nature genetics

Supplementary information

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

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

In the format provided by the authors and unedited

Contents

1 Supplementary Analyses

1.1 Detailed test set evaluation

Borzoi's test set performance was evaluated broadly across modalities and experiments in Fig. 1 (and Extended Data Fig. 1, Supplementary Fig. 1), including comparisons to measurements at the resolution of individual 32bp bins or at gene resolution after aggregating RNA-seq coverage across exon-overlapping bins. Additional evaluations were provided in Fig. 2 (and Supplementary Fig. 3) for GTEx RNA-seq tracks corresponding to 5 tissues, including gene expression fold changes and isoform-level metrics. The increase in performance due to replicate model ensembling was relatively small for bin-level predictions (average Pearson R increased from 0.74 to 0.75; Fig. 1c) or gene-level predictions (average Pearson R increased from 0.86 to 0.87; Fig. 1d). However, the gain was more noticeable when predicting gene-level fold changes across experiments (e.g. average Pearson R for quantile-normalized and mean-subtracted gene-level predictions increased from 0.55 to 0.58; Extended Data Fig. 1f), or when predicting isoform specificity (e.g. average Spearman R for predicting blood-specific TSS fold change increased by 0.05; Fig. 2d).

Pearson R dropped from 0.87 to 0.77 on average when evaluating gene-level RNA-seq predictions only on long transcripts (8kb - 65kb spliced length; Supplementary Fig. 1b) or on genes with many exons (20+ exons; Supplementary Fig. 1c). The average prediction error was highest for lowly expressed genes, but this may be due in part to noisier measurements (Supplementary Fig. 1d). Furthermore, the Pearson R between predicted and measured bin-level RNA coverage within the gene span was on average 0.70 across RNA-seq experiments when evaluated on all genes, but increased to 0.88 when evaluated only on the top 20% genes with highest variance (Supplementary Fig. 1e).

1.2 Attention matrix analysis and gene structure visualization

In order to predict RNA coverage, Borzoi must have internally learned a detailed representation of gene structure. We hypothesized that this learned structure is embedded in the higher-level attention layers (in their *attention maps*, which are dynamic matrices that specify from which input positions an output position will receive information; Methods). Extended Data Fig. 2a illustrates the average attention map of the two penultimate blocks for an example locus. The map highlights TSS regions, exon boundaries, and polyadenylation signals. Transcript elements tended to interact more within the same gene body, creating square-like structures that match annotated gene spans [1, 2]. The few positions attended to by many genes (seen as long vertical stripes) matched peaks from H3K4me3 data [3]. In Supplementary Fig. 2a, we extracted the average attention magnitude of regions overlapping annotated H3K4me3 peaks, exons, and

polyadenylation sites from held-out test loci and found a significant enrichment in intensity compared to matched background (p <1e-12).

Changes to global interactions in the attention map can be visualized through in-silico sequence perturbation. Furthermore, while the attention map has 128 bp resolution, we can elucidate base-resolution determinants by estimating the contribution that each nucleotide in the input sequence has on the total attention in a region of the map, e.g. based on *gradient saliency* ('Gradient x Input'; Methods). We can similarly compute the gradient of various statistics derived from the predicted RNA coverage to focus on determinants of tissue-specific expression or post-transcriptional regulation. For example, Extended Data Fig. 2b shows the attention map of the *SRSF11* gene. By dinucleotide-shuffling a 256 bp region overlapping the second TSS, the attention magnitude dropped (Supplementary Fig. 2b), as did predicted RNA coverage across the exons. The gradients highlighted TF binding motifs (e.g. putative FOXO- and SPI-like motifs) and promoter elements (e.g. SP motifs). Other attribution methods, e.g. *in-silico saturation mutagenesis* ('ISM') and window-shuffled ISM ('ISM Shuffle') displayed roughly equivalent interpretations (Supplementary Fig. 2c).

Attention gradients computed on smaller stripes within the gene body highlighted sequences relating to splicing and polyadenylation. Intronic regions attended to the closest neighboring splice junctions (Extended Data Fig. 2c, Supplementary Fig. 2d), suggesting that Borzoi is matching exon boundaries to infer low coverage across introns. When making hypothetical mutations in the splice donor of one of the constitutive exons, attention dropped along with RNA-seq coverage over the exon. Simultaneously, attention of the $3'$ acceptor extended to the next viable $5'$ donor site. Similarly, polyadenylation signals within SRSF11 attended to each other in order to adjust the coverage shape across the $3'$ UTR. After inducing hypothetical mutations in the distal site hexamer AATAAA, predicted RNA coverage dropped upstream of the distal site and increased at proximal sites (Extended Data Fig. 2d, Supplementary Fig. 2e). Proximal signals attend to the distal signal even for very long $3'$ UTRs (20 kb) (Supplementary Fig. 2f).

1.3 Splice site detection benchmark and alternative splicing evaluation

Splicing is a process by which the spliceosome identifies well-defined motifs near exon junctions and excises the intronic region [4, 5]. Splicing often produces alternative isoforms that may be cell- or tissue-specific [6]. Qualitatively, Borzoi predicts patterns of RNA-seq coverage in exons relative to introns well, indicating that the model has learned how sequence determines RNA splicing. As an example, Extended Data Fig. 4a shows predicted and measured coverage centered on an exon in the *SRSF11* gene. Researchers have previously trained neural networks to predict annotated splice sites as a function of the surrounding sequence [7, 8]. These methods are highly effective at classifying splice sites from non-splice sites and leave little room for improvement. Borzoi trains indirectly on splice sites, which are present in the quantitative RNA-seq coverage labels as 32 bp bins that transition between low intron and high exon coverage, rather than explicit nucleotide-resolution labels. Nevertheless, the predicted exon-to-intron coverage ratio surrounding a potential splice site discriminated between annotated splice acceptors and negatives with accuracies on par with the state-of-the-art model Pangolin [8], while performance was lower for canonical splice donors (GT; mean dAUPRC = -0.01) and worse for non-canonical donors (GC; mean dAUPRC = -0.04) (Extended Data Fig. 4b-c).

We also explored tissue-specific alternative splicing with Borzoi, but sequencing bias and other confounders made it challenging to definitively call such events from the predictions. Borzoi predictions tended to hedge similarly across tissues rather than capture switch-like examples of tissue-specific alternative splicing, as exemplified in Extended Data Fig. 4d-e for blood- and muscle-specific splicing patterns.

1.4 Comparison between gradient saliency and in-silico saturation mutagenesis

For variant interpretation and regulatory sequence analysis within targeted regions using Borzoi, attribution scores based on ISM are generally preferred; the scores are calculated by making syntactically valid perturbations to the sequence (point mutations) and there is a natural reference point (the original, unperturbed sequence). However, when discovering TF- or RBP-motifs *de novo*, which can potentially occur anywhere in the 524kb input sequence, it is computationally intractable to use an ISM-based approach and we instead resort to *gradient saliency*, even while recognizing the potential miscalibrations of this approximation.

In the GTEx motif analysis presented in Fig. 3, the results were based on gradient saliency scores computed for $5,000$ genes with respect to the predicted log expression in 5 distinct tissues. We assessed the quality of these attributions by comparing them to ISM-based scores and to the saliencies of other replicate models. Gradients computed from independent model replicates correlated reasonably well (Supplementary Fig. 4a-b; median Spearman R = 0.53 between replicate 0 and 1 for blood gradients in a $1,024$ bp window centered at the mode of importance), as did gradients and ISM maps (Supplementary Fig. 4c; Spearman R = 0.72 within a 192 bp window around the mode of importance). When evaluating replicate 0 on a subset of held-out test genes, its tissue-specific differential coverage predictions correlated strongly with measured TPM fold changes (Supplementary Fig. 4d; median Spearman R = 0.77). Test set correlation

was lower but still reasonably high when comparing the average difference in gradient saliency to TPM fold changes (median Spearman R $= 0.59$). All downstream analyses rely on the average gradient across the 4 model replicates.

1.5 Transcription Factor motif analysis in Esophagus and K562

When including gradient saliencies computed from GTEx 'Esophagus' RNA-seq in the unbiased motif analysis (related to Fig. 3), we identified only a few weakly tissue-specific MoDISco [9] motif clusters for esophagus (Supplementary Fig. 5c), such as TP53, COT2 and SRF. Furthermore, while we found GATA motif hits among the MoDISco clusters for GTEx 'Whole Blood', we were curious as to why GATA saliency scores were relatively small in comparison to other drivers of blood-specificity. Re-running the analysis using gradients derived from K562 RNA-seq tracks, we found GATA to be one of the highest-scoring clusters with significantly larger saliencies in K562 compared to whole blood (Supplementary Fig. 5d), suggesting that GATA is a significant driver of expression in the erythroleukemia cell line K562 but less so in a diverse mixture of blood. Directly comparing the measured expression quantiles of TF genes in GTEx whole blood and K562 experiments supported our inferences; we observed a considerable upregulation of SPI1, CEBPB and IRF motifs in whole blood alongside a decrease in GATA1/2 TPM quantiles relative to K562 (Supplementary Fig. 5e).

1.6 Additional gene-enhancer CRISPR perturbation evaluations

In Fig. 4, we calculated gradient saliency scores of exon-aggregated coverage predictions in K562 cells, derived scalar statistics from these nucleotide-level scores, and evaluated the ability to prioritize putative gene-enhancer interactions based on these statistics. The predictions were compared to ground truth labels of interacting / non-interacting pairs, as measured by large-scale CRISPR perturbation experiments (Fig. 4c-d) [10, 11, 12, 13, 14, 15].

We compared the gradient-based analysis to an in-silico perturbation approach where % change in RNA coverage across the exons of the target gene is recorded upon dinucleotide-shuffling the candidate enhancer. This allowed us to observe the total predicted change in expression due to enhancer disruption (which correlates with, but is not necessarily identical to, local gradient saliency) (Supplementary Fig. 6a). Manual inspection of false positive predictions suggest that loci near exons of the target gene, or other genes, may be a source of discrepancy between the model and measurements (Supplementary Fig. 6b). Furthermore, the in-silico perturbation method lets us examine the % change in coverage at individual exon-overlapping bins of a target gene. We found that the bin-level drop in coverage often exhibited a non-uniform pattern that is gene-specific (Supplementary Fig. 6c). While we observed a general trend where coverage dropped less in exonic bins furthest away from the enhancer, there were also many genes where this trend was reversed (Supplementary Fig. 6d-e). Overall, any global distance bias was minimal; averaged across all positive examples from Gasperini et al., the drop in coverage at the 20% most distal exonic bins was ~ 0.95x of the drop in coverage at the 20% proximal-most bins. Finally, the choice of shuffle window size only marginally affected overall accuracy (Supplementary Fig. 7a).

To understand what drives the improved performance in prioritizing enhancers with Borzoi, we performed a number of ablation studies. First, we tested how various parameter choices affected performance when aggregating the gradient- or shuffle-based scores, for both Enformer and Borzoi (Supplementary Fig. 7b). Briefly, we found that a fraction of Borzoi's improved accuracy was due to model ensembling, and that gaussian smoothing often was superior to averaging the gradient score in a discrete window, for both models. Next, we re-trained additional instances of Borzoi where various subsets of the original training data were removed, or where the architecture was changed (Supplementary Fig. 7c). Consistently, we observed diminished performance if only the RNA-seq data was used for training. We generally observed improved performance on the data from Gasperini et al. when training on mouse in addition to human assays, although human-only and K562-only models were performant on the data from Nasser et al. Including DNase-/ATAC-seq training data improved performance in both human-only and multi-species settings. However, including ChIP-seq only boosted performance of the human- or K562-specific models. Finally, we note that removing the U-net component and modeling coverage at 128 bp resolution does not degrade performance on this benchmark.

1.7 Detailed fine-mapped eQTL evaluations

Borzoi and Enformer were evaluated on their ability to classify fine-mapped causal GTEx eQTLs from matched negatives in Fig. 5b-c, based on an auxiliary classifier trained on summary statistics computed from each model's coverage predictions. Specifically, we evaluated two different *gene-agnostic* statistics: 'L2' (the L2-norm of the difference between predicted reference- and variant coverage tracks, after applying a log transform), and 'logSUM' (the aggregated difference in log-transformed reference- and variant coverage predictions; this is also the original statistic used by Enformer [16]). Using Borzoi and the L2 score resulted in an average AUROC of 0.794 per tissue, compared to Enformer's average AUROC of 0.747 (using the logSUM score). Enformer with an L2 score partially closed the gap

(AUROC = 0.770). Limited to the same tracks as Enformer (ChIP, DNase, ATAC, and CAGE), Borzoi already achieved greater accuracy (AUROC = 0.784). Limited to only the RNA-seq tracks achieved a similar 0.785 AUROC.

In Fig. 5d, we compared Borzoi's and Enformer's variant effect statistics derived from tissue-matched RNA-seq or CAGE predictions to fine-mapped eQTL effect sizes. For each model, we calculated a *gene-agnostic* variant effect score using 'logSUM' and compared this to the beta coefficient estimated for that SNP averaged across all significant eGenes. Note that we performed a gene-agnostic analysis in order to compare against Enformer without the need for annotations, which would be different per model (exon- vs TSS) and might bias the results.

When scoring the fine-mapped eQTLs with a *gene-specific* statistic, calculated as the difference in log-sum of exon coverage for a specific target eGene, and comparing to the beta coefficient of said target gene, the results changed in subtle yet interesting ways, as exemplified for GTEx tissue 'Whole Blood' in Supplementary Fig. 9b. Compared to the gene-agnostic comparison presented in Fig. 5d, there were fewer predictions with incorrect direction of effect. However, there were also many more 0 predictions, since some target eGenes fall outside the cropped output window (for which we assign a default 0 score). If the examples with too-distant target genes were removed from the analysis instead of assigned 0 predicted effect, the Spearman R for blood increased to 0.492 (from 0.425).

We also assessed whether variant effect predictions were significantly better or worse when the SNP position fell in genomic reference sequence windows that were trained on or held out during training (note that the variant allele was never seen during training regardless). When separating eQTL SNPs based on trained-on or held-out genomic windows in the *gene-agnostic* analysis, the average Spearman R was marginally worse in sequence windows that were trained on (Supplementary Fig. 9c, left). However, in the *gene-specific* analysis, this trend was reversed (Supplementary Fig. 9c, right). Due to the relatively small number of held-out SNPs ($\sim 1/4$ of all fine-mapped causal eQTLs), we caution against over-interpreting these observations and conclude that if there are differences, they appear marginal.

1.8 SNP-eGene prioritization evaluation

We hypothesized that Borzoi could be used to prioritize the affected eGene from the set of genes surrounding an eQTL association. As a baseline, we predicted the nearest gene to be the eGene, where distance between an eQTL and a gene is defined as the average inverse distance to an annotated TSS across all variants within the fine-mapped credible set, weighted by each variant's posterior causal probability (PP). Using Borzoi, we aggregated L2 scores for all variants within the eQTL's credible set, weighted by their fine-mapping PP and ranked the genes by their maximum aggregated score. Borzoi prioritized the true eGene at a similar rate as TSS distance across GTEx tissues (mean accuracy 63.4% vs 62.7%, Supplementary Fig. 9e), suggesting that the model has not learned complex determinants of specificity.

1.9 Saturation mutagenesis MPRA evaluation

We compared Enformer and Borzoi on a subset of the Massively Parallel Reporter Assay (MPRA) data from Kircher et al. (2019) [17] (Supplementary Fig. 10a-d). We found that Borzoi surpassed Enformer's concordance with measured variant fold changes on the majority of promoters tested, and matched the performance on enhancer loci. Borzoi's ensembling strategy contributes significantly to the improvement, as some motifs with large measured effects are not detected by the predictions of the first model replicate but discernible with the ensemble (Supplementary Fig. 10d).

1.10 Additional 3' UTR attribution analyses

Similar to using input gradients as an efficient method for discovering TF motifs anywhere in the sequence (which would be intractable using ISM), we computed gradient saliencies of polyadenylation coverage ratios for genes from the Gasperini set [18] and clustered the gradients using TF-MoDISco [9] to discover $3'$ UTR motifs (Fig. 6b). We recapitulated the core $3'$-end processing motifs as top hits (CFIm, CstF, Elavl) as well as putative auxiliary RBPs (e.g. SRSF1). We generally did not find concordance between $3'$ UTR attributions from Borzoi and Saluki, which is a sequence-based model of mRNA half-life [19], nor to mutagenesis data [20]. Furthermore, while ISM-based attribution scores generally are preferred when interpreting individual loci (due to their intuitive similarity with variant effect predictions), we noticed extensive buffering effects when interpreting certain polyadenylation-regulatory elements with ISM, resulting in hidden features (Supplementary Fig. 11b). Consequently, we used window-shuffled ISM as the primary interpretation method for $3'$ UTRs (but provide gradient and ISM scores in supplementary figures). Note that this issue did not appear to be as extensive when interpreting regions around splice junctions (Supplementary Fig. 12a). Consequently, ISM was used as the primary attribution method for splicing variants.

1.11 Fine-mapped intronic paQTL analysis

Extended Data Fig. 9a highlights a fine-mapped causal intronic paQTL exhibiting competition between a set of splice sites and an intronic PAS. A G>A change (s3830026) creates a stronger polypyrimidine tract, increasing splicing efficiency and the rate at which the intron is excised. Intriguingly, RNA coverage increased non-uniformly across the transcript, with a larger increase in density for downstream exons and for the exon immediately upstream of the SNP. Attribution scores of the coverage ratio between the set of exons increasing in coverage and those with no discernible increase identified a salient intronic PAS. As splicing efficiency increases, this PAS is excised more often (hence its negative attribution), resulting in fewer truncated transcripts and increased downstream coverage. These interactions across regulatory functions complicate variant interpretation, highlighting the need for an accurate joint model of post-transcriptional regulation. Borzoi was performant at the task of classifying fine-mapped intronic paQTLs from negatives with a mean AUPRC of 0.725 (Extended Data Fig. 9b). Supplementary Fig. 13a highlights another intronic paQTL where alterations to either the splice- or polyadenylation code result in asymmetric changes to RNA shape.

To emphasize the utility of modeling RNA coverage at 32 bp resolution, we visualized variant rs3830026 using a re-trained model at 128 bp resolution with no U-net component (Supplementary Fig. 13b). At this coarser resolution, the coverage across short intronic regions were diluted with neighboring exon coverage, making interpretation harder.

2 Supplementary Note References

[1] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[2] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[3] Epigenomics C Roadmap, A Kundaje, W Meuleman, J Ernst, M Bilenky, A Yen, A Heravi-Moussavi, P Kheradpour, Z Zhang, J Wang, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

[4] Nuno André Faustino and Thomas A Cooper. Pre-mrna splicing and human disease. *Genes & development*, 17(4):419–437, 2003.

[5] Marina M Scotti and Maurice S Swanson. Rna mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32, 2016.

[6] Yan Wang, Jing Liu, BO Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, Jing Zhang, Qing-Hua Min, Wei-Ming Yang, et al. Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158, 2015.

[7] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

[8] Tony Zeng and Yang I Li. Predicting rna splicing from dna sequence using pangolin. *Genome biology*, 23(1):1–18, 2022.

[9] Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416*, 2018.

[10] Charles P Fulco, Joseph Nasser, Thouis R Jones, Glen Munson, Drew T Bergman, Vidya Subramanian, Sharon R Grossman, Rockwell Anyoha, Benjamin R Doughty, Tejal A Patwardhan, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. *Nature genetics*, 51(12):1664–1669, 2019.

[11] Charles P Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R Grossman, Elizabeth M Perez, Michael Kane, Brian Cleary, Eric S Lander, and Jesse M Engreitz. Systematic mapping of functional enhancer–promoter connections with crispr interference. *Science*, 354(6313):769–773, 2016.

[12] Tyler S Klann, Joshua B Black, Malathi Chellappan, Alexias Safi, Lingyun Song, Isaac B Hilton, Gregory E Crawford, Timothy E Reddy, and Charles A Gersbach. Crispr–cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature biotechnology*, 35(6):561–568, 2017.

[13] Shiqi Xie, Jialei Duan, Boxun Li, Pei Zhou, and Gary C Hon. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Molecular cell*, 66(2):285–299, 2017.

[14] Jialiang Huang, Kailong Li, Wenqing Cai, Xin Liu, Yuannyu Zhang, Stuart H Orkin, Jian Xu, and Guo-Cheng Yuan. Dissecting super-enhancer hierarchy based on chromatin interactions. *Nature communications*, 9(1):943, 2018.

[15] Joseph Nasser, Drew T Bergman, Charles P Fulco, Philine Guckelberger, Benjamin R Doughty, Tejal A Patwardhan, Thouis R Jones, Tung H Nguyen, Jacob C Ulirsch, Fritz Lekschas, et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, 2021.

[16] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

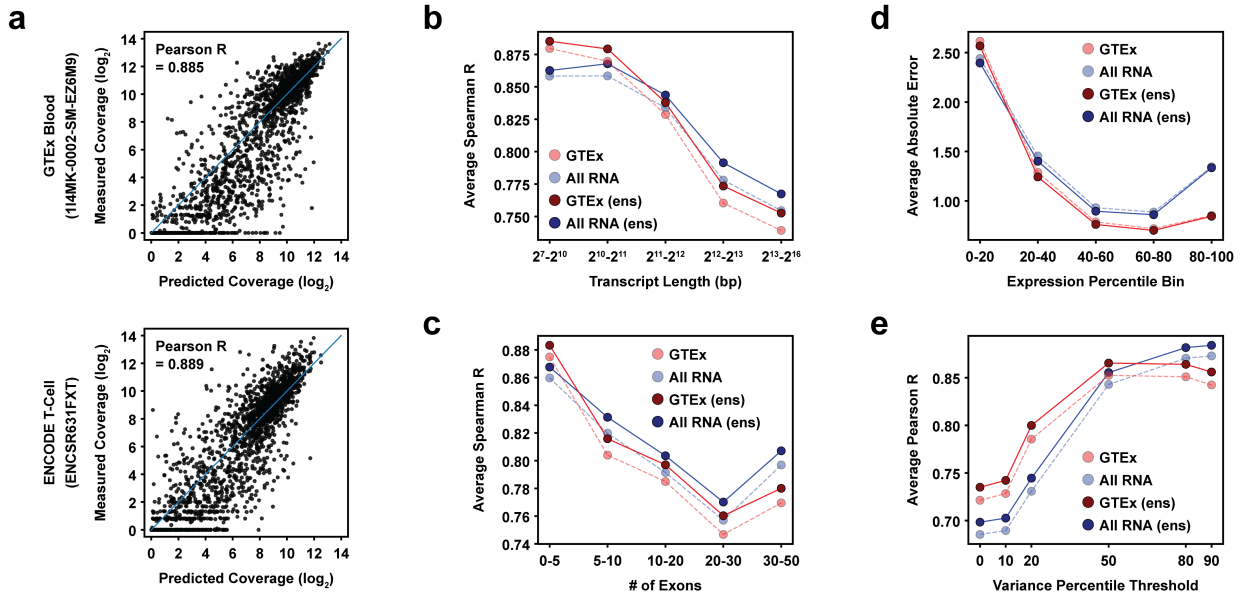
[17] Martin Kircher, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert JA Bell, Joseph F Costello, Jay Shendure, and Nadav Ahituv. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature communications*, 10(1):3583, 2019.

[18] Molly Gasperini, Andrew J Hill, José L McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S Noble, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(1):377–390, 2019.

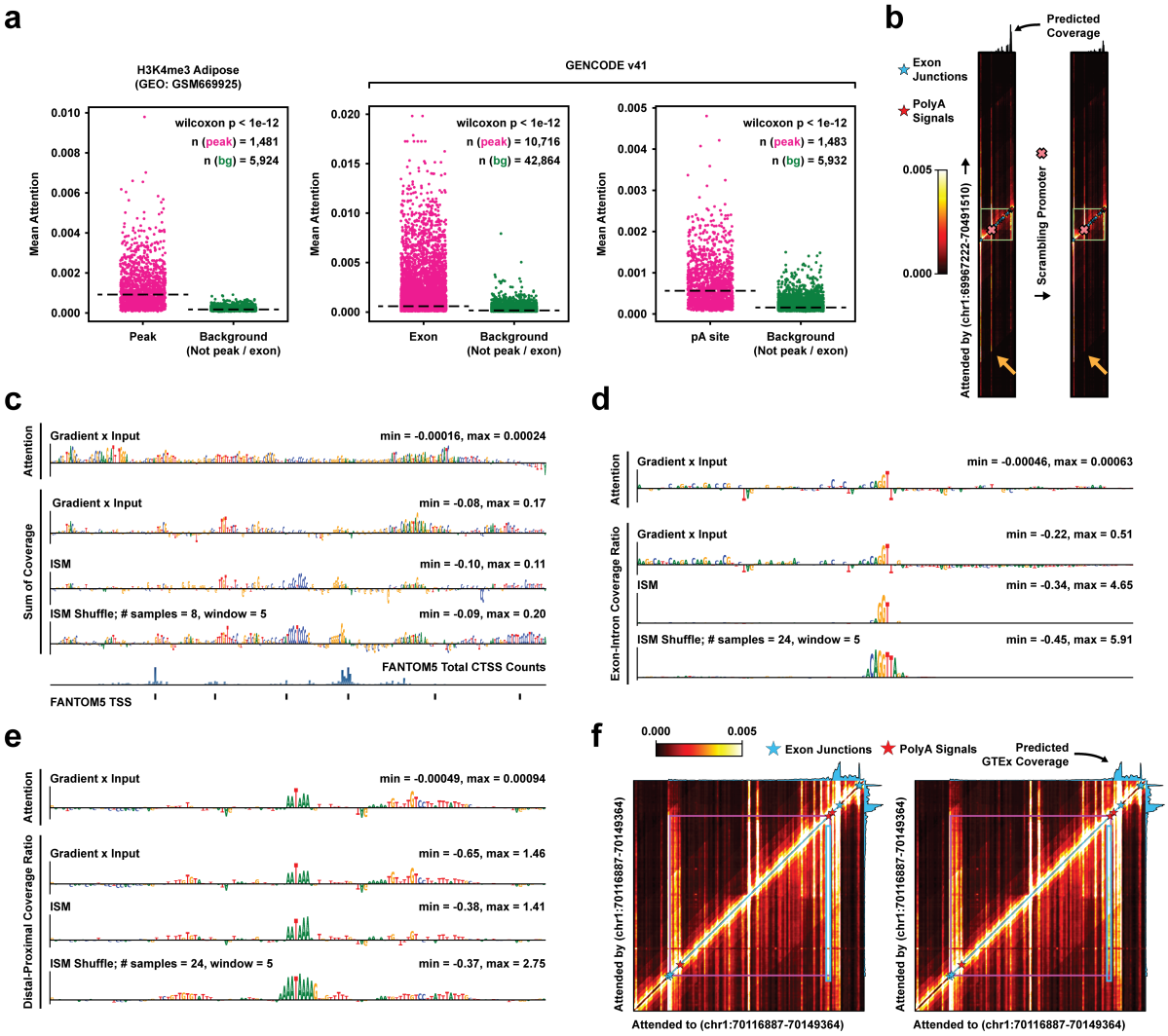
[19] Vikram Agarwal and David R Kelley. The genetic and biochemical determinants of mrna degradation rates in mammals. *Genome Biology*, 23(1):245, 2022.

[20] Wenxue Zhao, Joshua L Pollack, Denitza P Blagev, Noah Zaitlen, Michael T McManus, and David J Erle. Massively parallel functional annotation of 3' untranslated regions. *Nature biotechnology*, 32(4):387–391, 2014.

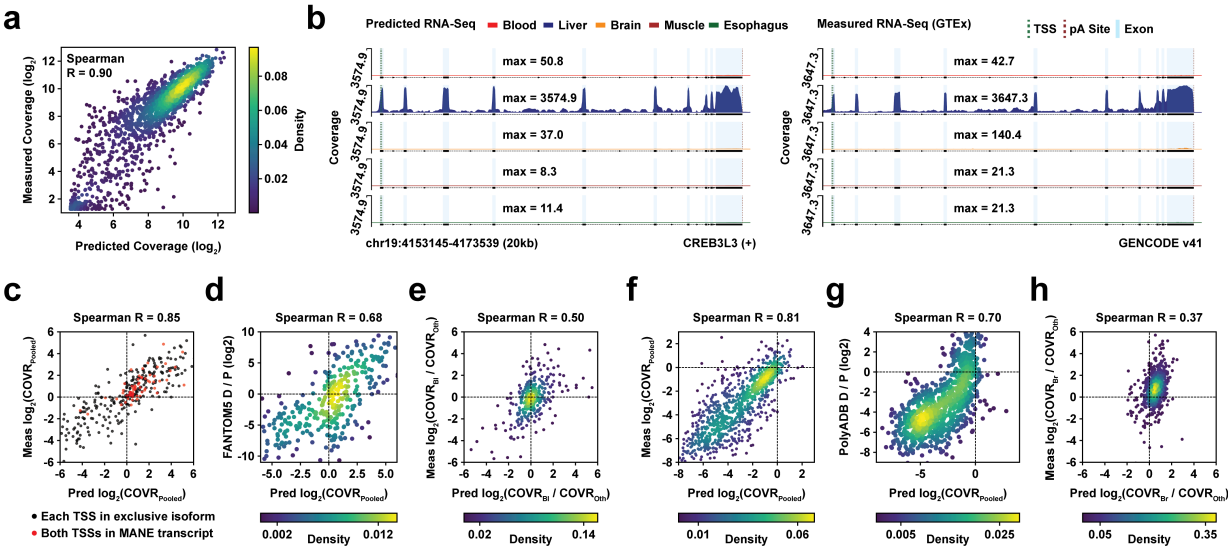
3 Supplementary Figures



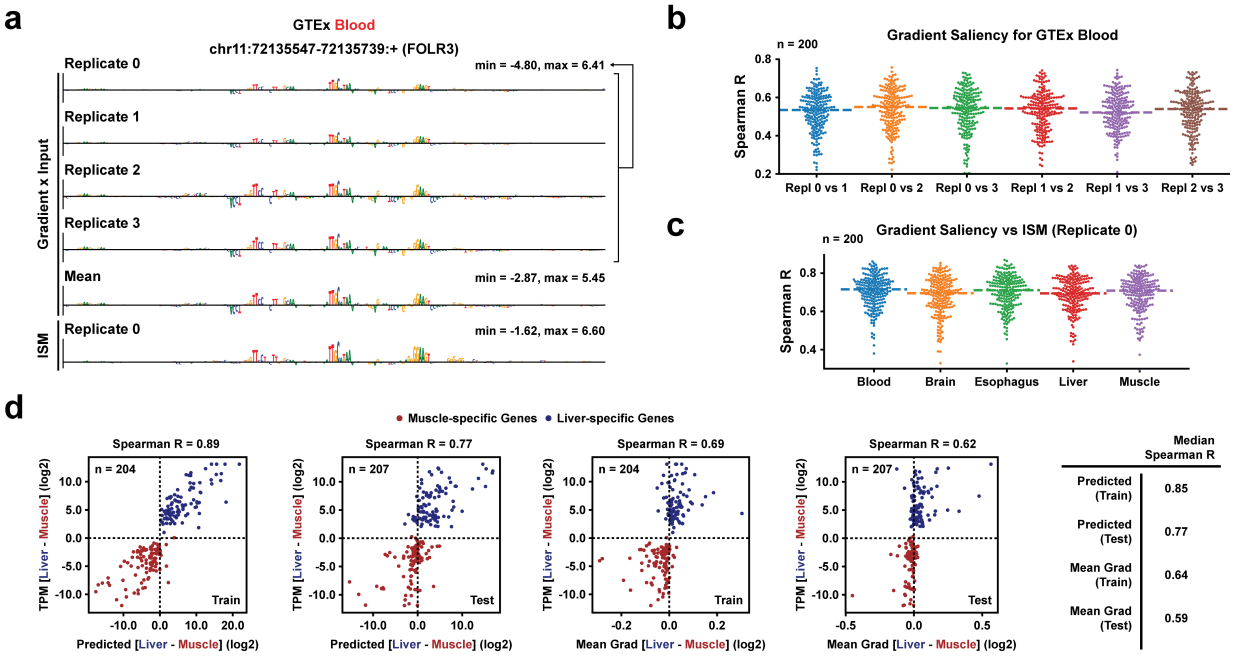
Supplementary Figure 1: **Additional gene-level test set performance comparisons.** **(a)** Comparison between predicted and measured aggregated exon coverage across held-out test genes, shown for two example assays and when averaging the predictions of 4 model replicates. Each data point represents a held-out gene. **(b)** Average Spearman correlation between predicted and measured aggregated exon coverage, calculated separately for held-out genes of different (spliced) transcript lengths. The metrics were averaged across all RNA-seq assays (blue), or GTEX tracks only (red). Solid lines = Spearman R when averaging the predictions of the 4 model replicates. Shaded dashed lines = Average Spearman R when using only a single model replicate to make predictions. **(c)** Average Spearman correlation across all RNA-seq assays (blue) or GTEx tracks only (red) when comparing predicted to measured aggregated exon coverage on held-out test genes, shown as a function of the total # of gene exons (for individual models and the ensemble). **(d)** Average absolute error between predicted and measured aggregated exon coverage across all RNA-seq assays (blue) or GTEx tracks only (red), shown as a function of measured aggregated coverage (percentile ranges). Results are shown for individual models and the ensemble. **(e)** Comparison between predicted and measured within-transcript coverage (bin-level), calculated at different lower thresholds of measured within-transcript variance (for example, Percentile Threshold $= 20$ corresponds to keeping the top 80% genes with highest within-transcript variance). Average Pearson R is computed across all RNA-seq assays (blue) or GTEx tracks only (red), using either one model (dashed lines) or the full ensemble (solid lines) to make predictions.



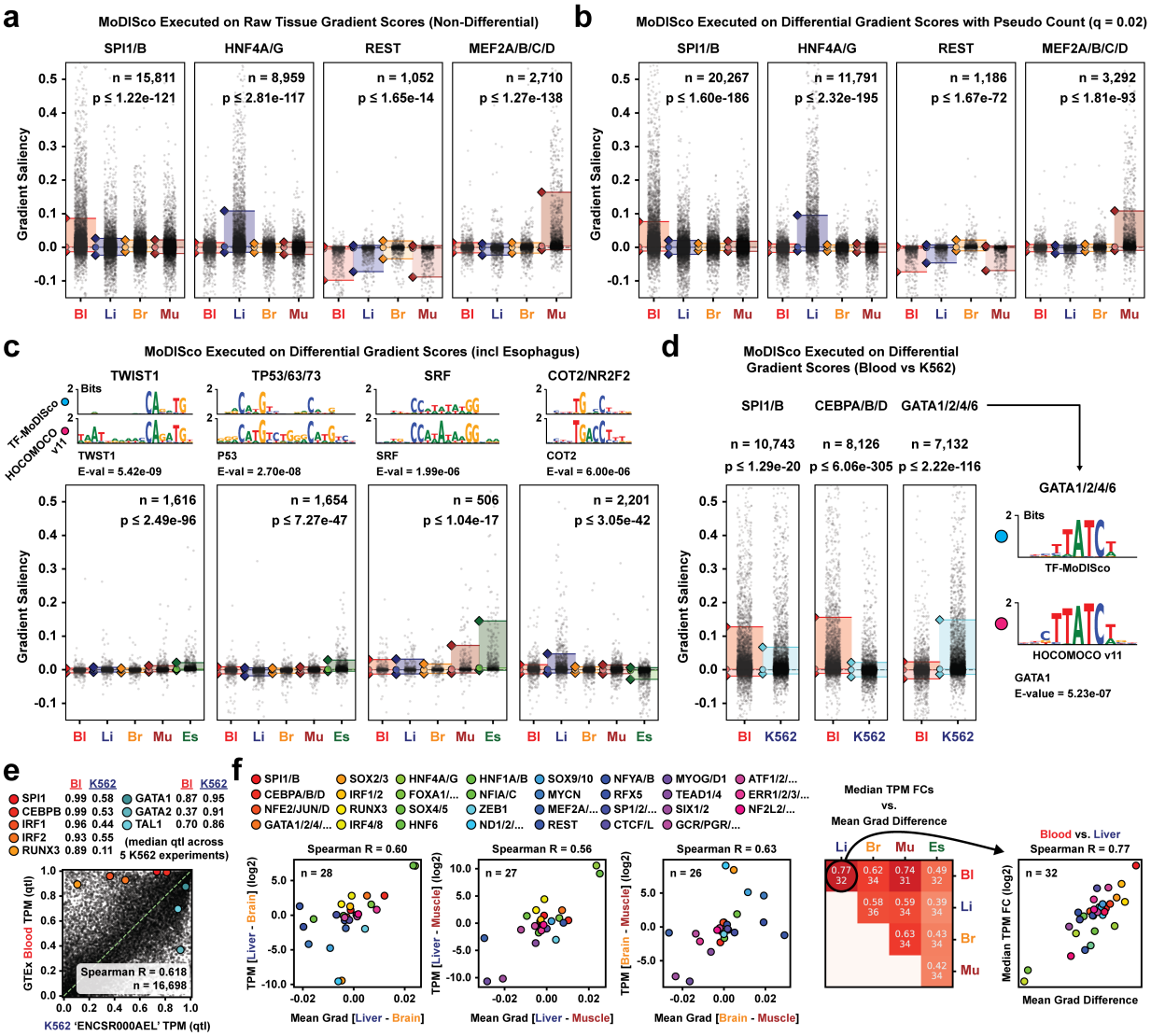
Supplementary Figure 2: **Detailed attention matrix evaluations.** **(a)** Mean attention within vertical stripes overlapping annotated H3K4me3 peaks (Adipose tissue), exonic regions, or in a 128-bp window around polyadenylation sites (as annotated in GENCODE v41). The coordinates of H3K4me3 peaks, exons and polyadenylation sites were extracted from the first 256 loci taken of the held-out test set. The attention weight matrices were generated by averaging the attention maps of all 4 model replicates. Size-matched negative background regions were sampled from attention bins that did not overlap any peak, exon, or polyadenylation site, at a rate of 4x the number of positives. P-values were calculated with two-sided wilcoxon tests (n = number of sampled regions). **(b)** A horizontal slice of the attention weight matrix averaged across all 8 heads of the final transformer layers, shown for example region chr1:69967222-70491510. The slice is centered on the second TSS of the *SRSF11* gene. Left attention map: Reference sequence. Right attention map: Result of dinucleotide-shuffling a 256 bp region overlapping the second TSS of the *SRSF11* gene (orange arrow). Average predicted RNA-seq coverage for 89 GTEx samples is shown above each heatmap. **(c)** Attribution scores of the *SRSF11* promoter, using a variety of attribution methods. The attribution scores were calculated with respect to the sum of attention overlapping the second TSS (top) or with respect to the predicted log-sum of exon coverage in GTEx 'Lung' tissue samples (bottom). Parameters specific to each attribution method are annotated above each sequence logo. Min / max annotated in the right corner. **(d)** Attribution scores of one of the splice donors of the *SRSF11* gene. The attribution scores were calculated with respect to the log ratio of coverage across the exon relative to coverage over the intron. Min / max annotated in the right corner. **(e)** Attribution scores of the distal-most PAS of the *SRSF11* gene. The attribution scores were calculated with respect to the log ratio of coverage immediately upstream of the distal PAS relative to coverage immediately upstream of the proximal PAS. Min / max annotated in the right corner. **(f)** Average attention weight matrix of the 3' UTR of the LRRC7 gene (coordinates chr1:70116887-70149364). Left: Reference sequence. Right: Result of mutating the core hexamer motif of the distal-most PAS of LRRC7.



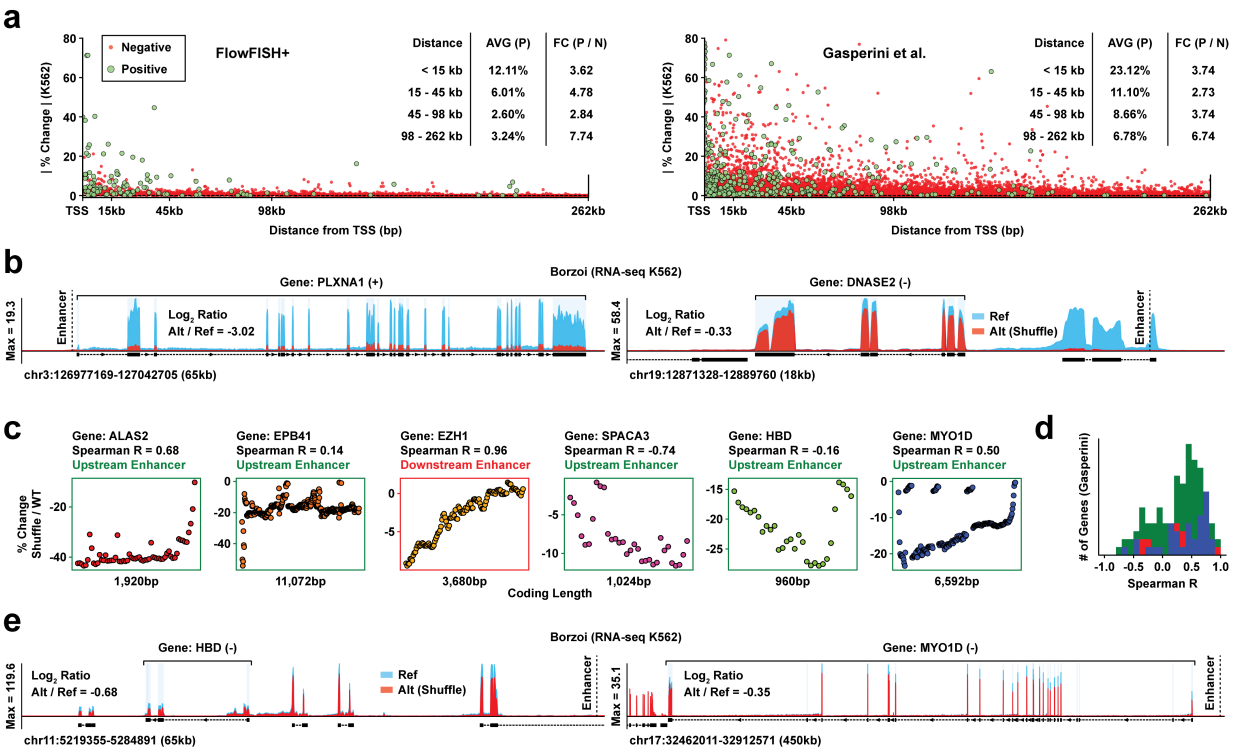
Supplementary Figure 3: **Additional gene- and isoform-level test set performance comparisons.** **(a)** Comparison between predicted and measured aggregated exon coverage for held-out test genes (n $= 1,940$). The predictions and measurements are averaged across all 89 trained-on GTEx RNA-seq coverage tracks (and the predictions are averaged across 4 model replicates). Points are colored according to a Gaussian kernel density estimate. **(b)** Predicted and measured RNA-seq coverage pattern in 5 GTEx tissues for test gene CREB3L3, which is highly expressed in tissue 'Liver' and lowly expressed in other tissues. The gene was chosen due to having among the highest measured expression fold change in liver compared to other tissues in GTEx. Predictions were averaged across all 4 model replicates. Exon-overlapping bins are shaded light-blue, and 'max' refers to the maximum bin value in the exonic regions. **(c)** Predicted vs measured 3'-to-5' TSS coverage ratio for held-out test genes, averaged across 89 trained-on GTEx RNA-seq assays and 4 model replicate predictions, excluding all genes whose TSS coordinates do not overlap the FANTOM5 annotation (Methods; n = 337). Genes whose alternative TSSs both reside in the GENCODE MANE transcript are colored red, while genes whose TSSs reside in exons of alternative isoforms are colored black. **(d)** Comparison between predicted TSS coverage ratios of held-out test genes (pooled across 89 GTEx assays) and measured tissue-pooled FANTOM5 TSS usage proportions. Points are colored according to a Gaussian kernel density estimate (in **d** - **h**). **(e)** Predicted vs measured fold change between the TSS coverage ratio in GTEx 'Whole Blood' and the average coverage ratio of 4 other tissues ('Brain', 'Liver', 'Muscle', 'Esophagus'), for held-out test genes. **(f)** Predicted vs measured coverage ratio between the distal-most and proximal-most 3' UTR polyadenylation signal (PAS) of each gene ('APA coverage ratio') for held-out test genes (n = 994), averaged across 89 GTEx assays and 4 model replicate predictions. **(g)** Predicted pooled-tissue APA coverage ratio vs measured isoform proportions from PolyADB (Wang et al., 2018). **(h)** Predicted vs measured fold change between the APA coverage ratio in GTEx 'Brain' and the average ratio of 4 other tissues ('Blood', 'Liver', 'Muscle', 'Esophagus'), for held-out test genes.



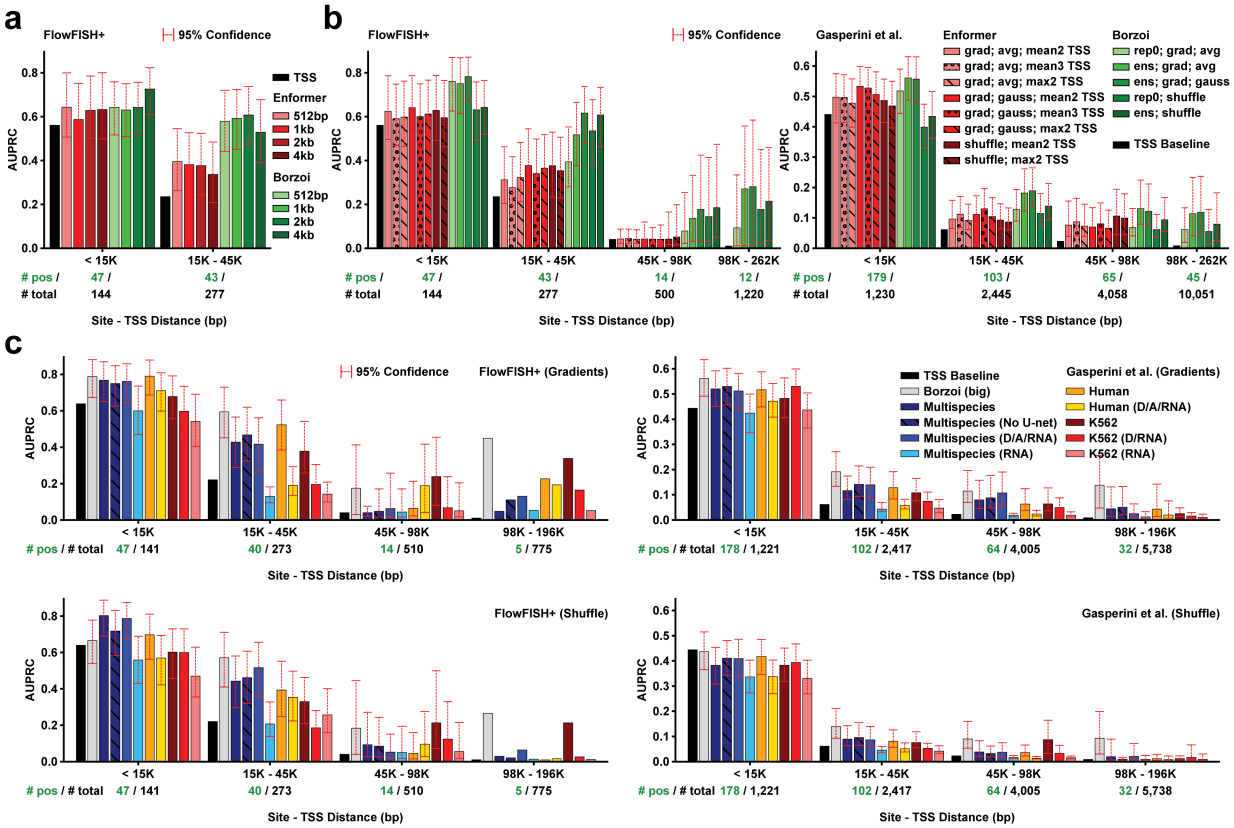
Supplementary Figure 4: **Detailed comparisons of gradient saliency and ISM scores.** **(a)** Gradient saliency scores calculated with respect to the log-sum of exon coverage for gene *FOLR3* (GTEx 'Whole Blood' tracks), displayed in a local sequence window centered on the region with maximal value. Gradients are shown for all 4 model replicates, as well as their average. In-silico saturation mutagenesis (ISM) is shown for the first model replicate ('Replicate 0') at the bottom. Min / max annotated in the right corner. **(b)** Distribution of correlations (Spearman R values) between gradient saliencies of different model replicates, computed over 200 blood-specific genes (the 200 genes with maximal fold change of 'Whole Blood' TPM relative to the TPM of other tissues in GTEx). The correlations were computed in a local window of $1,024$ bp centered on the position of maximal gradient score. **(c)** Distribution of correlations (Spearman R) between gradient saliency and ISM ('Replicate 0' only), computed for 200 blood-specific genes. The correlations were calculated in a local window of 192 bp centered on the position of maximal gradient score. **(d)** Left: Comparison between measured median TPM log2 fold changes in GTEx tissues 'Liver' and 'Muscle' and either (1) log2 fold changes of predicted sum of exon coverage in Borzoi's corresponding GTEx tracks, or (2) the average difference in input-gated gradient saliency within a 192-bp window centered at the mode of maximal differential saliency per gene. The predictions and gradients were computed using only a single model replicate ('Replicate 0'). The metrics were evaluated on the subset of $5,000$ previously chosen tissue-specific genes that (by random chance) happened to belong to the test set (denoted 'Test'; n = 207), or on a size-matched sample of trained-on loci (denoted 'Train'; n = 204). Right: Median Spearman correlations of the aforementioned evaluations taken across all combinations of tissue pairs (for tissues 'Whole Blood', 'Liver', 'Brain - Cortex', 'Muscle' and 'Esophagus - Muscularis').



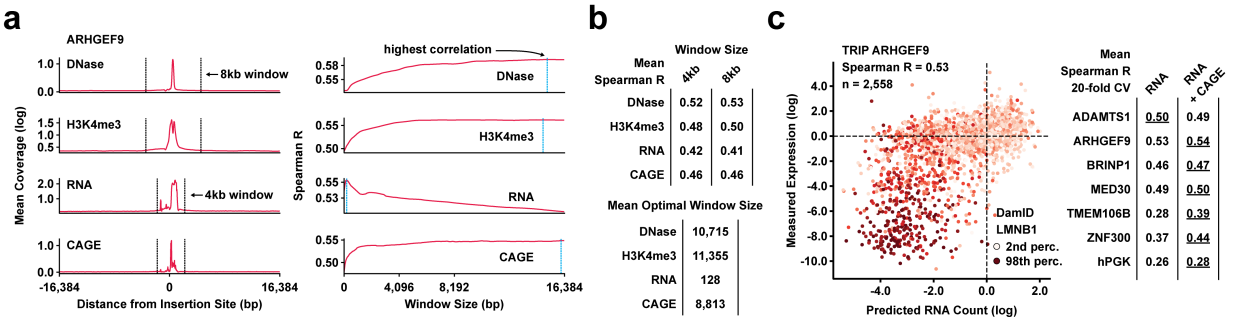
Supplementary Figure 5: **Additional motif-based analyses. (a)** A subset of motif clusters identified by MoDISco when executed on raw log-transformed exon-aggregated gradient saliencies computed from 4 GTEx RNA-seq coverage tracks ('raw' = no subtraction of the average gradient saliency of other tissues). Shown are the MoDISco PWMs, the best-matching HOCOMOCO v11 PWMs and the distribution of tissue-specific gradient saliencies for within-cluster seqlets. P-values were computed using a two-sided Wilcoxon test between the two tissues with largest absolute magnitude in gradient saliency (95th percentile). P-values ranged from 1.65×10^{-14} to 1.27×10^{-138} (n = number of seqlets). Each score distribution is annotated with the 5th, 50th and 95th percentile. **(b)** The same subset of motifs identified by running MoDISco on differential gradient saliencies with pseudocount added to the predictions before log-normalizing. The pseudocount was calculated as the 2nd percentile of expression predictions across all $5,000$ interpreted genes. P-values were computed using a two-sided Wilcoxon test, and ranged from 1.67×10^{-72} to 2.32×10^{-195} (n = number of seqlets). **(c)** Additional MoDISco motif clusters identified by interpreting Esophagus-specific genes (using predictions and gradients derived from Esophagus coverage tracks). The gradients were normalized by subtracting the average saliency of other tissues. P-values were computed using a two-sided Wilcoxon test, and ranged from 1.04×10^{-17} to 2.49×10^{-96} (n = number of seqlets). The E-values represent the significance of motif matches as computed by Tomtom (in **c** - **d**). **(d)** Motif clusters identified by MoDISco in gradients derived from Borzoi's GTEx 'Whole Blood' and K562 RNA coverage tracks, highlighting the differences in learned regulatory patterns. P-values were computed using a two-sided Wilcoxon test, and ranged from 1.29×10^{-20} to 6.06×10^{-305} (n = number of seqlets). **(e)** Top: Median TPM quantile of a subset of blood-specific TF genes in GTEx 'Whole Blood' compared to TPM quantiles in K562 (median quantiles calculated across 5 ENCODE experiments: 'ENCSR000AEL', 'ENCSR000AEM', 'ENCSR000AEN', 'ENCSR000AEO', 'ENCSR000CPH'). Bottom: Scatter plot comparing TPM quantiles of genes in GTEx 'Whole Blood' to K562 measurements (ENCODE experiment 'ENCSR000AEL'). Blood- and K562-enriched TF genes are annotated with red and blue color respectively. **(f)** Average residual gradient saliency of cluster seqlets for pairs of 4 distinct GTEx tissues, compared to the measured difference in log-TPM of the corresponding TF genes within the same GTEx tissues ('Whole Blood', 'Liver', 'Brain - Cortex' and 'Muscle'). The median TPM of genes belonging to the same TF subfamily (according to HOCOMOCO v11) were averaged. A subset of scatter plots for this analysis were shown in the main figure. **(g)** Spearman correlation between the average residual gradient saliency of MoDISco motifs and the difference in gene TF log-TPM for all 5 GTEx tissues (including 'Esophagus - Muscularis').



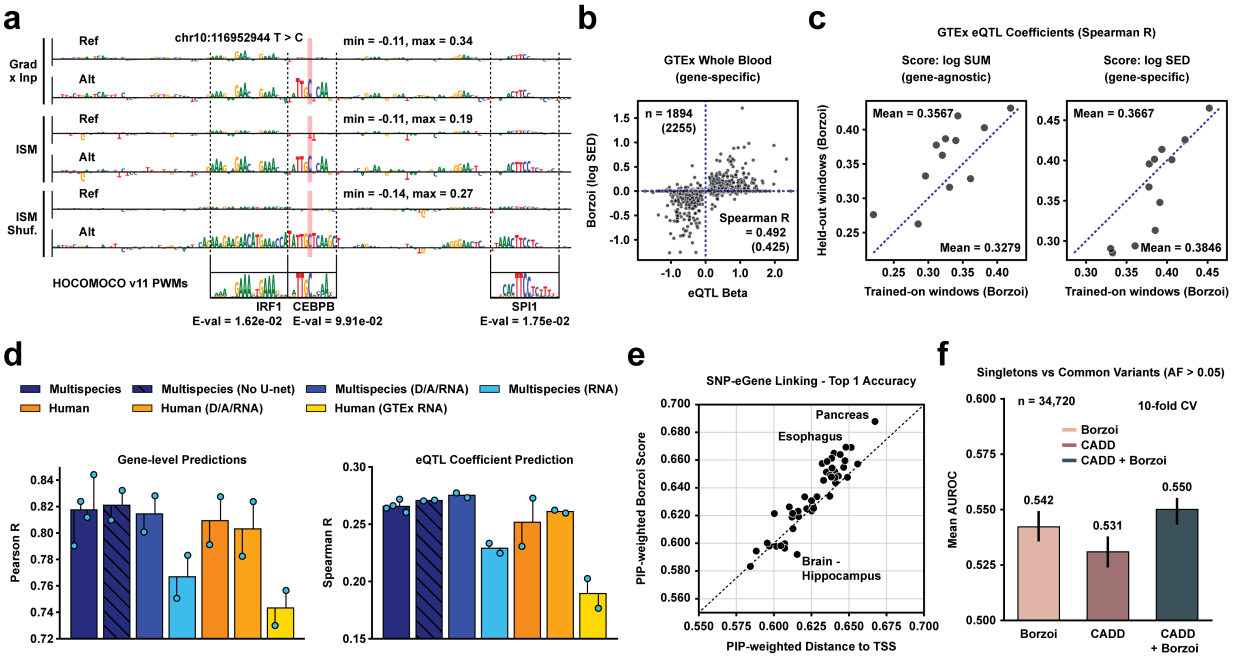
Supplementary Figure 6: **CRISPR perturbation analysis based on dinucleotide-shuffling putative enhancers.** **(a)** Predicted % change in sum of exon coverage of target genes when dinucleotide-shuffling a 2kb window centered on each of the CRE loci measured in Nasser et al., (2021) (top) or Gasperini et al. (2019) (bottom), using K562 RNA-seq output tracks. Green / red data points represent measured regulating / non-regulating loci respectively. Each dinucleotide-shuffle was repeated 16 times for each of the 4 model replicates. Inset table: ('AVG (P)') average % change of positive examples, and ('FC (P / N)') relative fold change between the average % change of positive examples and the average % change of negative examples. **(b)** Reference (blue) and perturbed (red) coverage prediction in K562 RNA-seq upon dinucleotide-shuffling putative enhancer loci. The enhancers are annotated as negative (non-regulating) examples for target genes *PLXNA1* and *DNASE2*, while Borzoi predicts a relatively large % change in RNA-seq coverage upon shuffling each respective locus. The log ratio of total predicted coverage between the alternate and reference sequences is annotated. Exon-overlapping bins are shaded light-blue. **(c)** Dinucleotide-shuffling a 2kb window to perturb the CRE locus of 6 positive CRE-gene pairs in the data from Nasser et al., (2021), while recording the predicted % change in coverage across all exon-overlapping bins of each target gene using Borzoi (4 model replicates). Each CRE is dinucleotide-shuffled 32 times. The x-axis corresponds to increasing distance of each exon bin to the CRE locus (exonic distance; excluding intergenic or intronic regions), i.e. left-most data points are closest to the CRE. Green / red border color corresponds to upstream or downstream enhancer locations respectively. **(d)** Distribution of Spearman correlation coefficients between the exonic bin distance to a putative CRE and the % change in coverage of the corresponding bin. Positive correlation corresponds to a decreasing drop in coverage as a function of distance from the CRE. Only positive CRE-gene pairs in the data from Gasperini et al. (2019) were considered. Green / red / blue color corresponds to upstream, downstream or intronic CRE locations relative to the gene (i.e. intronic = CRE is within one of the introns of the target gene). **(e)** Two example CRE-gene pairs (positive examples) in the data from Nasser et al., (2021). Shown are the predicted reference coverage tracks (blue; K562 RNA-seq) and the resulting coverage prediction after dinucleotide-shuffling the enhancer locus (red). The log ratio of total predicted coverage between the alternate and reference sequences is annotated. Exon-overlapping bins are shaded light-blue. Gene *HBD* exhibits an overall positive correlation between exonic bin distance to the enhancer and the absolute magnitude of drop in coverage, while *MYO1D* displays a negative trend (decreasing absolute magnitude in coverage drop with increasing distance to the exonic bin).



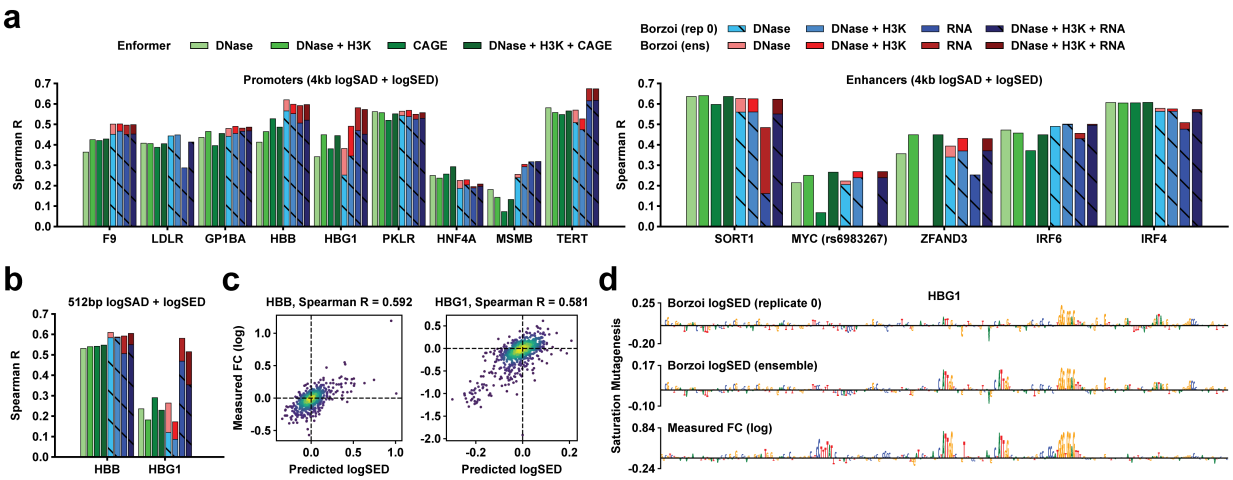
Supplementary Figure 7: **Additional CRISPR perturbation benchmark comparisons.** **(a)** Average precision (AUPRC) when classifying regulating / non-regulating CRE loci from Nasser et al., (2021) based on the absolute value of predicted % change of the target genes, using either Borzoi or Enformer, when dinucleotide-shuffling the CREs with increasingly larger window sizes. The % change is estimated from aggregated exon coverage in K562 RNA-seq tracks for Borzoi and from aggregated K562 CAGE signal for Enformer (using a window of 2 128bp bins centered at each TSS; the average is taken across all TSSs of a gene). 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 $= 277$. 95% confidence intervals were estimated from $1,000$-fold bootstrapping. **(b)** Average precision (AUPRC) when classifying regulating / non-regulating CREs in data from Nasser et al., (2021) (left) or Gasperini et al. (2019) (right), comparing a variety of method configurations for Borzoi and Enformer, including: 'grad' or 'shuffle' = attribution scores derived from gradient saliencies or 2kb dinucleotide-shuffling; 'avg' or 'gauss' = deriving the final statistics from the average gradient in a window centered on the CRE locus or from a smaller gaussian-smoothed window (Methods); 'mean2 TSS' or 'max2' TSS = using a window of 2 128bp bins centered at each TSS to compute Enformer CAGE gradients and either averaging or taking the maximum absolute-valued score for genes with multiple TSSs; 'mean3 TSS' = using a window of 3 128bp bins centered at each TSS, and averaging the gradient saliency score across multiple TSSs, to compute Enformer CAGE statistics; 'rep0' or 'ens' = using only the first Borzoi replicate or the full ensemble. The total number of examples in the FlowFISH+ data are: (< 15K) n $= 144$, (15K - 45K) n $= 277$, (45K - 98K) n $= 500$, (98K - 262K) n $= 1,220$. The total number of examples in the Gasperini data 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. **(c)** Average precision (AUPRC) when classifying regulating / non-regulating CREs, using scores derived either from gradient saliencies with gaussian smoothing (top) or 2kb dinucleotide-shuffles (bottom), comparing different Borzoi models trained on only a subset of the original training data (or trained on the full data but with architectural changes): 'Borzoi (big)' = original model; 'Multispecies' = all the original training data in human and mouse (baseline); 'Multispecies (No U-net)' = the u-net was removed (128bp bin resolution); 'Multispecies (D/A/RNA)' = only DNase, ATAC and RNA datasets were trained on (human and mouse); 'Multispecies (RNA)' = only RNA datasets were trained on (human and mouse); 'Human' = all the original human training data (no mouse data); 'Human (D/A/RNA)' = only human DNase, ATAC and RNA datasets were trained on; 'K562' = all the original training data in K562 were trained on (including ChIP-seq and CAGE); 'K562 (D/A/RNA)' = only DNase, ATAC and RNA datasets in K562 were trained on; 'K562 (RNA)' = only K562 RNA data were trained on. Note that, due to computational constraints, these models were trained on a smaller input window (393kb), with a smaller number of weights (~ 30 million), and only two model replicates were used for ensembling. Also note that the classification statistic of each model is always derived from the K562 RNA-seq output tracks, which suggests that any significant differences observed in model performances is a consequence of indirect interactions between trained-on data modalities. The total number of examples in the FlowFISH+ data are: (< 15K) n $= 141$, (15K - 45K) n $= 273$, (45K - 98K) n $= 510$, (98K - 196K) n $= 775$. The total number of examples in the Gasperini data are: (< 15K) n $= 1,221$, (15K - 45K) n $= 2,417$, (45K - 98K) n $= 4,005$, (98K - 196K) n $= 5,738$. 95% confidence intervals were estimated from $1,000$-fold bootstrapping (the confidence interval is undefined for the last FlowFISH+ distance category, as the number of positives is too small for bootstrapping).



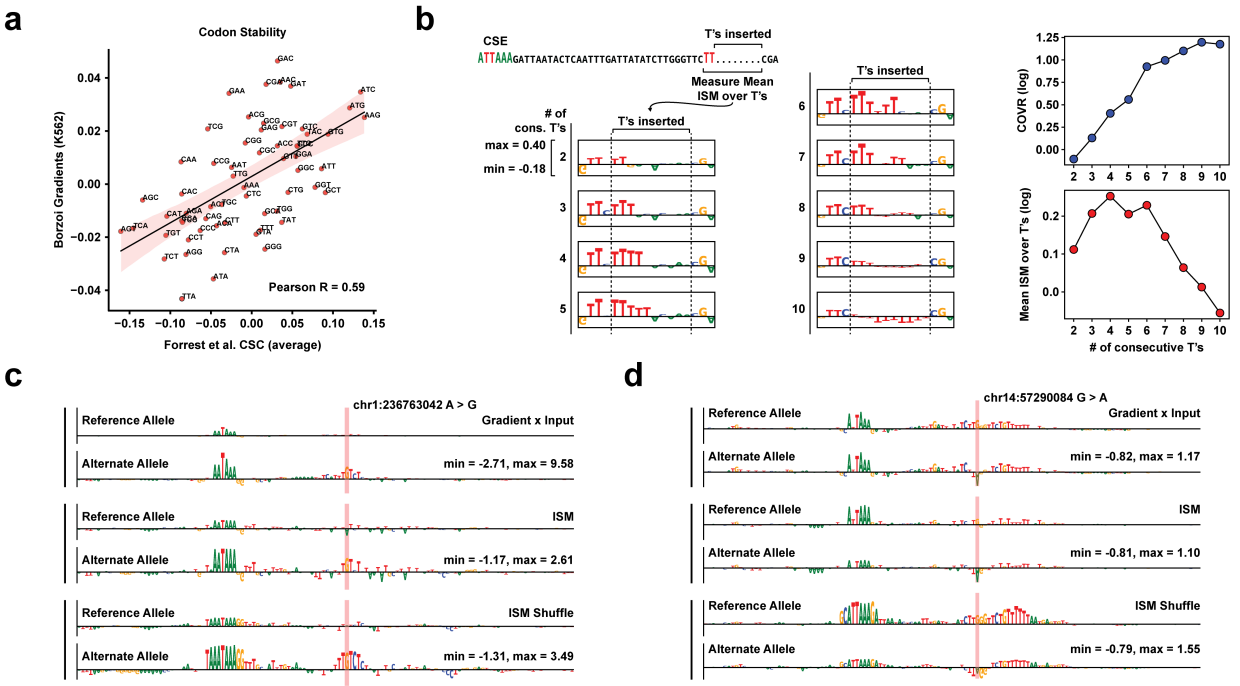
Supplementary Figure 8: **Additional TRIP comparisons.** **(a)** Left: Average predicted coverage across all insertions of the *ARHGEF9* TRIP reporter for various assay types. Distance = 0 corresponds to the position of insertion. All insertions on the negative strand have been reverse-complemented such that the promoter always extends to the right. All predictions were made using K562 output tracks only. The window sizes utilized for all TRIP analyses (8kb for DNase and histone tracks; 4kb for CAGE and RNA tracks) are annotated in the plots. Right: Spearman correlation between measured *ARHGEF9* reporter expression levels and predicted aggregated coverage as a function of the window size used for aggregation. **(b)** Top: Average Spearman correlation across all reporter constructs (promoter contexts) using either a 4kb or 8kb window of aggregation when computing the prediction statistic. Bottom: Average optimal (Spearman-maximizing) window sizes per assay. **(c)** Left: Predicted vs measured expression levels of TRIP reporter constructs based on Borzoi RNA-seq coverage in K562 (Promoter: *ARHGEF9*). Color = DamID measurements, n = number of TRIP insertions. Right: Average Spearman R (20-fold cross validation) when predicting TRIP expression based on K562 RNA-seq coverage, or when using coverage statistics derived from both RNA-seq tracks and CAGE tracks to make the prediction (Methods).



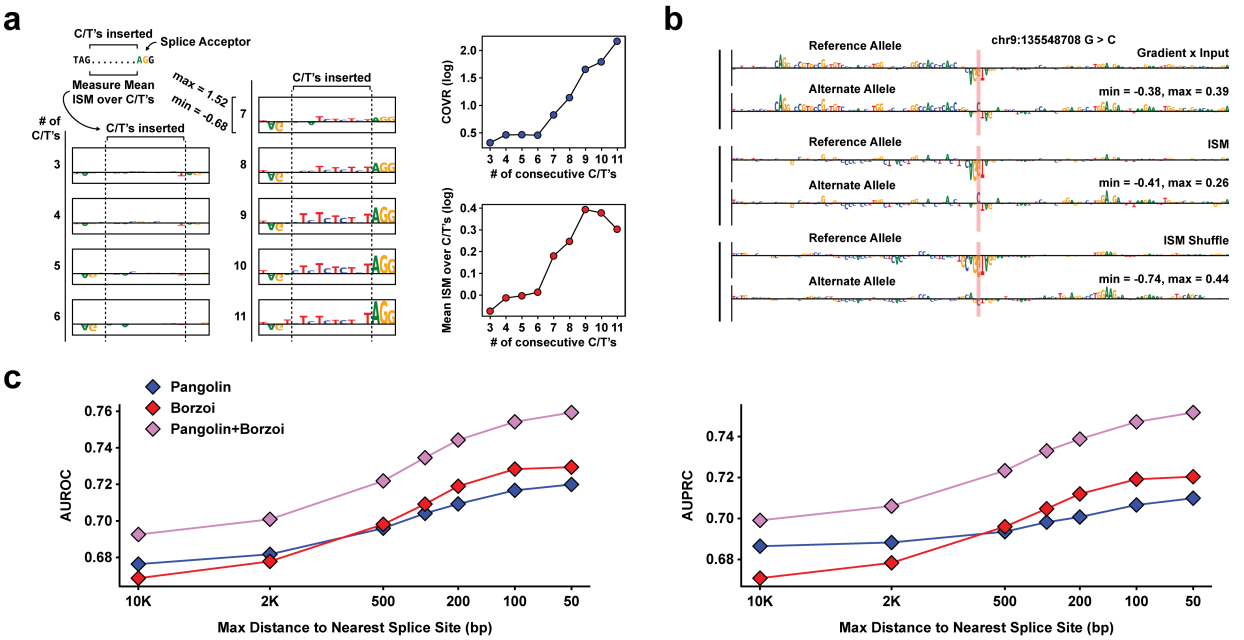
Supplementary Figure 9: **Additional fine-mapped eQTL analyses and gnomAD benchmark comparisons.** **(a)** Comparison of three different attribution methods (Input-gated gradient saliency, ISM, and ISM Shuffle) in a local sequence window centered on variant rs1905542 in the intron of the *SHTN1* gene. The sequence logo y-axes are equally scaled for both the reference and alternate alleles, but differently scaled for distinct attribution methods (min / max annotated in the logos). Putative TF binding motifs are displayed at the bottom and aligned with the input sequence (HOCOMOCO v11 PWMs with lowest p-values as determined by Tomtom MEME; Tomtom motif E-values annotated under each PWM). **(b)** Predicted vs observed effect sizes of fine-mapped causal GTEx eQTL SNPs for Whole Blood. The predictions are calculated as the log-fold change in exon-aggregated coverage of the target gene ('logSED' scores) and compared against that gene's eQTL coefficient (i.e. a gene-specific analysis). A Spearman R of 0.492 is obtained when removing SNPs with target genes that are outside Borzoi's output window (kept n = $1,894$). A lower Spearman R is obtained (0.425) when instead treating these predictions as 0 (total n = $2,255$). **(c)** Spearman correlation between predicted and observed effect sizes of fine-mapped causal eQTL SNPs from GTEx. SNPs are separated based on whether the variant coordinates occurred in trained-on or held-out hg38 sequence windows during training (Note: the alternate allele is never seen during training, regardless of whether the coordinate overlaps trained-on or held-out genomic windows). Each dot represents a tissue. Only tissues for which there were at least 200 SNPs in held-out sequence windows were included in the analysis. Left: Gene-agnostic eQTL coefficient analysis using the 'logSUM' score. Right: Gene-specific analysis using the 'logSED' score. Overall, only a small difference in average Spearman R is noted between trained-on and held-out genomic sequence windows (< 0.03). Mean Spearman R values for trained-on / held-out windows are annotated in the plot. **(d)** Performance comparison when making gene-level (exon-aggregated) predictions on the held-out test set (left) or when predicting fine-mapped eQTL coefficients (right) using different Borzoi models trained on only a subset of the original training data (or trained on the full data but with architectural changes): 'Multispecies' = all the original training data in human and mouse (baseline); 'Multispecies (No U-net)' = the u-net was removed (128bp bin resolution); 'Multispecies (D/A/RNA)' = only DNase, ATAC and RNA datasets were trained on (human and mouse); 'Multispecies (RNA)' = only RNA datasets were trained on (human and mouse); 'Human' = all the original human training data (no mouse data); 'Human (D/A/RNA)' = only human DNase, ATAC and RNA datasets; 'Human (GTEx RNA)' = only GTEx RNA data. Due to computational constraints, the models were trained on a smaller input window (393kb) with a smaller number of weights (~ 30 million). Each bar shows the average metric of two independently trained model replicates, except for the baseline 'Multispecies' model which was trained with 4 replicates. The gene-level test set predictions are evaluated only on the 89 trained-on GTEx RNA-seq tracks. The eQTL analysis was performed using the gene-specific 'logSED' score. Shown are the average Pearson R / Spearman R metrics across GTEx tissues. **(e)** The top-1-accuracy (percentage of eGenes that are correctly predicted) by posterior probability (PP)-weighted Borzoi scores compared to PP-weighted nearest gene across 49 GTEx tissues. **(f)** Left: Mean AUROC when classifying singleton variants from common variation (AF > 0.05) from gnomAD (10-fold CV). The error bars indicate the 95% confidence interval, estimated from $1,000$-fold bootstrapping.



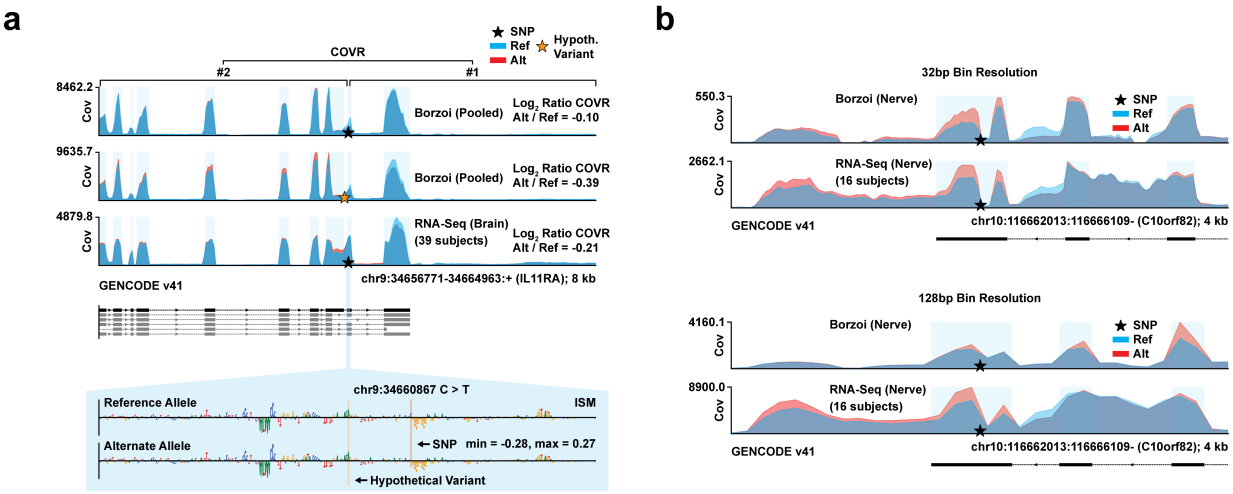
Supplementary Figure 10: **Saturation mutagenesis MPRA benchmark comparisons. (a)** Performance comparison when using Borzoi or Enformer to predict saturation mutagenesis MPRA data from Kircher et al. (2019), with scores derived from various subsets of data modalities predicted by the two models. All variants were evaluated within their native context in the reference human genome. The variant effect scores were derived from log-fold changes of total coverage aggregated in a 4kb window centered on the variant (for DNase, histone and CAGE predictions), or from log-fold changes of aggregated exon coverage of the target gene (for RNA predictions). Spearman R was computed between the predicted variant scores and the expression log fold changes measured in the MPRA. The original cell type mappings used in the Enformer paper by Avsec et al. (2021) were used here for Borzoi, except when there was a different cell type mapping whose predictions consistently increased performance across all assay types and when that cell type was reasonable to utilize (e.g. using K562 for the *F9* gene, which encodes a coagulation factor; Methods). For every cell type mapping swapped for Borzoi, we tested the same mapping with Enformer and switched to the new cell type for Enformer as well if it led to an increase in performance. Only promoters and enhancers with a better 'cell-type matched' performance than the 'cell-type agnostic' performance in Figure 4a of the original Enformer paper were included in this analysis (for either the CAGE or DNase assays). Blue dashed bars correspond to Borzoi predictions using only a single model replicate. **(b)** Repeat performance comparison for a subset of promoters tested by Kircher et al. (2019), using a smaller 512bp window size to aggregate coverage at the variant position. This was the original window size used by Enformer. **(c)** Variant effect predictions for the *HBB* and *HBG1* promoters using Borzoi (4-replicate ensemble). The scores were derived from predicted log-fold changes of aggregated gene exon coverage in the K562 RNA-seq tracks, and compared to MPRA measurements. **(d)** Predicted and measured saturation mutagenesis of the *HBG1* promoter, visualized as sequence logos (where letter height is proportional to the average effect of the three alternative alleles). Borzoi predictions are derived from K562 RNA-seq tracks, using either a single model replicate or the full ensemble.



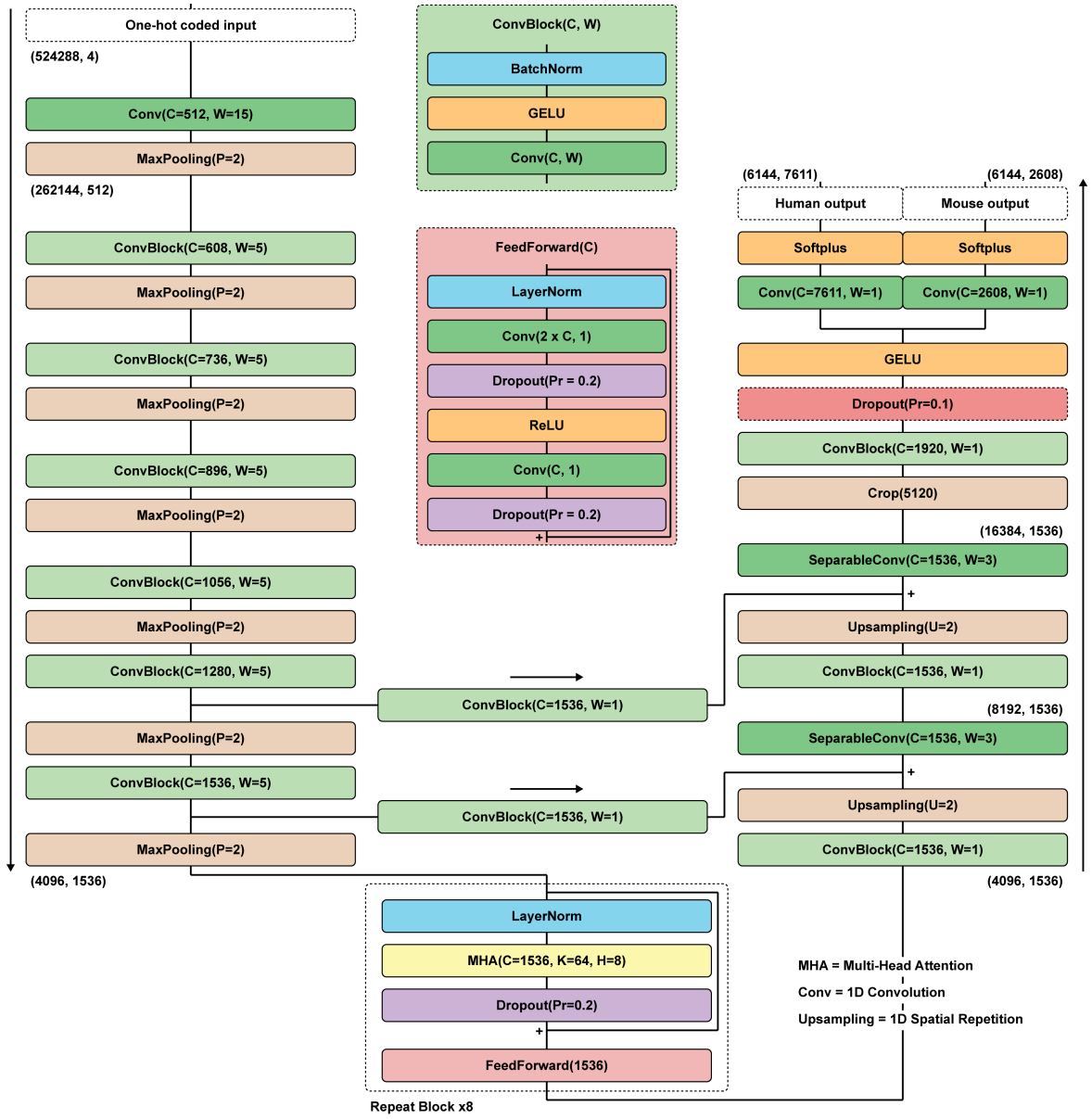
Supplementary Figure 11: **Additional 3' UTR analyses and fine-mapped paQTL variant interpretations.** **(a)** Comparison between the average gradient saliency of codons in genes from the Gasperini set and measured codon stability in data from Forrest et al. (2020). The gradient saliencies are calculated with respect to the predicted sum of exon coverage in K562 RNA-seq samples. The plot includes a 95% confidence area of the linear regression line, estimated from $1,000$-fold bootstrapping. **(b)** Left: ISM-based analysis of buffering effects of T-rich motifs in a polyadenylation signal in the *AP5M1* gene. Starting from a downstream region (DSE) with neutral ISM contribution scores, T nucleotides are sequentially and incrementally substituted for the neutral bases. Simultaneously, ISM scores are computed (from pooled-tissue GTEx RNA-seq coverage) and the average contribution score of T's in this region are recorded (starting from 2 consecutive T's up to 10 T's). Annotated are the min / max values of the ISM logos, and the number of T's inserted in the region. Right: Plots showing the predicted polyadenylation site 'strength' (RNA-seq coverage ratio) (top) and mean ISM contribution score across T's (bottom), as a function of the number of consecutive T's inserted in the DSE region. Note: The starting sequence was modified from the reference DSE region of the polyadenylation signal in the *AP5M1* gene (it was altered to have ISM contribution scores close to 0). **(c)** Attribution scores for a polyadenylation signal (PAS) in the *ACTN2* gene when inducing variant rs114880747. The resulting scores of three separate attribution methods are displayed (equally scaled y-axes for the reference and alternate allele; min / max displayed in the right corner). The attribution scores are calculated with respect to the log ratio of coverage immediately upstream of the mutated PAS relative to coverage upstream of the distal-most PAS. **(d)** Attribution scores for a PAS in the *AP5M1* gene when inducing variant rs80168986, using three attribution methods. The y-axes are equally scaled for the reference and alternate allele (min / max displayed in the right corner)



Supplementary Figure 12: **Additional fine-mapped sQTL analyses and benchmark comparisons. (a)** Left: ISM-based analysis of buffering effects in the polypyrimidine-tract of a splice acceptor site in the *ACRBP* gene. The nucleotides of an initially neutral starting sequence (according to ISM contribution scores) are sequentially replaced by repeating C and T nucleotides. For every position substituted in the sequence, ISM scores are computed (from pooled-tissue GTEx RNA-seq coverage) and the average C/T contribution score is recorded (starting from 3 consecutive C's or T's up to 11 C's or T's). Annotated are the min / max values of the ISM logos, and the number of C/T's inserted in the region. Right: Plots showing the predicted splice site 'strength' (RNA-seq coverage ratio) (top) and mean ISM contribution score of C's and T's (bottom), as a function of the number of consecutive C's or T's inserted in the polypyrimidine tract. Note: The starting sequence is modified from the reference polypyrimidine tract of the original splice site in the *ACRBP* gene (it was altered to have ISM contribution scores close to 0). **(b)** Attribution scores of exon-to-intron log coverage ratio in a local window centered on the variant rs55695858, comparing three separate methods (equally scaled y-axes for the reference and alternate allele; min / max displayed in the right corner). **(c)** Tissue-pooled variant effect prediction benchmark, comparing the performance of Pangolin, Borzoi and an ensemble of the two models (their average rank predictions) on the task of classifying between fine-mapped causal sQTLs (PP > 0.9) and expression- and distance-controlled negatives. The AUROC and AUPRC is reported for each model as a function of decreasing distance threshold to the nearest splice junction. A similar type of benchmark is presented in the main figure, but the results are separated by GTEx tissue in that analysis.



Supplementary Figure 13: **Additional fine-mapped intronic paQTL analyses. (a)** Predicted RNA-seq coverage (GTEx pooled-tissue) and measured coverage in 39 individuals with the alternative (hetero- or homozygous) or reference allele of variant rs11575580. Attributions (bottom) are calculated with respect to the log ratio of downstream-to-upstream coverage (equally scaled y-axes; min / max displayed in the right corner). A hypothetical variant (which was not observed in the population) is also visualized (orange star). The log ratio between the coverage ratio (COVR) statistics computed for the alternate and reference alleles is annotated in the plot. **(b)** Predicted and measured RNA-seq coverage in GTEx tissue 'Nerve' for variant rs3830026. The measurements were aggregated from 16 individuals who were homozygous for the reference allele and 16 individuals who were hetero- or homozygous for the alternative allele. Top: Predictions and measurements at 32 bp bin resolution. Bottom: 128 bp bin resolution. Predictions were made using smaller Borzoi models (393kb input window, $\sim$ 30 million weights) trained on human and mouse data (DNase, ATAC, CAGE, ChIP, RNA). The predictions at 128bp bin resolution were made using a version of the Borzoi model trained without the U-net component. See Methods for a detailed description of these smaller Borzoi models.



Supplementary Figure 14: **Borzoi model architecture.** Neural network diagram. C = Channels. W = Kernel width. P = Pooling size. H = Attention heads. Pr = (Dropout) Probability. MHA = Multi-Head Attention. All convolutions are 1-dimensional.