, 245 (2022).
24. Hallier, M., Tavitian, A. & Moreau-Gachelin, F. The transcription factor Spi-1/PU.1 binds RNA and interferes with the RNA-binding protein p54nrb. *J. Biol. Chem.* **271**, 11177–11181 (1996).
25. Oksuz, O. et al. Transcription factors interact with RNA to regulate genes. *Mol. Cell* **83**, 2449–2463.e13 (2023).
26. Kwon, B. et al. Enhancers regulate 3′ end processing activity to control expression of alternative 3′UTR isoforms. *Nat. Commun.* **13**, 2709 (2022).
27. Vaswani, A. et al. Attention is all you need. In *Proc. 31st Conference on Neural Information Processing Systems (NIPS 2017)* (eds. von Luxburg, U. et al.) 6000–6010 (Curran Associates, 2017).
28. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics* 464–468 (Association for Computational Linguistics, 2018).
29. Lin, T. Y. et al. Feature pyramid networks for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)* 936–944 (IEEE, 2017).
30. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015): Proc. 18th International Conference* (eds Navab, N. et al.) 234–241 (Springer, 2015).
31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
33. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
34. Wilks, C. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
35. Wilks, C. et al. Megadepth: efficient coverage quantification for BigWigs and BAMs. *Bioinformatics* **37**, 3014–3016 (2021).
36. Kimura, K. et al. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).
37. de Klerk, E. & AC't Hoen, P. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* **31**, 128–139 (2015).
38. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018).
39. Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).
40. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
41. Shi, Y. Alternative polyadenylation: new insights from global analyses. *RNA* **18**, 2105–2117 (2012).
42. 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).
43. Zhang, H., Hu, J., Recce, M. & Tian, B. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33**, D116–D120 (2005).
44. Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.* **46**, D315–D319 (2018).
45. Talukder, A., Barham, C., Li, X. & Hu, H. Interpretation of deep learning in genomics and epigenomics. *Brief. Bioinform.* **22**, bbaa177 (2021).
46. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proc. International Conference on Learning Representations (ICLR) 2014* (ICLR, 2014).
47. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. 34th International Conference on Machine Learning (PMLR 70)* (eds Precup, D. & Teh, Y. W.) 3319–3328 (Curran Associates, 2017).
48. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. In *International Conference on Machine Learning, Workshop on Visualization for Deep Learning* (ICML, 2017).
49. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proc. 34th International Conference on Machine Learning (PMLR 70)* (eds Precup, D. & Teh, Y. W.) 3145–3153 (Curran Associates, 2017).
50. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st Conference on Neural Information Processing Systems (NIPS 2017)* (eds. von Luxburg, U. et al.) 1–10 (Curran Associates, 2017).

51. Shrikumar, A. et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. Preprint at https://doi.org/10.48550/arXiv.1811.00416 (2018).
52. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
53. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP–seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
54. Luyten, A., Zang, C., Liu, X. S. & Shivdasani, R. A. Active enhancers are delineated de novo during hematopoiesis, with limited lineage fidelity among specified primary blood cells. *Genes Dev.* **28**, 1827–1839 (2014).
55. Soler, E. et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev.* **24**, 277–289 (2010).
56. Wilson, N. K. et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532–544 (2010).
57. Stadhouders, R. et al. Transcription regulation by distal enhancers: who's in the loop? *Transcription* **3**, 181–186 (2012).
58. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390.e19 (2019).
59. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
60. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
61. Fulco, C. P. et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
62. Klann, T. S. et al. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
63. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285–299.e5 (2017).
64. Huang, J. et al. Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, 943 (2018).
65. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
66. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 56 (2023).
67. Akhtar, W. et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914–927 (2013).
68. Leemans, C. et al. Promoter-intrinsic and local chromatin features determine gene repression in lads. *Cell* **177**, 852–864.e14 (2019).
69. Burda, P., Laslo, P. & Stopka, T. The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia* **24**, 1249–1257 (2010).
70. Oikawa, T. et al. The role of Ets family transcription factor PU.1 in hematopoietic cell differentiation, proliferation and apoptosis. *Cell Death Differ.* **6**, 599–608 (1999).
71. Yanai, H., Negishi, H. & Taniguchi, T. The IRF family of transcription factors: inception, impact and implications in oncogenesis. *Oncoimmunology* **1**, 1376–1386 (2012).
72. Tamura, T., Yanai, H., Savitsky, D. & Taniguchi, T. The IRF family transcription factors in immunity and oncogenesis. *Annu. Rev. Immunol.* **26**, 535–584 (2008).
73. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
74. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
75. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
76. 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).
77. Danckwardt, S., Hentze, M. W. & Kulozik, A. E. 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.* **27**, 482–498 (2008).
78. Forrest, M. E. et al. Codon and amino acid content are associated with mRNA stability in mammalian cells. *PLoS ONE* **15**, e0228730 (2020).
79. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
80. Alasoo, K. et al. Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *Elife* **8**, e41673 (2019).
81. Tian, B., Pan, Z. & Lee, J. Y. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* **17**, 156–165 (2007).
82. Zhao, W. et al. Massively parallel functional annotation of 3′ untranslated regions. *Nat. Biotechnol.* **32**, 387–391 (2014).
83. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).
84. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).
85. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
86. Van Nostrand, E. L. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
87. 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).
88. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213 (2014).
89. Ross, J. mRNA stability in mammalian cells. *Microbiol. Rev.* **59**, 423–450 (1995).
90. Loflin, P. T., Chen, C. Y., Xu, N. & Shyu, A. B. Transcriptional pulsing approaches for analysis of mRNA turnover in mammalian cells. *Methods* **17**, 11–20 (1999).
91. Joung, J. et al. A transcription factor atlas of directed differentiation. *Cell* **186**, 209–229.e26 (2023).
92. Kowalski, M. H. et al. Multiplexed single-cell characterization of alternative polyadenylation regulators. *Cell* **187**, 4408–4425.e23 (2024).
93. Kaplow, I. M. et al. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **380**, eabm7993 (2023).
94. Sasse, A. et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
95. Huang, C. et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.* **55**, 2056–2059 (2023).
96. Chen, B. et al. Scatterbrain: unifying sparse and low-rank attention. *Adv. Neural. Inf. Process. Syst.* **34**, 17413–17426 (2021).

Methods

The experiments conducted in this study did not require approval from a specific ethics board.

Training data

The training data for this analysis consisted of a large set of human and mouse RNA-seq experiments. To help the model use important sequence features for making its RNA coverage predictions, we also included the experimental assays studied by the Enformer and Basenji models in the training data[8,9,13]. This includes a curated set of human and mouse CAGE assays from the FANTOM5 consortium, which we reasoned would help the model relate TSS usage and strength to RNA-seq coverage between multiple (alternative) TSSs[97,98], as well as DNase-seq and ChIP–seq from ENCODE and the Epigenomics Roadmap[31,99] and pseudo-bulk single-cell ATAC–seq data from CATlas[100,101], which focuses the model towards distal regulatory elements. We processed the data slightly differently relative to prior analyses[9,13]. First, we aggregated the aligned read counts here at 32 bp resolution. Second, we split the CAGE-aligned reads by strand, requiring that the model predict both the forward and anti-sense coverage.

We collected 867 human and 278 mouse RNA-seq coverage tracks from ENCODE. This set includes samples from a diverse set of tissues and cell types, with measurements spanning the developmental spectrum for both human and mouse. The tracks available for download represent normalized coverage from the STAR alignment program of uniquely mapping reads[102]. Most experiments used a protocol to enable stranded analysis, creating a forward and anti-sense coverage track. We trained Borzoi to directly predict these continuous coverage values in 32 bp genomic bins. Owing to the relatively large dynamic range of RNA-seq, we normalized each coverage track by exponentiating its bin values by 3/4. If bin values were still larger than 384 after exponentiation, we applied an additional square-root transform to the residual value. These operations effectively limit the contribution that very highly expressed genes can impose on the model training loss. The formula below summarizes the transform applied to the j^{th} bin for tissue t of target tensor $\boldsymbol{y}$:

$$\boldsymbol{y}_{j,t}^{(\text{squashed})} = \left\{ \boldsymbol{y}_{j,t}^{(3/4)} \text{ if } \boldsymbol{y}_{j,t}^{(3/4)} \leq 384, \text{ otherwise } 384 + \sqrt{\boldsymbol{y}_{j,t}^{(3/4)} - 384} \right\}$$

We refer to this set of transformations as 'squashed scale' in the main text. The parameters were chosen such that most genes had bin values of <1,000 (a reasonably large maximum value that is handled well by standard tensorflow data types). For most downstream tasks, for example, when calculating log fold changes from predicted values because of a mutation, we first undo the normalization by applying inverse transforms to the predictions (thus operating in 'count' space). One exception is when visualizing reference predictions of test sequences, in which all transforms except the residual exponentiation at 384 are inverted, as small amounts of noise near the threshold would otherwise be amplified.

We supplemented the training data with 89 tracks from GTEx whole-tissue samples[33], uniformly processed by the recount3 project[34] (GTEx v.8 release). recount3 clustered the 49 GTEx tissues into 30 meta-tissues, combining highly related physiological regions (such as regions of the brain). For each meta-tissue, we chose a subset of samples to include as training data by performing k-means clustering on the gene expression profiles of all samples with $k = 3$ (although several meta-tissues collapsed to $k = 2$). For each cluster, we chose to include the sample with the minimum average distance to all cluster members. These data were processed without consideration of strand information in recount3, which means the GTEx training tracks are non-stranded whereas most other RNA-seq tracks are stranded. For these tracks, we scaled the aligned fragment counts by the inverse of their average length to weight each fragment as a single event, in addition to the exponentiation transform described above.

We fragmented the human (hg38) and mouse (mm10) chromosomes and randomly divided these fragments into eight roughly evenly sized partitions, pairing orthologous regions into the same partition. One partition was held out for validation and another for testing, and the remainder of the data (~75%) was used for training. Note that all coverage measurements of all experimental assays (RNA, DNase, CAGE, ATAC, ChIP) are held out (and not seen by the model) whenever a particular 524 kb sequence window is not in the training set.

Model

The model is based on the Enformer network architecture but introduces a number of simplifications and enhancements to optimize for RNA-seq prediction[13]. Supplementary Fig. 14 shows the full architecture. Enformer comprises two main stages. First, repeated application of a convolution block that achieves a twofold reduction of the sequence length extracts local sequence patterns until each position in the sequence represents 128 bp. Second, repeated application of a self-attention (or transformer) block enables long-range interaction and exchange between every pair of sequence positions[27,28]. Enformer accepts a 196 kb input sequence and predicts coverage data aggregated at 128 bp resolution.

RNA-seq is a base-resolution readout of transcribed RNAs. We believed that it was important to both increase the sequence length and decrease the prediction resolution to model RNA-seq well. Mammalian genes regularly exceed a full span of >100 kb, and if the 5′ or 3′ end of a gene extends outside of the training sequence window (such that its promoter and other regulatory signals are not captured in the receptive field of the network), it will probably obstruct learning. Conversely, mammalian exons regularly cover fewer than 128 bp, and modeling the coverage patterns around these exons at such a coarse resolution can obstruct splice site learning. However, computational limitations make these joint objectives challenging. Therefore, we aimed for a compromise of 524 kb input sequences, predicting at 32 bp resolution.

Halting the convolution and pooling blocks in the vanilla Enformer architecture at 32 bp would mean that the self-attention blocks processed 16,384-length sequences. These blocks require quadratic memory complexity, which exceeds the capability of contemporary GPU/TPU hardware without complicated optimizations. Therefore, we chose to remain at 128 bp resolution for the self-attention blocks. To predict at 32 bp resolution, we instead make use of U-net upsampling techniques from the image segmentation and object detection literature[29,30], which solve an analogous problem of determining image-level content and communicating it back down to pixel resolution annotations. In brief, the output embeddings predicted by the self-attention blocks at 128 bp resolution are upsampled two times by duplicating the embedding vector at each position. We then apply point-wise convolutions to match the number of channels to those of the original convolution tower output (preceding the self-attention blocks) at 64 bp resolution. Finally, we add the upsampled feature map from the self-attention blocks and the intermediate feature map from the convolution tower and apply a separable convolution with a width of three. This workflow is repeated once more using the intermediate feature map with 32 bp resolution from the convolution tower.

As this architecture is still very computationally expensive, we simplified several Enformer components. First, we used max pooling instead of attention pooling, which requires an additional convolution but generally only minimally boosts performance. Second, we apply only a single convolution with a width of five in each block of the initial convolution tower, forgoing the second convolution added in with a residual connection used by Enformer. Third, we reduced the number of self-attention blocks from 11 to 8 to reduce memory usage. Fourth, we used only central mask relative position embeddings given that additional distance functions minimally affected performance.

Training

We trained the model in a multi-task setting to predict coverage for all assays from one species, with a species-specific head attached to the shared model trunk. During training, we alternated human and mouse training batches by dynamically swapping in the corresponding species-specific head. To avoid less accurate predictions on the sequence boundaries (owing to asymmetric visibility), we cropped from each side to focus the loss computation on the center 196,608 bp. We used a Poisson loss function but decomposed the loss analogous to BPnet to separate magnitude and shape terms[7]. Having independent Poisson distributions at each sequence position is mathematically equivalent to a single Poisson distribution representing their sum, followed by allocating the counts to sequence positions using a multinomial distribution. Thus, we apply a Poisson loss on the sum of the observed and predicted coverage and a multinomial loss on the normalized observed and predicted coverage across the sequence length. This decomposition allows us to weight the multinomial shape loss by a greater amount (five times), which we found boosts performance.

Using TensorFlow (v.2.11), backpropagation of this model on a 524 kb sequence maxes out the 40 GB of RAM of a standard NVIDIA A100 GPU. Each model instance was trained using the Adam optimizer with a batch size of two, split across two GPUs for ~25 days, and training stopped when the validation set accuracy plateaued.

We trained four replicate models with random weight initialization and sequence training order. We constructed an ensemble predictor from these four replicates that generally performed better than any individual model. Note that for all analyses in Figs. 1 and 2 in which we evaluate model performance, we do so strictly on fragments from the held-out test set. In subsequent analyses (for example, variant effect prediction in Fig. 5), we make no distinction between train or test splits of hg38. This technically means that the ensemble is applied to genomic loci seen during training. We argue that these are still unbiased analyses, as the evaluations are done on out-of-domain measurements not trained on (for example, the alternative alleles of fine-mapped QTLs and their estimated effects were not part of the training data).

Model ablation experiments

Instances of the Borzoi model were trained on smaller subsets of the original training data to assess the contribution of various data modalities to final performance. We varied whether or not the model was trained on mouse data in addition to human experiments, whether or not the model was trained on additional assays (for example, DNase-seq, ATAC–seq, ChIP–seq and CAGE) in addition to the core RNA-seq modality and whether or not the model used a U-net component to increase the output resolution. Owing to the large number of combinations, it was difficult to acquire a sufficient set of NVIDIA A100 GPUs that would allow training them as full-sized Borzoi models in a reasonable amount of time. Therefore, we reduced their size (393,192 bp input length, ~30 million trainable parameters, four self-attention heads per layer) such that we could fit them with a batch size of two on either NVIDIA RTX 4090 GPUs or NVIDIA TITAN RTX GPUs. We trained two cross-validation folds per ablation condition, choosing a different held-out validation and test set from the eight genomic hg38 or mm10 partitions per fold. We trained four folds for the baseline condition (with all features included). Training lasted 30–90 days, depending on condition, and was stopped when the validation accuracy saturated.

The following model instances were trained: **['Multispecies']** Training data - CAGE, DNase-, ATAC-, ChIP- and RNA-seq in human (hg38) and mouse (mm10). Architecture changes - N/A (baseline model). **['Multispecies (No U-net)']** Training data - CAGE, DNase-, ATAC-, ChIP- and RNA-seq in human and mouse. Architecture changes - U-net removed. Trained at 128 bp output resolution. **['Multispecies (D/A/RNA)']** Training data - DNase-, ATAC- and RNA-seq in human and mouse. Architecture changes - N/A. **['Multispecies (RNA)']** Training data - RNA-seq in human and mouse. Architecture changes - N/A. **['Human']** Training data - CAGE, DNase-, ATAC-, ChIP- and RNA-seq in human. Architecture changes - N/A. **['Human (D/A/RNA)']** Training data - DNase-, ATAC- and RNA-seq in human. Architecture changes - N/A. **['Human (GTEx RNA)']** Training data - GTEx RNA-seq (human). Architecture changes - N/A. **['K562']** Training data - CAGE, DNase-, ChIP- and RNA-seq in K562 cells. Architecture changes - N/A. **['K562 (D/A/RNA)']** Training data - DNase-, and RNA-seq in K562 cells. Architecture changes - N/A. **['K562 (RNA)']** Training data - RNA-seq in K562 cells. Architecture changes - N/A.

Enformer comparison

Our research objective was to extend this modeling framework to new data (that is, RNA-seq) and not to exceed Enformer performance on the set of overlapping tracks, which includes CAGE, DNase, ATAC and ChIP assays. Several modeling decisions make comparisons between Borzoi and Enformer imperfect. First, working with larger sequences required reprocessing the genome so that the held-out test set of Borzoi does not exactly match that of Enformer. Second, we aggregated the data at 32 bp resolution, whereas Enformer works with 128 bp, thus altering the distribution of bin values. Third, we split the aligned reads from the CAGE datasets by strand. Nevertheless, we examined test accuracies for Borzoi versus Enformer (v.3.0) on these overlapping datasets and found them to be broadly similar despite these modifications (Extended Data Fig. 1a–d).

Tissue-specific expression, TSS and APA predictions

We evaluated three different statistics derived from the predicted GTEx RNA-seq coverage tracks to quantify (tissue-specific) gene expression, alternative TSS usage and APA isoform abundance (Fig. 2). Gene expression is quantified as the sum of predicted coverage overlapping exonic bins. Alternative TSS usage is quantified by taking the maximum coverage among the nine bins immediately downstream of each annotated TSS in GENCODE (v.41) (maximum given that the exon may be shorter than nine bins) and computing the ratio between the 3′-most and 5′-most TSSs of each gene. Only TSSs that were within 50 bp of an annotated TSS in FANTOM5 were included[97]. APA site usage is quantified by calculating the ratio of average coverage between the four bins immediately upstream of the distal-most PAS and the four bins upstream of the proximal-most PAS, based on polyadenylation sites annotated in PolyADB[44].

Examples visualized in Fig. 2 and Extended Data Fig. 3 were chosen as follows: (1) differentially expressed examples were selected from the genes with the largest measured fold change between exon-aggregated coverage in the target tissue and the average coverage in the four other tissues, based on the GTEx RNA-seq data; (2) tissue-specific TSS examples were selected from the set of genes with largest measured differential TSS usage according to tissue-matched FANTOM5 CAGE data; and (3) tissue-specific APA examples were selected from the genes with the largest measured fold change in coverage ratio in the target tissue with respect to the average coverage ratio in the four other tissues. To reduce the risk of picking genes in which the perceived APA is driven by 3′ bias in the GTEx RNA-seq data, we required that the genes also exhibited differential distal polyadenylation in cell-type-matched experiments from the PolyASite 2.0 database[42]. All example genes were picked from the held-out test set, and coverage was predicted using the four-replicate ensemble.

Input sequence attribution

To visualize important features in the input sequence (such as TF or RNA-binding protein motifs) and quantify their contribution to the prediction (their saliency score), we apply a number of different attribution methods, each with their own strengths and limitations. In summary, we either use methods based on gradient saliency, which are computationally efficient for single outputs but tend to be noisier owing to moving off the one-hot coding simplex, or in-silico

mutagenesis, which often give better-calibrated attributions for all outputs, but are too computationally expensive to run on long sequences. The shared goal of these methods is to estimate the contribution of each nucleotide in the input with respect to scalar statistics derived from the predicted coverage tracks, resulting in a matrix $\boldsymbol{s} \in \mathbb{R}^{524{,}288\times 4}$ of saliency scores for each coverage track. In this study, we focus solely on interpreting Borzoi's RNA-seq tracks. Furthermore, by computing distinct summary statistics from the predicted RNA coverage tracks, we dynamically isolate distinct regulatory mechanisms in the attribution scores; namely, transcription, polyadenylation and splicing.

As preliminaries, let $\mathcal{M}$ be the Borzoi model, $\boldsymbol{x} \in \{0,1\}^{524{,}288\times 4}$ be the one-hot coded input sequence, $\boldsymbol{y} = \mathcal{M}(\boldsymbol{x}) \in (0,+\infty]^{16{,}384\times 7{,}611}$ be the (human) coverage prediction and $\mathcal{T} = \{t_0, \ldots, t_T\}$ be the set of T indices of the coverage tracks in $\boldsymbol{y}$ that we want to average over (for example, to combine all blood-specific tracks) and compute the attribution scores for. Note that Borzoi's raw prediction $\boldsymbol{y}$ is based on training data that had been subjected to various transforms intended to stabilize training (exponentiating by 3/4, additional exponentiation of residuals above a target value and re-scaling). Here, we assume that we have applied the inverse transforms to $\boldsymbol{y}$ such that the tensor can be reasonably assumed to reflect counts (also note that these transforms are differentiable, which means gradient saliency can be propagated through the inverse operations).

Below are the definitions of three distinct summary statistics used for expression attribution, polyadenylation attribution and splicing attribution, respectively:

Log sum of exon coverage (expression attribution). The summary statistic $u \in \mathbb{R}$ is computed by aggregating the set of 32 bp bins $\mathcal{B} = \{b_0, \ldots, b_B\}$ in $\boldsymbol{y}$ overlapping the exons of the gene of interest (with optional pseudo count $C \in \mathbb{R}$):

$$u = \log\left(C + (1/T) \times \sum_{t\in\mathcal{T}} \sum_{b\in\mathcal{B}} \boldsymbol{y}_{b,t}\right)$$

Log ratio of PAS coverage (polyadenylation attribution). The statistic $u \in \mathbb{R}$ is computed by summing coverage in five adjacent bins immediately upstream of bin b_{prox}, which overlaps the PAS of interest, and dividing by the coverage of a matched set of bins upstream of bin b_{dist}, where a competing PAS is located (or immediately downstream of b_{prox} if the gene of interest is not subject to APA):

$$u = \log\left(\frac{C + (1/T) \times \sum_{t\in\mathcal{T}} \sum_{b=b_{\text{prox}}-5}^{b_{\text{prox}}} \boldsymbol{y}_{b,t}}{C + (1/T) \times \sum_{t\in\mathcal{T}} \sum_{b=b_{\text{dist}}-5}^{b_{\text{dist}}} \boldsymbol{y}_{b,t}}\right)$$

Note that the formula above assumes that the gene is on the forward (plus) strand. Coverage must be summed from $b_{\text{prox}}+1$ to $b_{\text{prox}}+5+1$ (and from $b_{\text{dist}}+1$ to $b_{\text{dist}}+5+1$) if the gene is on the minus strand.

Log ratio of exon-to-intron coverage (splicing attribution). The statistic $u \in \mathbb{R}$ is computed by summing coverage in bins $\mathcal{B}_{\text{exon}} = \{b_0, \ldots, b_E\}$ overlapping the exon and dividing by the sum of coverage in a matched number of bins $\mathcal{B}_{\text{intron}} = \{b_0, \ldots, b_I\}$ overlapping the adjacent intron or, alternatively, a neighboring exon (which occasionally resulted in less noisy attributions when intronic polyadenylation sites created non-uniform intronic coverage):

$$u = \log\left(\frac{C + (1/T) \times \sum_{t\in\mathcal{T}} \sum_{b\in\mathcal{B}_{\text{exon}}} \boldsymbol{y}_{b,t}}{C + (1/T) \times \sum_{t\in\mathcal{T}} \sum_{b\in\mathcal{B}_{\text{intron}}} \boldsymbol{y}_{b,t}}\right)$$

The summary statistics defined above are used in conjunction with the following attribution methods:

Gradient × input (gradients). Given summary statistic $u(\boldsymbol{x})$, the attribution scores $\boldsymbol{s} \in \mathbb{R}^{524{,}288\times 4}$ are computed by taking the gradient with respect to input $\boldsymbol{x}$ and subtracting the mean at each position across nucleotides[103]:

$$\boldsymbol{s}_{i,j} = \frac{\partial u(\boldsymbol{x})}{\partial \boldsymbol{x}_{i,j}} - (1/4) \times \sum_{k=1}^{4} \frac{\partial u(\boldsymbol{x})}{\partial \boldsymbol{x}_{i,k}}$$

When visualizing $\boldsymbol{s}$, we extract the score at position i corresponding to the reference nucleotide j only (which is easily implemented by multiplying with $\boldsymbol{x}$ and aggregating across nucleotides):

$$\boldsymbol{s}_i^{(\text{vis})} = \sum_{j=1}^{4} \boldsymbol{s}_{i,j} \times \boldsymbol{x}_{i,j}$$

ISM. Given a start and end position, p_{start} and p_{end}, in $\boldsymbol{x}$ to compute ISM over, the attribution scores $\boldsymbol{s} \in \mathbb{R}^{524{,}288\times 4}$ are computed as follows: create a new tensor $\tilde{\boldsymbol{x}} \in \{0,1\}^{(p_{\text{end}}-p_{\text{start}})\times 4\times 524{,}288\times 4}$ and let each matrix $\tilde{\boldsymbol{x}}_{u,v}$ hold a mutated copy of $\boldsymbol{x}$ where the reference nucleotide at position u is substituted for nucleotide v. Then compute the ISM scores $\boldsymbol{s}$ as:

$$\boldsymbol{s}_{i,j} = u(\boldsymbol{x}) - u(\tilde{\boldsymbol{x}}_{i-p_{\text{start}},j}), \text{ if } p_{\text{start}} \le i \le p_{\text{end}}, 0 \text{ otherwise}.$$

When visualizing $\boldsymbol{s}$, we average the scores across the four nucleotides:

$$\boldsymbol{s}_i^{(\text{vis})} = (1/4) \times \sum_{j=1}^{4} \boldsymbol{s}_{i,j}$$

Window-shuffled ISM (ISM shuffle). Given a start and end position, p_{start} and p_{end}, a window size M and a number of re-shuffles N, the attribution scores $\boldsymbol{s} \in \mathbb{R}^{524{,}288\times 4}$ are computed as follows: create tensor $\tilde{\boldsymbol{x}} \in \{0,1\}^{(p_{\text{end}}-p_{\text{start}})\times N\times 524{,}288\times 4}$ containing $(p_{\text{end}} - p_{\text{start}}) \times N$ copies of input pattern $\boldsymbol{x}$. For each matrix $\tilde{\boldsymbol{x}}_{u,v}$ (where v denotes one of N independent samples), either dinucleotide-shuffle the local region $[u - M/2, u + M/2 + 1]$ or replace the reference nucleotides in this region with uniformly random nucleotides. Dinucleotide shuffling (with $M = 7$ and $N = 24$, or $N = 8$ for large window sizes) is performed when computing enhancer saliency, whereas uniform random substitution ($M = 5$ and $N = 24$, or $N = 8$ for large window sizes) is used for promoters, splice sites and PASs (where salient features are often stretches of repeating nucleotides). Then compute the attribution scores $\boldsymbol{s}$ as:

$$\boldsymbol{s}_{i,n} = u(\boldsymbol{x}) - u(\tilde{\boldsymbol{x}}_{i-p_{\text{start}},n}), \text{ if } p_{\text{start}} \le i \le p_{\text{end}}, 0 \text{ otherwise}.$$

When visualizing $\boldsymbol{s}$, we average the scores across the N samples:

$$\boldsymbol{s}_i^{(\text{vis})} = (1/N) \times \sum_{n=1}^{N} \boldsymbol{s}_{i,n}$$

Tissue-specific motif discovery

We visualized learned tissue-specific *cis*-regulatory motifs driving RNA coverage in GTEx tracks through a combination of (1) picking a large set of (measured) highly tissue-specific genes, (2) computing their gradient saliencies and normalizing out tissue-shared saliency and (3) clustering and annotating the saliency scores using TF-MoDISco (v.0.5.14.1)[51] and Tomtom MEME suite (v.5.5.2)[52]. We first downloaded measured TPMs for GTEx (v.8) (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz). We heuristically cleaned the data by adding a small pseudo-TPM that was roughly the first percentile of all values (to avoid zeros), followed by clipping at a value slightly larger than the 99$^{\text{th}}$ percentile per tissue (to avoid extremely large numbers). Then, for each of the five prospective GTEx tissues whole blood, liver, brain - cortex, muscle - skeletal and esophagus - muscularis,

we computed gene-specific log fold changes of TPM expression for the tissue of interest relative to the average TPM expression of the four other tissues. For each tissue, we sorted the TPM matrix in descending order of this metric and selected the top 1,000 most differentially expressed genes, resulting in a total of 5,000 genes.

We computed nucleotide-level attribution scores (input gradients) with respect to the log of aggregated exon coverage for each of the 5,000 genes, repeating the gradient computation for each of the five GTEx tissues. Specifically, we matched each GTEx tissue to the corresponding two to three RNA coverage tracks obtained from recount3 that we trained on (for example, for brain - cortex, we computed the input gradient saliency with respect to the three GTEx brain meta-tissue tracks). The gradient computation was repeated for all four model replicates, for both forward-complemented and reverse-complemented input sequences, and averaged.

The gradient computation outlined above produces five separate sets of saliency scores for all 5,000 genes (one set of scores per tissue). Next, we performed de novo motif discovery for tissue *x* by slicing out the 1,000 genes originally selected to be differentially upregulated in tissue *x* and running TF-MoDISco on the residual gradient scores for tissue *x*. The residual scores were calculated by subtracting the average gradient of the four other tissues from those of tissue *x*, thus dampening the saliency of shared regulatory motifs and accentuating motifs specific to tissue *x*. Additionally, before running MoDISco, we first re-weighted the gradients by computing the standard deviation at each position across the four nucleotides, applying a Gaussian filter (s.d. = 1,280; truncate = 2) to the resulting vector of standard deviations and dividing the gradient scores by this smoothed vector. This operation results in down-weighting of regulatory regions with long contiguous stretches of large magnitude (often promoter regions) and up-weights sparser regulatory regions (transcriptional enhancers). To increase computational efficiency, we extracted the centered-on 131 kb gradient scores (as opposed to the full 524 kb) before calling MoDISco. TF-MoDISco was executed with the following parameters: 'revcomp = true', 'trim_to_window_size = 24', 'initial_flank_to_add = 8', 'sliding_window_size = 18', 'flank_size = 8' and 'max_seqlets_per_metacluster = 40,000'. Other parameters were kept at their default values.

The five tissue-specific MoDISco result objects were filtered and pooled as follows: Tomtom MEME was used to match the position weight matrices of each MoDISco cluster to HOCOMOCO (v.11)[53] motifs (each position weight matrix was trimmed by an information content threshold of >0.1). Only matches with *E* values of ≤0.1 were retained. The match with the lowest *P* value was chosen as the representative motif for that cluster. The five MoDISco objects were pooled by matching clusters with identical HOCOMOCO motifs and merging the seqlet coordinates, resulting in a single list of seqlet coordinates for each putative motif. A scalar tissue-specific saliency score was then computed for each seqlet by averaging the input-gated gradients overlapping its coordinates. The distributions of these seqlet-level gradient saliencies were used to assess the tissue-specificity of each motif.

Replicating the entire analysis with pseudo counts added to the predicted sum of exon coverage before applying log and computing gradients resulted in nearly identical results. Replicating the analysis without running TF-MoDISco on residual attribution scores but rather using the raw gradients from each tissue-specific coverage track as input to TF-MoDISco similarly produced negligible differences.

Tissue-pooled splice motif discovery

Splice-regulatory motifs were generated by computing input gradients with respect to the splicing attribution statistic (log ratio of exon-to-intron coverage) for one randomly chosen exon in each of the 4,778 genes from the Gasperini dataset[58]. The gradients were computed with respect to the average predicted coverage taken across all 89 of Borzoi's GTEx RNA-seq tracks. The gradients were normalized across genes as follows: we first computed the standard deviation across the four nucleotides and found the maximum standard deviation across all 524,288 positions per gene. We clipped the lower end of the 4,778 maximum deviations at the 25th percentile (to avoid up-weighting gradients with very low magnitudes) and divided each gene's gradient by this number. We tried varying the percentile threshold (from 1 to 100) and the results were robust to this parameter (the same motif clusters were identified with roughly the same number of supporting seqlets). Finally, to obtain 5′ splice motifs, we extracted a 192 bp window centered on the splice donor from each of the gradients. To obtain 3′ splice motifs, we extracted a 192 bp window around the splice acceptor.

TF-MoDISco was executed on the resulting 4,778 × 192 × 4 hypothetical scores, using custom parameter settings that we empirically found worked better for degenerate RNA-binding protein motifs: 'revcomp = false', 'trim_to_window_size = 8', 'initial_flank_to_add = 2', 'sliding_window_size = 6', 'flank_size = 2', 'max_seqlets_per_metacluster = 40,000', 'kmer_len = 5', 'num_gaps = 2' and 'num_mismatches = 1'.

Tissue-pooled polyadenylation motif discovery

Salient motifs related to PASs were obtained in a process similar to the procedure for splice-regulatory motif discovery. We computed tissue-pooled gradients with respect to the polyadenylation statistic (log ratio of PAS coverage) for the distal-most PAS of each gene from the Gasperini dataset[58]. The gradients were normalized by the (clipped) maximum standard deviation per gene. Finally, a 192 bp window centered on the mode of saliency in the 3′ UTR of each gene was used to extract short gradient slices. These gradient slices were used as hypothetical scores for TF-MoDISco, which was executed using the same custom parameters as was used for splice motif discovery.

Attention matrix visualization

We visualized higher-order structures and long-range interactions learned by Borzoi directly through the attention score matrices of the self-attention layers. Examples of such higher-order structures include intronic and exonic regions, UTRs, promoters and gene spans. Long-range interactions describe relationships or dependencies between these structures learned by Borzoi, which would be observed as off-diagonal intensities in the attention matrix. Such examples include phenomena in which an intron attends to its nearest exon junction, a 3′ UTR attends to its PASs or gene spans attend to promoters and transcriptional enhancers. After exploring the predicted attention maps for several different loci, we noticed that higher-order structures matching GENCODE annotations[104] were generally found in the later self-attention layers. However, to mitigate capturing potential assay-specific or experiment-specific biases and focus on general knowledge, we decided not to use the two final attention layers and instead used the two penultimate self-attention layers for all analyses. We further noted that different attention heads tended to capture mostly the same trends, leading us to analyze the mean attention of all eight heads.

Let $\boldsymbol{a}_{i,j}^{l,h} = \mathrm{softmax}\left(\boldsymbol{q}_i \boldsymbol{k}_j^T / \sqrt{K} + \boldsymbol{r}_{i,j}\right) \in \mathbb{R}^{N \times N}$ be the attention matrix for head h of layer l, where $\boldsymbol{q}_i$ is the i^{th} query vector, $\boldsymbol{k}_j$ is the j^{th} key vector, $\boldsymbol{r}_{i,j}$ is the positional encoding and K is the key or query size. We obtain the final attention matrix to be visualized as an unweighted average of all heads of the two penultimate layers: $(1/16) \times \sum_{l=6}^{7} \sum_{h=1}^{8} \boldsymbol{a}_{ij}^{lh}$. When zooming in on smaller sections of the attention matrix, we apply a small Gaussian filter to smooth out high-frequency noise ($\sigma = 0.5$, truncate = 2.0). We further average the attention matrix over four independent model replicates and reverse-complemented input sequences. Promoters generally had higher magnitude attention values than exons, leading us to clip individual entries in the average attention matrix at 0.005 (each row of 4,096 entries sums to 1.0).

Fine-mapped eQTL classification and regression tasks

eQTL studies deliver valuable data for evaluating whether Borzoi identifies the correct nucleotides driving expression and their sensitivity

to specific alternative alleles. We studied GTEx (v.8) eQTL results from 49 tissues of varying sample sizes. We made use of summary statistics and fine-mapping results generated with SuSiE in a previous publication[1]. Only fine-mapped causal eQTLs with a posterior causal probability (PIP) of ≥0.9 were kept as positives. We focused all analyses on single nucleotide variants only because insertions and deletions (indels) introduce technical variance caused by shifted prediction boundaries, which we aspire to alleviate in future work. To visualize the measured RNA-seq coverage tracks in individuals with or without the minor allele(s) of interest, we also made use of whole genome sequencing genotyping data of GTEx subjects obtained through dbGAP (http://www.ncbi.nlm.nih.gov/gap).

Inspired by the expression modifier score construction presented in a previous work[1], in which the authors demonstrated that functional eQTL classification probabilities enable improved fine-mapping, we evaluated Borzoi and other models at the task of discriminating fine-mapped causal eQTLs from a negative set chosen to control for TSS distance. To compare against models with multiple generic outputs, we constructed a feature vector based on the model predictions for each variant and trained a random forest classifier with the eQTL causal and non-causal labels. We considered a 'SUM' score and an 'L2' score to define these SNP features. For both score types, we start by centering the 524 kb input window on the SNP of interest and predict coverage $\boldsymbol{y}^{(\mathrm{ref})} = \mathcal{M}(\boldsymbol{x}^{(\mathrm{ref})}), \boldsymbol{y}^{(\mathrm{alt})} = \mathcal{M}(\boldsymbol{x}^{(\mathrm{alt})}) \in \mathbb{R}^{16{,}384\times 7{,}611}$ for the reference and variant patterns, respectively. When computing the SUM score vector $\boldsymbol{u}(\boldsymbol{x}^{(\mathrm{ref})}, \boldsymbol{x}^{(\mathrm{alt})}) \in \mathbb{R}^{7{,}611}$ for the 7,611 distinct Borzoi tracks, we aggregate the difference between coverage predictions $\boldsymbol{y}^{(\mathrm{ref})}$ and $\boldsymbol{y}^{(\mathrm{alt})}$ across the length axis independently per track:

$$\boldsymbol{u}_t = \sum_{j=1}^{16{,}384} \left(\boldsymbol{y}_{j,t}^{(\mathrm{alt})} - \boldsymbol{y}_{j,t}^{(\mathrm{ref})}\right)$$

For the L2 score vector, we compute the L2 norm of the difference between predictions $\boldsymbol{y}^{(\mathrm{ref})}$ and $\boldsymbol{y}^{(\mathrm{alt})}$ across the length axis independently for each track. Before applying the L2 norm, we first log transform the coverage track bins to focus on fold change rather than absolute change. The final metric is calculated as:

$$\boldsymbol{u}_t = \sqrt{\sum_{j=1}^{16{,}384} \left(\log_2(1+\boldsymbol{y}_{j,t}^{(\mathrm{alt})}) - \log_2(1+\boldsymbol{y}_{j,t}^{(\mathrm{ref})})\right)^2}$$

The L2 score extracts more information and achieves greater performance on this task for Borzoi. All previous Enformer work uses the SUM score, but we observed here that it also benefits from L2, though less than Borzoi.

For the second task, we evaluated models on their ability to predict eQTL effect sizes, which is a critical component of a system tasked with predicting gene expression values across a population of individuals. Given that the Borzoi and Enformer models make use of gene annotation differently to map predictions to genes, we chose to perform a gene-agnostic analysis for a less biased comparison. Thus, we filtered the variant set for only those with a consistent sign of the estimated eQTL effect sizes across genes and chose the effect size with maximum absolute value as the representative effect size for that particular fine-mapped SNP. For a subset of GTEx tissues, we were able to select an appropriately matched CAGE experiment from Enformer's outputs and computed the SUM score. For Borzoi, we selected the matching GTEx tissue RNA-seq output and computed a 'logSUM' score, in which we transformed the bin predictions $\boldsymbol{y}$ by $\log_2(\boldsymbol{y}+1)$ before taking a sum over the length axis. In supplementary analyses, we performed gene-specific coefficient analyses using a variant statistic termed 'logSED' ('sum of expression differences'), in which we aggregated predicted coverage in the bins $\mathcal{B} = \{b_1, \ldots, b_K\}$ overlapping the exons of the target gene, and compared the log fold change between alternate and reference alleles: $\log_2\left(\sum_{k=1}^{K} \boldsymbol{y}_{\mathcal{B}(k)}^{(\mathrm{alt})}\right) - \log_2\left(\sum_{k=1}^{K} \boldsymbol{y}_{\mathcal{B}(k)}^{(\mathrm{ref})}\right)$.

For the third task, we evaluated Borzoi's ability to identify the gene(s) affected by an eQTL from the set of local genes, which is intended to estimate how accurately the model can prioritize the correct gene at more general GWAS loci. We downloaded fine-mapped eQTL credible sets and their associated eGenes for 49 GTEx tissues from the eQTL catalog (release 5)[79,80]. The credible set files were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/credible_sets/ (e.g. ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/credible_sets/GTEx_ge_adipose_subcutaneous.purity_filtered.txt.gz).

Note: These file paths have since changed but historical versions can be found at https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/00ea8a7abca895f26c3aee74ece1307dc5054ace/tabix/tabix_ftp_paths.tsv. To download credible sets with the latest file path table, use column 'ftp_cs_path' (e.g. for adipose_subcutaneous, download file ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/susie/QTS000015/QTD000116/QTD000116.credible_sets.tsv.gz).

For each variant within a credible set, we predicted a gene-specific L2 score, which considers only sequence positions overlapping the genes' exons, for all genes within a 360,448 bp sequence window centered on the variant. For each credible set, we computed a single score for each surrounding gene by averaging the gene's score across variants weighted by their posterior causal probabilities. For each GTEx tissue, we computed a variant's L2 score using model predictions for the matched GTEx RNA-seq tracks. We analyzed only credible sets associated with protein-coding genes. Owing to the indel challenge described above, we further removed credible sets in which a fine-mapped variant (PIP > 0.1) is an indel. We predicted a credible set's target gene as the gene with the highest aggregate PIP-weighted L2 score for that credible set. As a baseline, we predicted a credible set's target gene as the nearest gene. We define 'nearest gene' as the gene with the maximum PIP-weighted inverse distance from the credible set. Maximizing the PIP-weighted inverse distance outperforms the previously described approach of minimizing the PIP-weighted distance[105]. Notably, a single distal credible set variant can inflate the minimum average distance statistic, resulting in an incorrect eGene prediction, whereas maximizing the inverse distance does not lead to this problem.

Fine-mapped paQTL classification task

We benchmarked Borzoi's ability to predict genetic variants that alter the relative abundance of mRNA 3′ isoforms using fine-mapped 3′ QTLs (referred to in this paper as polyadenylation QTLs) obtained from the eQTL catalog via txrevise processing[79,80]. The file paths to the fine-mapping results were obtained from https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/master/tabix/tabix_ftp_paths.tsv.

Table rows were filtered by study = 'GTEx' and quant_method = 'txrev'. The resulting sumstat files (for example, 'GTEx_txrev_adipose_subcutaneous.all.tsv.gz') were changed to fine-map files ('GTEx_txrev_adipose_subcutaneous.purity_filtered.txt.gz') and downloaded from ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/credible_sets/ (e.g. ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/credible_sets/GTEx_txrev_adipose_subcutaneous.purity_filtered.txt.gz).

Note: These file paths have since changed but a historical version of the file path table can be found at https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/00ea8a7abca895f26c3aee74ece1307dc5054ace/tabix/tabix_ftp_paths.tsv. To download credible sets with the latest file path table, use column 'ftp_cs_path' (e.g. for adipose_subcutaneous, download file ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/susie/QTS000015/QTD000119/QTD000119.credible_sets.tsv.gz).

To build negative sets of GTEx SNPs that are not part of any txrevise credible set, we obtained rows from the file path table where quant_method = 'ge' and downloaded the full sumstat files from ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GTEx/ge/ (e.g. ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GTEx/ge/GTEx_ge_adipose_subcutaneous.all.tsv.gz). These file paths have also changed;

to download sumstat files with the latest file path table, use column 'ftp_path' (e.g. for adipose_subcutaneous, download file ftp://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/QTS000015/QTD000116/QTD000116.all.tsv.gz).

Fine-mapped causal paQTLs for a given tissue were obtained from the corresponding fine-mapping file ('XYZ.purity_filtered.txt.gz') by filtering on rows in which molecular_traid_id contained the substring '.downstream.', the SNP occurred at most 50 bp outside of a gene span (GENCODE v.41), the distance to the nearest annotated 3′ UTR PAS in PolyADB (v.3)[44] was at most 10,000 bp and PIP was ≥0.9. Valid negatives were obtained from the tissue's sumstat file ('XYZ.all.tsv.gz') with identical gene-span and PAS distance filters as the fine-mapped paQTLs. Negative SNPs had to be either absent from all credible sets or have PIP < 0.01 across all GTEx tissues. Finally, we selected one negative SNP for each fine-mapped causal paQTL by requiring that they have identical distances to an annotated PAS and that the negative SNP occurs in a gene with expression levels that are within (and less than) 1.5-fold the expression level of the paQTL gene (in the same tissue). This resulted in 1,058 retained unique fine-mapped causal paQTLs. The following procedure was used to efficiently search for negative SNPs fulfilling these requirements for a given tissue:

Step 1. Discretize and bin the $\log_2$(TPM) values of all genes (GTEx v.8) into buckets of size 0.4 (in $\log_2$-space). Step 2. For a given query gene (and its associated $\log_2$(TPM) value), take all candidate genes that map into the same bucket. Scan this subset of genes for any gene that contains a distance-matched non-causal SNP. Step 3. If none of the genes in the bucket are suitable candidates (none have a non-causal distance-matched SNP), then subtract 0.15 from the query $\log_2$(TPM) value and take all candidate genes that were binned into the new bucket (if subtracting 0.15 does not change the bucket, skip to Step 4). Scan this new bucket for suitable genes. Step 4. If no suitable gene has been found, repeat Step 3 but instead add 0.15 rather than subtract 0.15 to the original $\log_2$(TPM) value. Scan this (potentially) new bucket for suitable genes. Step 5. If no suitable gene has been found, exit with an error (unmatchable).

The maximum $\log_2$ fold change that two genes can be within and still match is 0.4 + 0.15 = 0.55 (~1.464-fold). With these parameter settings, each bucket contained at least 100 genes, and we never exited Step 5 with an error.

Note that owing to the relatively small number of fine-mapped paQTLs, we decided to pool all tissues rather than benchmark separately per tissue. Given that many of the positives are shared between tissues (there are a total of 1,058 unique paQTLs, each occurring in at least one tissue), we end up with ~2.5× the amount of unique negative SNPs after merging across tissues. Hence, for the benchmark, we performed 100 permutations of randomly matching one of the multiple valid negative SNPs (from different tissues) to each corresponding positive SNP and evaluated performance on each permutation set of 1,058 positives and 1,058 sampled negatives.

Intronic paQTLs (and matched negatives) were extracted from the same files as above but had to occur in intronic regions and be closer to an annotated intronic polyadenylation site than any 3′ UTR polyadenylation site. Negatives were now matched by distance to the nearest intronic PAS. A total of 567 fine-mapped causal intronic paQTLs were retained.

Polyadenylation variant effect prediction

We compute polyadenylation-centric variant effect scores from Borzoi's predicted RNA coverage tracks as the maximum ratio of coverage fold change between any annotated 3′ cleavage junction within the UTR of the same gene as the SNP. Specifically, we center the 524 kb input window on the SNP, predict coverage tracks $\boldsymbol{y}^{(\mathrm{ref})} = \mathcal{M}(\boldsymbol{x}^{(\mathrm{ref})}), \boldsymbol{y}^{(\mathrm{alt})} = \mathcal{M}(\boldsymbol{x}^{(\mathrm{alt})}) \in \mathbb{R}^{16{,}384\times 7{,}611}$ given the reference and alternate allele sequences $\boldsymbol{x}^{(\mathrm{ref})}$ and $\boldsymbol{x}^{(\mathrm{alt})}$ as input and compute the statistic $\boldsymbol{u}(\boldsymbol{y}^{(\mathrm{ref})}, \boldsymbol{y}^{(\mathrm{alt})})_t$ for coverage track t as follows:

$$\boldsymbol{u}_t = \max_{k=1}^{K-1} \left| \log_2 \left(\frac{(1/k)\times \sum_{u=1}^{k} \left(\sum_{j=\mathcal{B}(u)-4}^{\mathcal{B}(u)} \boldsymbol{y}_{j,t}^{(\mathrm{alt})} \right) / \left(\sum_{j=\mathcal{B}(u)-4}^{\mathcal{B}(u)} \boldsymbol{y}_{j,t}^{(\mathrm{ref})} \right)}{(1/(K-k-1)) \times \sum_{u=k+1}^{K} \left(\sum_{j=\mathcal{B}(u)-4}^{\mathcal{B}(u)} \boldsymbol{y}_{j,t}^{(\mathrm{alt})} \right) / \left(\sum_{j=\mathcal{B}(u)-4}^{\mathcal{B}(u)} \boldsymbol{y}_{j,t}^{(\mathrm{ref})} \right)} \right) \right|$$

K in the equation above denotes the total number of PASs within the UTR. $\mathcal{B} = \{b_1, \ldots, b_K\}$ is the ordered set of bin indices in $\boldsymbol{y}$ overlapping the K PASs. The final score used in the benchmarks was the average statistic computed from all of Borzoi's 89 GTEx coverage tracks. The score was also averaged over all four model replicates in both forward-complemented and reverse-complemented input formats.

Comparison to APARENT2 and Saluki

We compare Borzoi's classification performance to APARENT2 (v.1.0.2) in two ways. First, we score the reference and alternate PAS sequence affected by the variant using APARENT2 and simply use the absolute value of the predicted log odds ratio as the variant effect score. Second, we use the predicted odds ratio to scale the tissue-pooled reference PAS usage (as reported in PolyADB) and use the absolute value of the difference in PAS usage as the final variant effect score. The latter statistic effectively dampens the magnitude of variants, which, based on APARENT2's prediction, has a large predicted fold change but, according to measurements, occur in lowly used PASs (owing to competing PASs).

When comparing performance to an ensemble consisting of both APARENT2 and Saluki (v.1.0.0) on the paQTL classification task, we follow the methodology from the APARENT2 paper[22]. In brief, we curate the PAS sequences and corresponding mRNA isoforms of each gene (at most 30) based on annotations from PolyADB and fit a logistic regression model to predict tissue-pooled distal isoform proportions (as reported in PolyADB) given both APARENT2's PAS scores (at most 30 scalars) and Saluki's isoform scores (at most 30 vectors of top four PCA components extracted from the penultimate layer of Saluki) as input. Using this calibrated ensemble model, we predict the reference and alternate distal proportions of a gene when inducing a particular variant (which may affect multiple PAS- and isoform sequences). We estimate a final odds ratio from the predicted distal proportions and use the odds ratio to recalculate the alternate distal proportion based on the measured reference distal proportion. Finally, we subtract the alternate distal proportion from the reference proportion and use the absolute value of this difference as the final variant effect score.

Fine-mapped sQTL classification task

Fine-mapped sQTLs and matched negatives were obtained from the eQTL catalog[79,80] using the same sumstat and fine-mapping files as were used for the paQTL classification task. The fine-mapped causal sQTLs were extracted by filtering on rows in which molecular_trait_id contained the substring '.contained.'. These QTLs were further filtered on PIP ≥ 0.9 and on a maximum distance of ≤10,000 bp to an annotated splice junction (GENCODE v.41). A set of expression-matched and distance-matched negatives were constructed per tissue in an identical fashion to the paQTL task, with the exception of matching by nearest distance to splice junctions. We retained a total of 4,105 unique fine-mapped causal sQTL SNPs.

Splicing variant effect prediction

Purely isolating splicing impact from other mechanisms proved challenging. We focus on a simple statistic that worked well in practice; namely, the maximum difference in normalized coverage across the gene span. Specifically, we center the 524 kb input window on the SNP, predict coverage tracks $\boldsymbol{y}^{(\mathrm{ref})} = \mathcal{M}(\boldsymbol{x}^{(\mathrm{ref})}), \boldsymbol{y}^{(\mathrm{alt})} = \mathcal{M}(\boldsymbol{x}^{(\mathrm{alt})}) \in \mathbb{R}^{16{,}384\times 7{,}611}$ and compute the statistic $\boldsymbol{u}(\boldsymbol{y}^{(\mathrm{ref})}, \boldsymbol{y}^{(\mathrm{alt})})_t$ for coverage track t as follows:

$$\boldsymbol{u}_t = \max_{j=b_{\mathrm{start}}}^{b_{\mathrm{end}}} \left| \frac{\boldsymbol{y}_{j,t}^{(\mathrm{alt})}}{\sum_{k=b_{\mathrm{start}}}^{b_{\mathrm{end}}} \boldsymbol{y}_{k,t}^{(\mathrm{alt})}} - \frac{\boldsymbol{y}_{j,t}^{(\mathrm{ref})}}{\sum_{k=b_{\mathrm{start}}}^{b_{\mathrm{end}}} \boldsymbol{y}_{k,t}^{(\mathrm{alt})}} \right|$$

The indices b_{start} and b_{end} refer to the bins in $\boldsymbol{y}$ overlapping the start and end positions of the gene span. The relatively large number of fine-mapped causal sQTLs allows for a tissue-specific benchmark comparison. To that end, for a given SNP and GTEx tissue, we average the statistic only over the subset of tracks corresponding to the tissue.

Comparison to Pangolin

We used the pre-packaged command-line utility to score sQTL SNPs with Pangolin (v.1.0.1)[16]. To make comparisons easier, we modified the program to output scores with six rather than two decimals. We used the following command to score the positive and negative vcf files: pangolin -d 2,000 -m False < sqtl file > .vcf hg38.fa gencode41_basic_nort_protein.db < out_dir >.

Although this command allows at most a distance of 2,000 bp from an annotated splice junction, Pangolin will also score potential de novo splice gains at the variant position, meaning that the command will produce variant effect scores for all variants (even those separated by >2,000 bp from a splice site). We parsed the command-line output and matched the gene identifier of the Pangolin output to the gene in which the SNP occurs. The final variant effect score is calculated as the sum of the absolute values of the predicted maximum increase and decrease.

Splice site identification task

Identifying splice sites in DNA sequences has formed the basis for a successful approach to interpreting the splicing code and prioritizing pathogenic splicing variants[15,16]. To evaluate Borzoi's ability to identify splice sites, we constructed an analogous classification task and compared it to Pangolin[16]. We downloaded the splicing junction counts for all GTEx samples from recount3 and selected positive examples from annotated junctions with coverage above the 50th percentile of aligned read counts. We filtered this set for those that fall in the intersection of Pangolin's and Borzoi's test sets. For each positive example, we selected a matching negative site that had the same tri-nucleotide context, was between 100 bp and 2,000 bp away and lacked evidence of being a splice junction itself. For Borzoi, we scored each site as the predicted log ratio of exon-to-intron coverage around the junction, averaged across samples from the corresponding GTEx tissue. For Pangolin, we scored each site with its predicted splice site probability, averaged across all tissues.

Classifying rare and common variation from gnomAD

We sampled a set of 14,198 singletons and 14,198 matched common variants (allele frequency > 5%) from the GnomAD (v.3.1) database (https://gnomad.broadinstitute.org), with sampling restricted to regions overlapping ENCODE candidate *cis*-regulatory elements. To control for sequence mutability, we excluded variants within CpG islands and low-complexity regions. For each singleton sampled, we sampled a negative example as a matched common variant with the same reference and alternate allele as the singleton. We also matched the variants' background DNA contexts, sampling common variants that lie within the same tri-nucleotide as the singleton. Finally, we removed variants overlapping gene exons in coding sequences (GENCODE v.41), focusing only on regulatory variants for our evaluation. For all sampled variants, we used their CADD raw score and CADD phred scores (v.1.6) from the GnomAD (v.3.1) dataset. We trained ridge regression models to discriminate common variants from singletons and used tenfold cross-validation to evaluate the models. The CADD-based model uses the CADD scores as features, whereas the Borzoi-based model uses the L2 scores across all RNA-seq tracks as features, averaged across the four model replicates. We derived a third (combined) model by averaging predicted variant ranks for the Borzoi-based and CADD-based models. For a second genome-wide benchmark, we sampled uniformly from across the genome instead of restricting the variant sampling to ENCODE candidate *cis*-regulatory elements. This resulted in a variant set containing 17,360 singletons and 17,360 matched common variants.

Predicting TRIP expression

We downloaded TRIP insertion coordinates and measured expression levels for seven distinct promoters from the supplementary material of a previous publication[68]. The promoter sequences are listed in Table S1 and the insertion coordinates (and measurements) are listed in Data S2 of that paper. To predict the activity of TRIP reporters, we iterated over each promoter sequence and coordinate, centered the 524 kb input window on the insertion coordinate and inserted the sequence. When deriving statistics from Borzoi's RNA-seq or CAGE predictions, we inserted the entire TRIP reporter into the genomic location (including the promoter sequence, the GFP CDS, the PAS and the PiggyBac terminal repeat regions). By contrast, when deriving statistics from Borzoi's DNase or histone modification tracks (for example, H3K4me3) we only inserted the promoter, as these predictions became marginally worse when inserting the full reporter. We attribute this phenomenon to the PiggyBac transposable elements flanking the reporter, which Borzoi inherently does not predict well owing to the clipping of unmappable regions during the original training data processing.

Given the predicted coverage $\boldsymbol{y} = \mathcal{M}(\boldsymbol{x}) \in \mathbb{R}^{16,384\times T}$ for the T coverage tracks considered (for example, K562 DNase tracks), we calculate a scalar prediction $u(\boldsymbol{x}) \in \mathbb{R}$ by averaging the coverage tracks, aggregating the signal in a local window of size W centered at the insertion site and applying a $\log_2$ transform:

$$u = \log_2\left((1/T) \times \sum_{t=1}^{T} \sum_{j=-W/2}^{W/2} \boldsymbol{y}_{8,192+j,t}\right)$$

For CAGE and RNA-seq outputs, we used a 4,096 bp window size that tightly covered the full reporter construct (and tightly covered the average signal profile, as exemplified in Supplementary Fig. 8a for promoter ARHGEF98). Although this was technically a sub-optimal choice (a narrow 128 bp window maximized the average Spearman's R across promoters for RNA-seq; see Supplementary Fig. 8b), the difference in Spearman's R was small (for example, <0.02 for ARHGEF98) and a 4,096 bp window size was a more intuitive choice. Similarly, the average optimal CAGE window size was 8,813 bp, but the 4,096 bp window had near-identical performance (<0.01 difference in average Spearman's R across promoter types). For DNase and histone ChIP tracks, we used a slightly wider 8,192 bp window size as we noticed that the correlation to measured expression saturated less quickly than CAGE as a function of window size (for example, ~0.02 difference in average Spearman's R across promoter types when comparing a window size of 4,096 bp to 8,192 bp for H3K4me3).

Gene–enhancer prioritization task

We evaluated Borzoi's ability to link distal regulatory elements to genes by analyzing experiments in which CRISPRi was used to block the regulatory element followed by measuring gene expression. These experiments have been performed on a small set of specifically chosen genes in which expression was measured by various techniques[60–65] and a large set of all expressed genes in which perturbation and expression were measured by single-cell RNA-seq (scRNA-seq)[58]. These datasets were analyzed to consider whether each tested regulatory element significantly altered gene expression, defining a set of binary labels. The flow/proliferation dataset contains 117 positives out of 2,194 tested within 262 kb of the gene's TSS. After filtering for genes with ≥3 elements tested, the scRNA-seq dataset contains 404 positives of 19,104 tested within 262 kb of the gene's TSS. These numbers shrunk further on a per-analysis basis after requiring that each enhancer–gene pair is within the input window of the current set of models evaluated in a given benchmark.

For both Enformer and Borzoi, we scored putative enhancers using input gradient analysis. For Enformer, we computed the gradient of the K562 CAGE prediction in the two 128 bp bins centered at the gene's

TSS. Computing the gradient using three bins (as in the original paper) resulted in marginally worse performance. The gradient score statistic was averaged for genes with multiple TSSs, which performed better than taking either the max or sum. For Borzoi, we computed the gradient of the K562 RNA-seq prediction for all bins overlapping the gene's exons in GENCODE (v.41). For each nucleotide, we took the absolute value of the reference nucleotide gradient. For each regulatory element, we computed a weighted average of the nucleotide scores using Gaussian weights (s.d. = 300), centered at the element's midpoint. This approach improved performance for both Enformer and Borzoi compared to a simpler strategy of averaging the absolute-valued gradients in a 2 kb window centered on the enhancer. To calibrate scores across genes with different expression levels, we normalized the scores by the mean nucleotide score across the entire region.

The analysis was repeated using an in-silico perturbation approach instead of input gradients. The putative enhancers were independently dinucleotide-shuffled with a 2 kb window. Using Borzoi, each shuffle was repeated 16 times for both forward and reverse orientations and for all four model replicates (128 times total). For Enformer, each shuffle was repeated 64 times in forward and reverse orientations. For Borzoi, the absolute-valued percent change in exon-aggregated RNA-seq coverage was used as the final statistic. For Enformer, the absolute-valued percent change in aggregated CAGE signal was used (within two or three output bins). Smaller or larger window sizes only marginally affected the results (as shown in Supplementary Fig. 7a).

Saturation mutagenesis MPRA benchmark

The saturation mutagenesis experiment from a previous publication[106] was used to compare Borzoi to Enformer on non-QTL variation data. Each measured variant was induced in the hg38 reference sequence and centered on when making predictions. For DNase, CAGE and histone ChIP tracks, variant effects were estimated as the log fold change in coverage within a 4 kb window, whereas scores for RNA-seq were computed as the log fold change in exon-aggregated coverage. The final predictions were calculated as an unweighted average of (potentially a subset of) the different assays' scores. Using a narrow 512 bp window for aggregation as in the Enformer paper[13] resulted in worse concordance with measured effects for some promoters and better concordance for other promoters. We settled on the wider 4 kb window as it led to better performance on the majority of promoters. Only promoters and enhancers with better performance using cell-type-matched outputs in the Enformer paper were included to simplify the benchmark. The same cell-type mappings were used except for promoters F9 (K562 instead of HepG2), LDLR (adrenal gland instead of HepG2) and HNF4A, MSMB, TERT and MYC (adrenal gland RNA-seq instead of HEK293T RNA-seq). These changes led to better performance for Borzoi and were reasonable choices with respect to the target genes' expression patterns. The same changes were made to Enformer's mappings if they resulted in an improvement.

Codon stability comparison

Prior work has demonstrated strong relationships between codon usage and mRNA half-life[23,78]. We constructed a Borzoi codon statistic to compare to those previously measured. For the Gasperini[58] scRNA-seq enhancer screen, we computed input gradients for a set of 4,778 genes for K562 gene expression. We made use of these gradients here to quantify codon contributions to expression. For each reference codon in these genes, we used the gradients to approximate the predicted effect of changing it to all alternative codons with a single base-pair mutation. We used least squares regression to fit a coefficient for each codon on this set of possible codon mutations and effects. Finally, we compared these coefficients to codon stability coefficients computed in previous work[78] as the Pearson correlation between codon frequency and mRNA half-life in three mammalian cell lines: HeLA, mouse embryonic stem cells and CHO cells[78].

Statistics and reproducibility

All data from ENCODE, FANTOM5 and CATlas matching the target assay types and passing quality metrics, as established by each respective source, were included in the training data. Only a subset of GTEx RNA-seq samples were included; namely, the most representative samples as determined by expression profile clustering (details above). No statistical methods were used to predetermine sample size. No data were otherwise excluded from analyses.

Computational experiments and statistical tests were conducted as indicated in relevant sections. The experiments were not randomized and the authors were not blinded to outcome assessment. Confidence intervals of performance metrics were obtained by bootstrapping or permutation tests.

Reporting summary

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

Data availability

The processed Borzoi training data (including one-hot coded sequences and coverage tracks) are available for download at 'gs://borzoi-paper/data' (Google Cloud Storage). Gene annotations were obtained from https://www.gencodegenes.org (v.41). Human variation data were obtained from gnomAD (v.3.1) (https://gnomad.broadinstitute.org). Annotations of polyadenylation sites were obtained from PolyADB (v.3.2) (https://exon.apps.wistar.org/polya_db/v3) and PolyASite (v.2.0) (https://polyasite.unibas.ch/atlas). CRISPRi data were obtained from Nasser et al. (2021)[65] and from GEO accession GSE120861 for the Gasperini et al. (2019)[58] data. DNase-seq, ChIP–seq and RNA-seq data were downloaded and processed from ENCODE (https://www.encodeproject.org); see the ENCODE portal for details and statistics on the RNA-seq experiments. Processed RNA-seq samples for GTEx individuals were downloaded from recount3 (https://rna.recount.bio). CAGE data were downloaded from FANTOM5 (https://fantom.gsc.riken.jp/5). ATAC–seq data were downloaded from CATlas (http://catlas.org/catlas_hub). All experiments used for training, including their unique identifiers, are enumerated for human samples at https://storage.googleapis.com/seqnn-share/borzoi/hg38/targets.txt and for mouse samples at https://storage.googleapis.com/seqnn-share/borzoi/mm10/targets.txt. Fine-mapped eQTLs were obtained from the supplementary material of Wang et al. (2021)[1]. Fine-mapped eQTL credible sets and other QTLs (sQTLs and paQTLs) were downloaded from the eQTL catalog (https://www.ebi.ac.uk/eqtl). The positive (fine-mapped causal) and negative eQTL, sQTL and paQTL sets used in this study are available at 'gs://borzoi-paper/qtl/' (Google Cloud Storage). TRIP data were downloaded from the supplementary material of Leemans et al. (2019)[68].

Code availability

The code repository for training RNA-seq deep learning models, including example code to use the model as well as scripts for variant scoring, is available under the Apache 2.0 open source license at https://github.com/calico/borzoi[107]. Pre-trained Borzoi model weights are available through GitHub. A separate GitHub repository (also licensed under Apache 2.0 open source) contains code relevant to the analyses and results presented in the manuscript, located at https://github.com/calico/borzoi-paper[108].

References

97. Noguchi, S. et al. FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
98. Lizio, M. et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).
99. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

100. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001.e19 (2021).
101. Li, Y. E. et al. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
102. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
103. Majdandzic, A., Rajesh, C. & Koo, P. K. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biol.* **24**, 109 (2023).
104. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
105. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
106. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
107. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. calico/borzoi: Borzoi release (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.13984114 (2024).
108. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. and Kelley, D. R. calico/borzoi-paper: Borzoi paper release (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.13984155 (2024).

Acknowledgements

This work was funded by Calico Life Sciences. The funder had no role in study design, data collection or analysis. Publication of the manuscript was approved after an internal scientific review process. We thank A. Korsakova, X. Huang, M. Yilmaz and J. Xu for helpful discussions and valuable feedback. We thank L. Ruiz for help with code repositories and cloud computing infrastructures. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v9.p2.

Author contributions

D.R.K. conceptualized the project. J.L., D.R.K., D.S., H.Y. and V.A. analyzed the data. J.L., D.R.K., D.S., H.Y. and V.A. wrote the manuscript.

Competing interests

D.R.K., J.L., D.S. and H.Y. are employees of Calico Life Sciences. V.A. is an employee of Sanofi Pasteur but was involved in this work independently of Sanofi.

Additional information

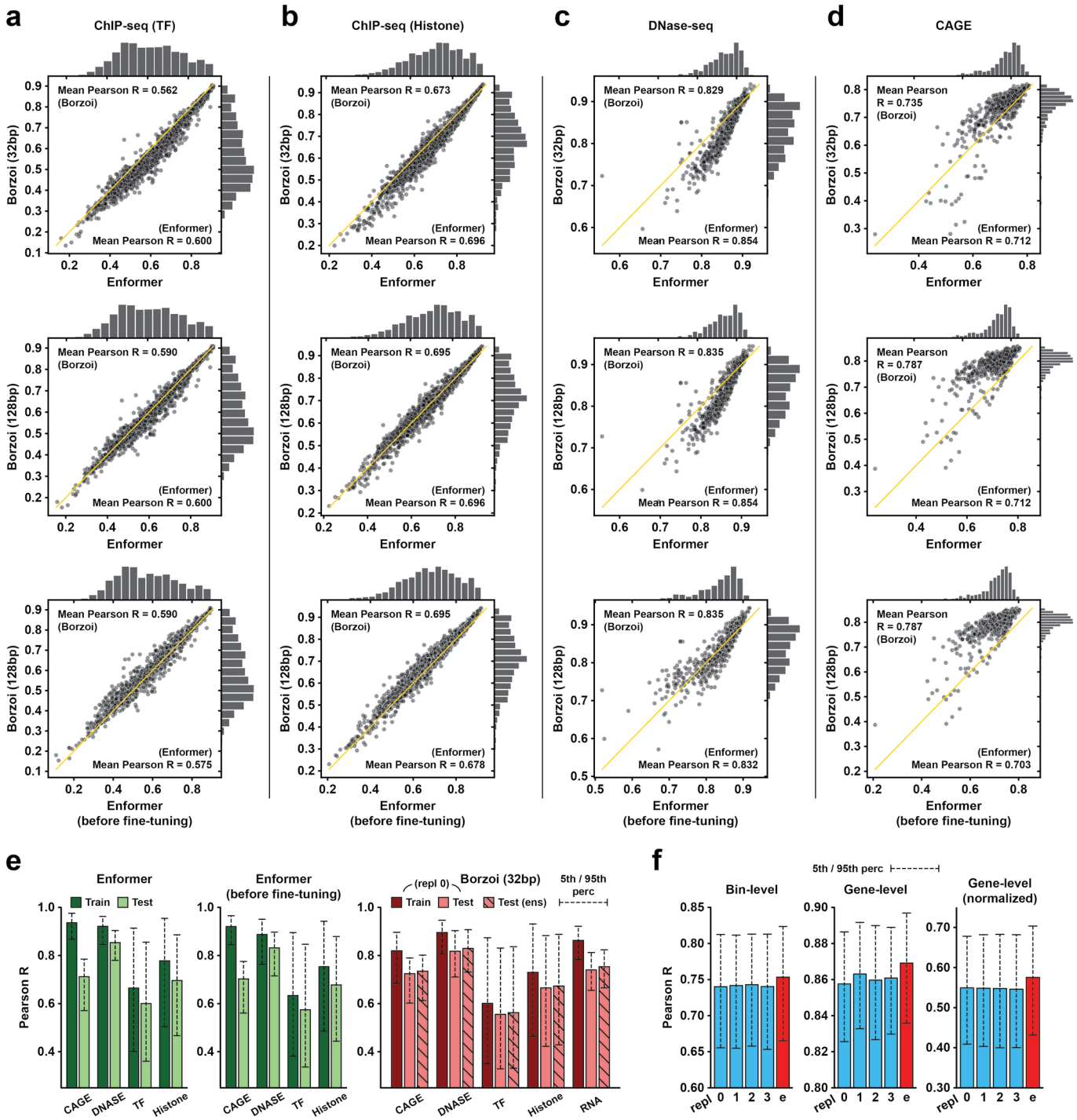
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-02053-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-02053-6.

Correspondence and requests for materials should be addressed to Johannes Linder or David R. Kelley.

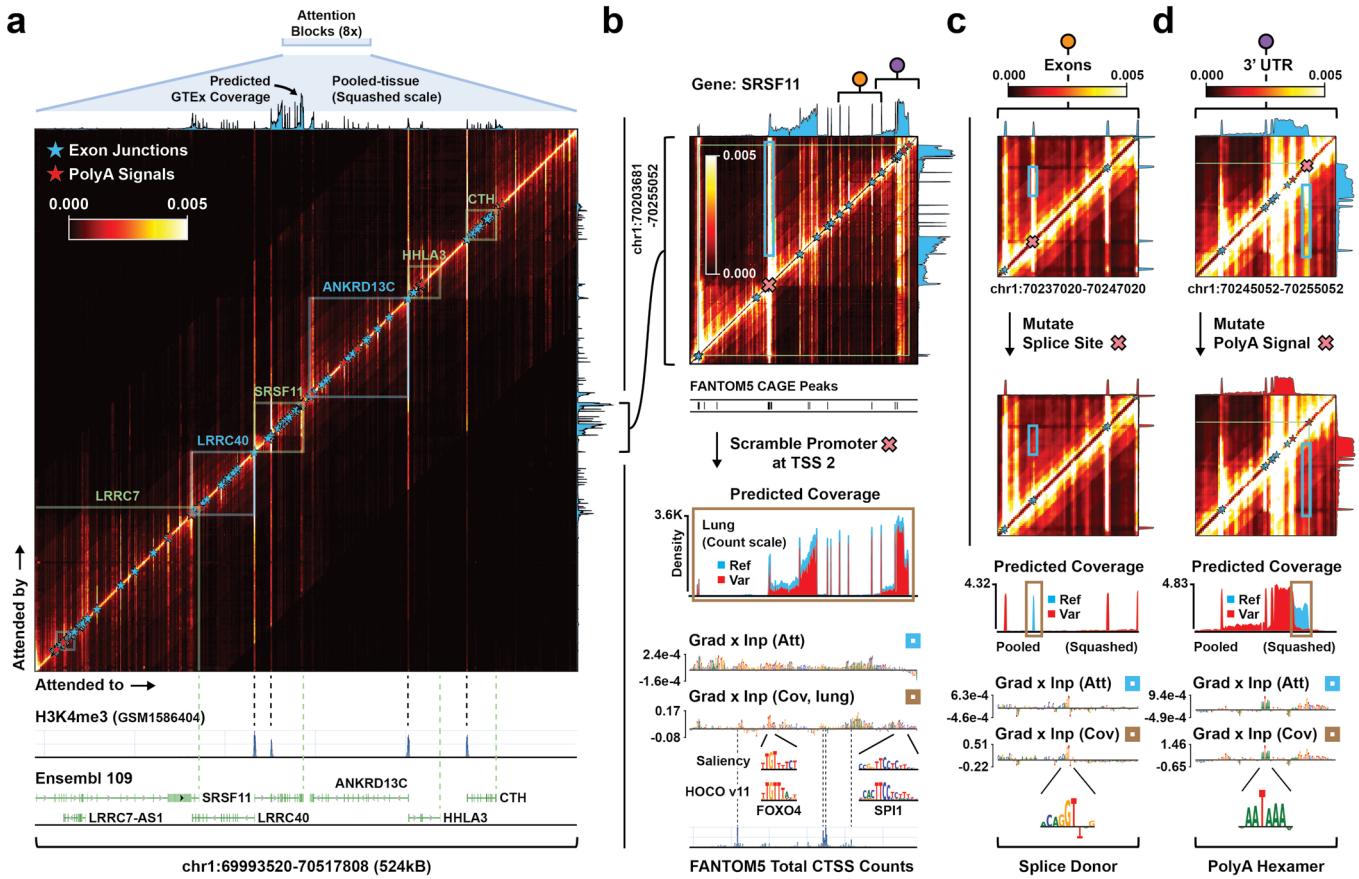
Peer review information *Nature Genetics* thanks Sandra Cooper, Julien Gagneur and Matthew Weirauch for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-02053-6



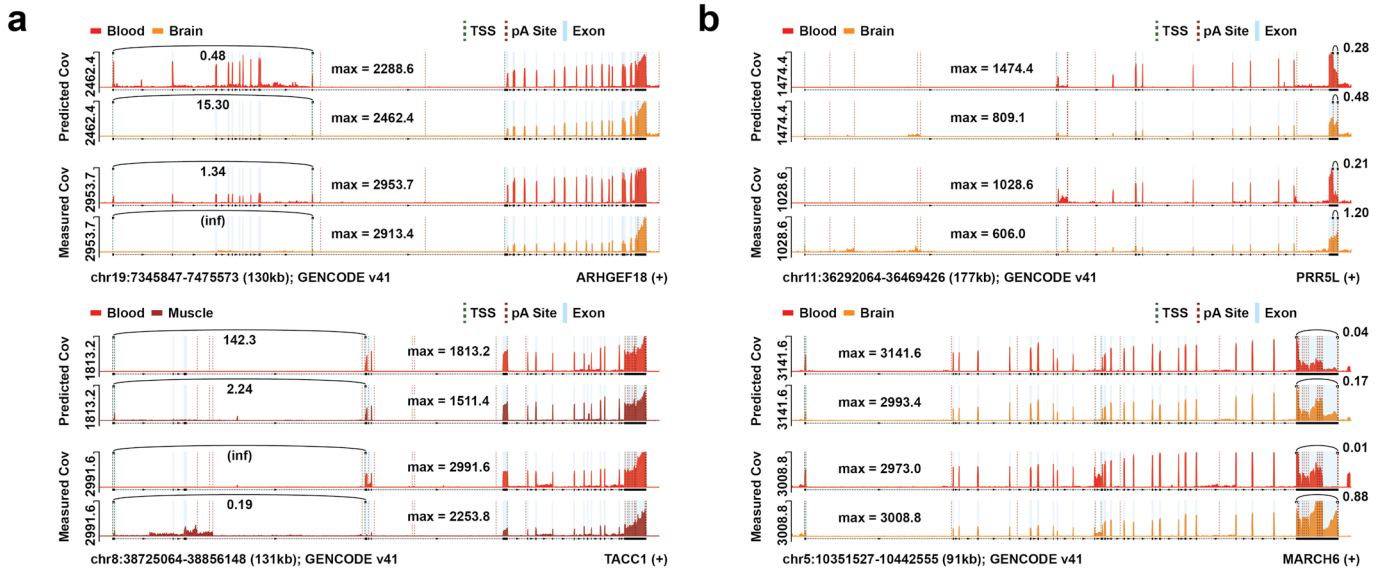
Extended Data Fig. 1 | Additional test set evaluations and comparisons to Enformer. (**a**) - (**d**) Performance comparison between Borzoi and Enformer on held-out genomic (human) sequences when tasked with inferring (**a**) ChIP-seq TF, (**b**) ChIP-seq Histone, (**c**) DNase-seq, or (**d**) CAGE coverage. Each dot in the scatter plot represents the Pearson correlation between predicted and observed bin-level coverage values. The mean Pearson R of each model is annotated in the plot. The top row of plots displays performance when comparing Borzoi's predictions to measurements at the original 32bp resolution, while the middle row of plots shows the result of aggregating the predicted and measured bins to 128bp resolution before computing Pearson R. Additionally, the published version of Enformer was fine-tuned on human assays without mouse data as a final step, while Borzoi was not. The bottom row of plots compares Borzoi at 128bp resolution to Enformer before fine-tuning, which is comparable to how Borzoi was trained. (**e**) Distribution of Pearson correlation metrics, for Enformer (green) or Borzoi (red), when comparing predicted to observed bin-level coverage values on held-out data. Bars with darker/lighter shades correspond to train/test performances respectively. Each bar displays the mean correlation across experiments, and the intervals mark the 5th and 95th percentiles. Only tracks shared by both models are included for CAGE (n = 638), DNase (n = 546), TF- (n = 1,203) and Histone (n = 1,634) ChIP. A total of 1,543 tracks were included for Borzoi's RNA-seq bar. For Enformer, the results of the published (fine-tuned) version of the model are shown alongside the performance metrics before fine-tuning. For Borzoi, performance metrics for a single replicate ('repl 0') and the ensemble ('ens') are shown at 32bp resolution. (**f**) Distribution of Pearson correlation metrics for RNA-seq tracks when making bin-level, gene-level, or quantile-normalized and mean-subtracted (gene-level) predictions on the held-out test set. Results are shown for each individual Borzoi replicate (blue; '0'-'3') or the full ensemble (red; 'e'). Each bar displays the mean correlation, and the intervals mark the 5th and 95th percentiles. Bin- / gene-level bars are estimated from 1,543 and 955 distinct (stranded) tracks / experiments respectively.

https://doi.org/10.1038/s41588-024-02053-6



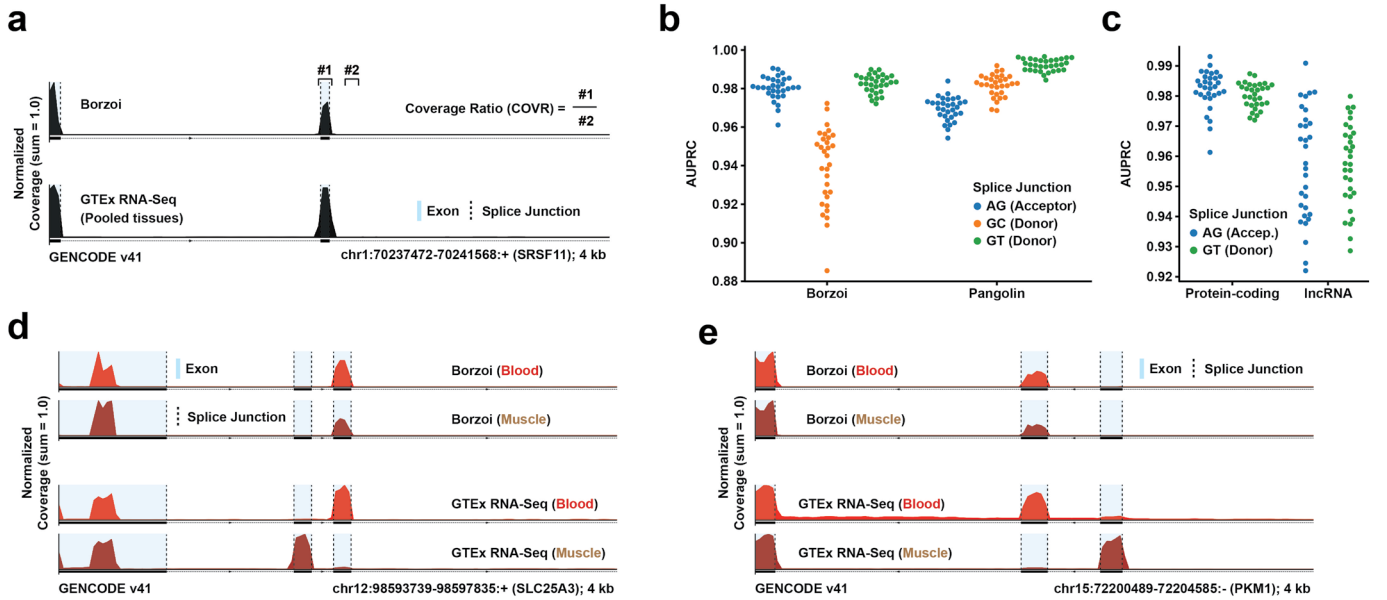
Extended Data Fig. 2 | Attention matrix visualization. (**a**) Attention weight matrix averaged across all 8 heads of the final transformer layers, shown for example region chr1:69993520-70517808. Average predicted RNA-seq coverage for 89 GTEx samples is shown above the attention heatmap. Ensembl gene models and H3K4me3 tracks are shown below. Exon junctions (blue stars) and polyadenylation signals (red stars) are annotated in the plot. Genes and their bounding boxes are also annotated (green = forward strand, blue = reverse strand). (**b**)-(**d**) Enlarged view of the attention weight matrix for the SRSF11 gene, highlighting (**b**) a promoter region (and alternative TSS), (**c**) several introns and exons, and (d) the 3' UTR. Gradient saliencies ('Grad x Inp') of either the output coverage tracks ('Cov', within the brown boxes) or the attention matrix ('Att', within the blue boxes) are displayed below each vignette. The regions highlighted in the saliency logos are either dinucleotide-shuffled (promoter) or mutated (exon and 3' UTR) and the resulting coverage predictions are depicted above each logo (blue = reference, red = variant). The altered attention matrices due to the mutations are also shown in (**c**) and (**d**). Exon junctions (blue stars) and polyadenylation signals (red stars) are annotated in the plots. FANTOM5 CAGE peaks / counts are annotated below the heatmap and sequence logo in (**b**).

https://doi.org/10.1038/s41588-024-02053-6



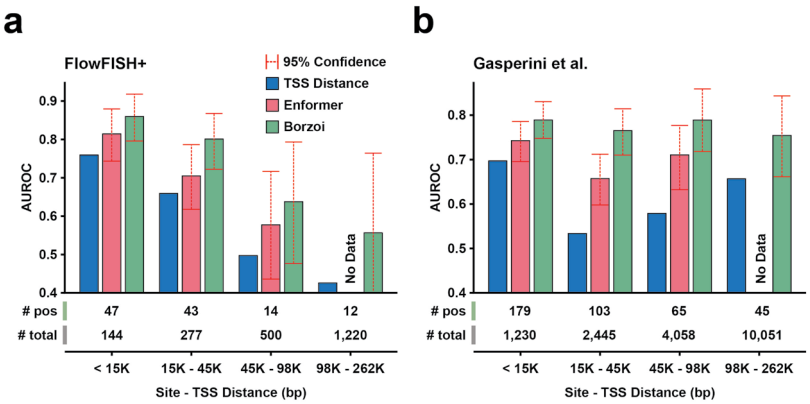
Extended Data Fig. 3 | Example coverage predictions for genes that exhibit differential isoform usage. (**a**) Predicted and measured RNA-seq coverage patterns for pairs of GTEx tissues exhibiting differential TSS usage, for two example genes: ARHGEF18 (GTEx whole blood vs brain) and TACC1 (whole blood vs muscle). Exon-overlapping bins are shaded light-blue, TSSs and pA sites are drawn as dashed lines, and 'max' refers to the maximum bin value in the exonic regions (in **a** - **b**). TSS usage is estimated as a coverage ratio between bins overlapping each alternative start site (the ratio is annotated above each track). The examples were selected by searching for test genes with largest measured fold change in TSS usage between each pair of tissues, where measured usage was estimated from FANTOM5 TSS counts (Methods). (**b**) Predicted and measured coverage in GTEx tissues whole blood and brain for two test genes with increased coverage over the distal polyadenylation signal in brain compared to blood: PRR5L and MARCH6. Distal usage is estimated as a coverage ratio between bins overlapping the distal site relative to the proximal site (the ratio is annotated above each track). The genes were chosen from the set of genes with maximal fold change of distal-to-proximal coverage ratio in brain. Their brain-specific distal polyadenylation bias were verified in bulk 3'-sequencing data obtained from the database PolyASite 2.0 (Methods).

https://doi.org/10.1038/s41588-024-02053-6



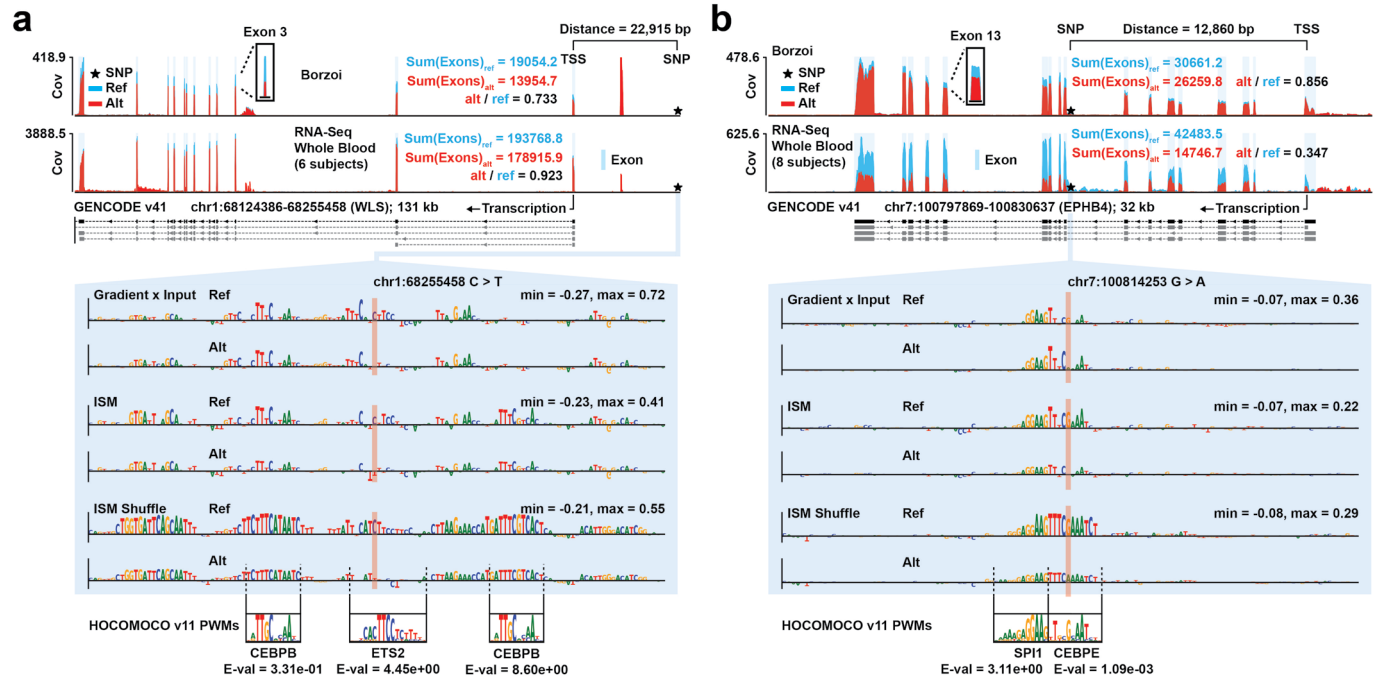
Extended Data Fig. 4 | Splice site detection and alternative splicing examples. (**a**) Predicted and measured RNA-seq coverage across an exon in the SRSF11 gene (GTEx pooled-tissue), centered on its splice donor. Each coverage track is independently normalized. Calculation of exon-to-intron coverage statistics (COVR) is illustrated in the figure. Exon-overlapping bins are shaded light-blue, and junctions are drawn with dashed lines. (**b**) Comparison between Borzoi's predicted exon-to-intron coverage ratio statistic and Pangolin's predicted splice usage when classifying annotated splice donors/acceptors from matched negatives in the reference genome. Average precision (AUPRC) is displayed separately for each type of splice junction (AG - Acceptor, GT/GC - Donor) and each dot corresponds to a GTEx tissue. (**c**) Average precision when using Borzoi to classify annotated splice junctions, calculated separately for protein-coding and long non-coding RNA. (**d**) Predicted and measured RNA-seq coverage across an alternative splicing event in the SLC25A3 gene for GTEx tissues whole blood and muscle (both predictions and measurements are pooled across 3 tissue-specific samples for blood and muscle). Each coverage track is independently normalized. Exon-overlapping bins are shaded light-blue, and junctions are drawn with dashed lines (in **d** - **e**). (**e**) Predicted and measured blood- and muscle RNA-seq coverage across an alternative splicing event in the PKM1 gene (coverage tracks are pooled across 3 samples per tissue).

https://doi.org/10.1038/s41588-024-02053-6



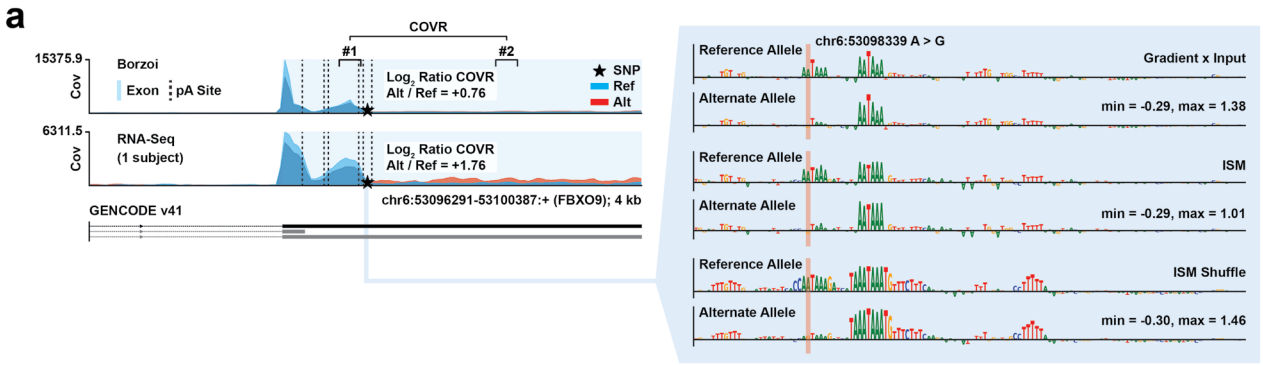
Extended Data Fig. 5 | Prioritizing gene-enhancer pairs from CRISPR perturbation data. (**a**) Area under the receiver operating characteristic curve (AUROC) when using a statistic computed from the Borzoi or Enformer gradient saliencies to classify whether or not a given CRE locus regulates a target gene (measurements from Fulco et al., 2016, 2019 and Klann et al., 2017, and others)[60–65]. The baseline performance (blue bars) corresponds to using only TSS distance when performing the classification. The number of positives and total number of examples are displayed below each distance bin. The total number of examples are: (< 15K) n = 144, (15K - 45K) n = 277, (45K - 98K) n = 500, (98K - 262K) n = 1,220. 95% confidence intervals were estimated from 1,000-fold bootstrapping. (**b**) AUROCs when using the Borzoi or Enformer gradient scores to classify regulating / non-regulating CRE loci in the data from Gasperini et al. (2019)[58]. The total number of examples are: (< 15K) n = 1,230, (15K - 45K) n = 2,445, (45K - 98K) n = 4,058, (98K - 262K) n = 10,051. 95% confidence intervals were estimated from 1,000-fold bootstrapping.

https://doi.org/10.1038/s41588-024-02053-6



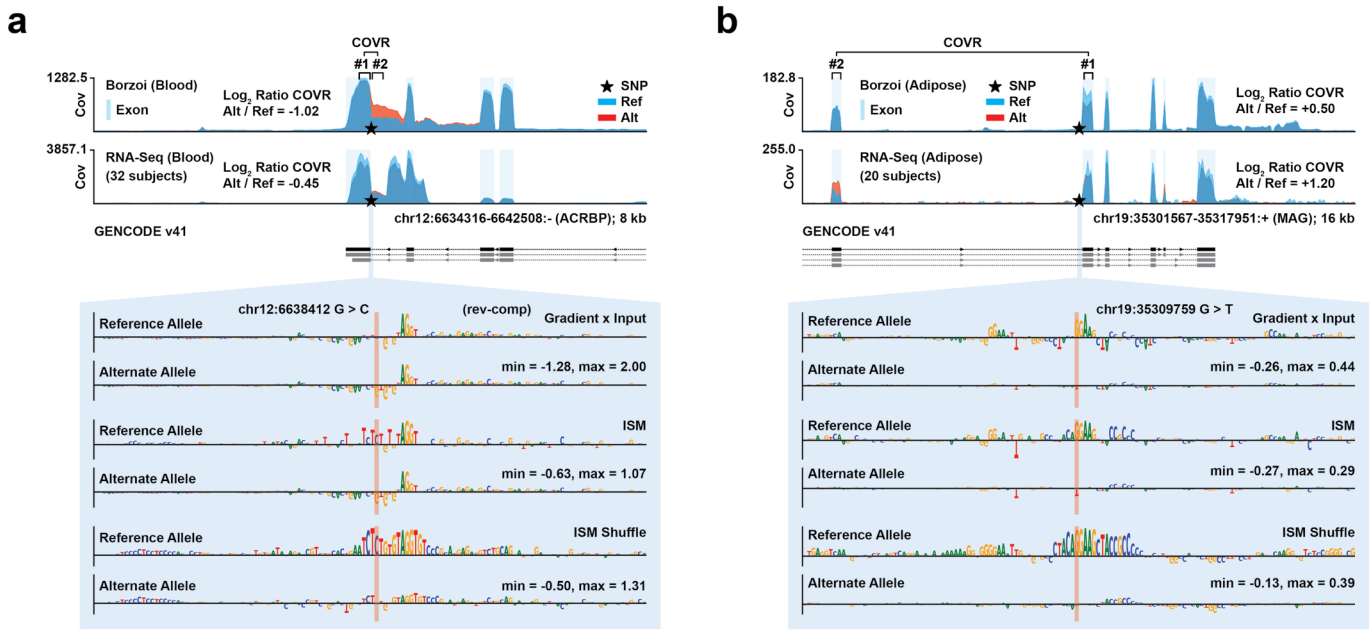
Extended Data Fig. 6 | Variant interpretation of fine-mapped eQTLs. (**a**) Predicted RNA-seq coverage (GTEx tissue whole blood) for the WLS gene when introducing variant rs72670481. Exon-overlapping bins are shaded light-blue. Exon-aggregated coverage for the alternate and reference alleles, and their ratio, are annotated in the coverage plot (in **a** - **b**). Measured coverage in 6 individuals with the reference allele and 6 hetero- or homozygous individuals for the alternative allele is displayed below the predictions, along with attribution scores computed in a local window centered on the variant. The attributions scores are calculated with respect to the log-sum of exon coverage for the WLS gene. The sequence logo y-axes are equally scaled for both the reference and alternate alleles (min / max annotated in the right corner). Likely motif hits are displayed below the sequence logos (the E-values represent the significance of the motif match, as computed by Tomtom). (**b**) Predicted RNA-seq coverage (GTEx tissue whole blood) for variant rs3890144, along with measured coverage in 8 individuals with the reference allele and in 8 individuals who are either hetero- or homozygous for the alternative allele. Attribution scores in a window centered on the variant, calculated with respect to EPHB4 coverage, displayed at the bottom (equally scaled y-axes for the reference and alternate allele; min / max annotated in the right corner). Likely motifs and Tomtom E-values shown below the sequence logos.

https://doi.org/10.1038/s41588-024-02053-6



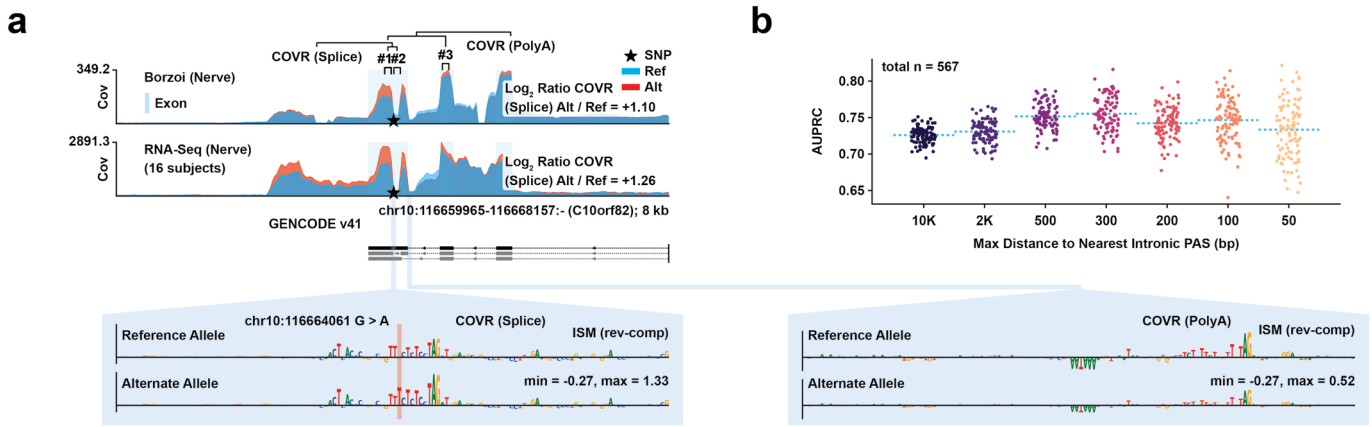
Extended Data Fig. 7 | Variant interpretation of fine-mapped paQTLs. (**a**) Predicted RNA-seq coverage (GTEx tissue-pooled) for variant rs74327114 in the FBXO9 gene, along with measured coverage in 1 individual with the reference allele and 1 heterozygous individual (averaged across two tissues each). Exon-overlapping bins are shaded light-blue, and pA sites are drawn with black dashed lines. The log ratio between the coverage ratio (COVR) statistics computed for the alternate and reference alleles is annotated in the plot. Attribution scores of the predicted COVR statistic, computed using three separate methods, are displayed to the right and indicate loss of an extra hexamer motif, resulting in moderate reduction in polyadenylation efficiency. The sequence logo y-axes are equally scaled for both the reference and alternate alleles (min / max annotated in the right corner).

https://doi.org/10.1038/s41588-024-02053-6



Extended Data Fig. 8 | Variant interpretation of fine-mapped sQTLs. (**a**) Predicted RNA-seq coverage (GTEx tissue whole blood) for variant rs1882553 using Borzoi, along with measured coverage in 32 individuals with the reference allele and 32 hetero- or homozygous individuals for the alternative allele (whole blood samples). Exon-overlapping bins are shaded light-blue. The log ratio between the coverage ratio (COVR) statistics computed for the alternate and reference alleles is annotated in the plot (in **a** - **b**). Attribution scores (bottom) are computed with respect to the predicted log ratio of exon-to-intron coverage, comparing three methods (equally scaled y-axes for the reference and alternate allele, displayed in reverse-complemented form; min / max annotated in the right corner). (**b**) Predicted RNA-seq coverage (GTEx tissue adipose) for variant rs10411704, along with measured coverage in 20 individuals with the reference allele and 20 hetero- or homozygous individuals for the alternative allele (adipose samples). Attribution scores of the predicted exon-to-exon log coverage ratio are displayed at the bottom (equally scaled y-axes for the reference and alternate allele; min / max annotated in the right corner).

https://doi.org/10.1038/s41588-024-02053-6



Extended Data Fig. 9 | Variant interpretation of fine-mapped intronic paQTLs. (**a**) Predicted RNA-seq coverage (GTEx tissue nerve) for variant rs3830026 and measured coverage in 16 individuals with the reference allele and 16 individuals who are hetero- or homozygous for the alternative allele (the QTL is significant in nerve samples). Exon-overlapping bins are shaded light-blue. The computation of polyadenylation- and splice-centric coverage ratio (COVR) statistics is illustrated. The log ratio between the splice-centric COVR statistics computed for the alternate and reference alleles is annotated in the plot. Bottom: Attribution scores of the exon-to-intron coverage ratio (COVR Splice) and the exon-to-exon coverage ratio (COVR PolyA), plotted in reverse-complement with equally scaled y-axes for the reference and alternate allele (min / max annotated in the right corner of each sequence logo). (**b**) Average AUPRC when using Borzoi to classify fine-mapped intronic paQTLs (tissue-pooled). Each dot represents a permutation test and the dashed line shows the mean (n = 100; Methods).

nature portfolio

Corresponding author(s): Michael Johnson

Last updated by author(s): Nov 7, 2024

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	No software was used for data collection.
Data analysis	Packages or software used in this manuscript include the following; Michigan Imputation Server (version 1.6.3), Eagle (2.4), plink (2.0), bcftools (1.18), Cellranger (5.0.1), Cellranger-ARC (2.0.2), DropletUtils (1.22), Seurat (v4), DoubletFinder (2.0), edgeR (3.4.2), MatrixEQTL (2.3), lmerTest (3.1),qvalue (2.34),MungeSumstats (1.10.1),ieugwasr (1.0.1), coloc (5.2.3), MendelianRandomization (0.10). Scripts used for data analysis are available here. https://github.com/johnsonlab-ic/singlecell-MR

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

Raw snRNA-seq and genotype data from the Bryois_192 dataset is available as per their publication at the European Genome-Phenome Archive (EGA) under accession code EGAS00001006345 [7]. Raw snRNA-seq and genotype data from the MATTHEWS dataset is hosted at Synapse under accession code syn54083444. Newly generated raw snRNA-seq and associated genotype data (MRC_60 and Roche_PD) is available under accession code EGAS50000000687. Genotype data is considered personal data and is therefore under protected access by the host repository (EGA), where access is subject to the submission of an application delineating the scope of the project and the data required (full details on the portal). Applications are aimed to be reviewed within two weeks.

Processed single-cell expression counts for each dataset and the full set of eQTL summary statistics for both the full and control-only datasets are available at https://zenodo.org/records/13343729 . The full set of published GWAS summary statistics used for the colocalisation and MR analysis as well as links to the original publications are described in Supplementary Table 3.

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	Yes, sex is included as clinical covariates for the eQTL mapping.
Reporting on race, ethnicity, or other socially relevant groupings	Yes, there is a description of genetic ancestry (white european).
Population characteristics	Yes, clinical covariates have been included and contain age, sex, genotypic information, disease diagnosis as assessed by neuropathology. All samples were collected post-mortem.
Recruitment	This research was conduceted under the oversight of Imperial College Research ethics.
Ethics oversight	Imperial College Research Ethics reference: ICREC_14_2_11

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 performed snRNA-seq on all brain samples available to us, yielding N=409 individuals (391 post quality control). It is the largest dataset to date with almost equal sizes of controls (N = 183) and disease cases (N = 208), allowing to isolate disease-specific effects of feQTLs at cell-type specific level.
Data exclusions	Nuclei with less than 500 UMIs in 300 features, and more than 5% Mitochondrial content were excluded. Related individuals based on genotypic data were excluded, and individuals with less than 10 nuclei for a single cell-type were removed.
Replication	Replication was made by comparing eQTL discovery to a large-scale eQTL study performed in bulk brain tissue (N = 6,523). Between 72.9-88.7% of cell-type eQTLs (depending on cell type) replicated at FDR < 5%, of which 90.0–98.3 had the same direction of effect.
Randomization	Grouping was done based on diagnosis, determined by neuropathology. eQTL discovery was conducted on the full dataset (N = 391) and on the controls-only dataset (N = 183). No other grouping or selection was made.
Blinding	Blinding was not implemented to group allocation. However, in our case, the analysis focused on objective genetic and expression data, where researcher bias is unlikely to influence the outcome. Our study design required knowledge of group allocation to conduct separate analyses for controls and the full cohort, which is standard in genetic studies aiming to capture eQTLs across different biological conditions.

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

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.