

Supplementary Information for Paper: "Deciphering the Impact of Genetic Variation on Human Polyadenylation using APARENT2"

Johannes Linder^{1*}, Samantha E. Koplik², Anshul Kundaje^{1,3} and Georg Seelig^{4,5}

¹ Department of Genetics, Stanford University, Stanford, USA

² Department of Bioengineering, University of Washington, Seattle, USA

³ Department of Computer Science, Stanford University, Stanford, USA

⁴ Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA

⁵ Department of Electrical & Computer Engineering, University of Washington, Seattle, USA

* Correspondence: jlinder2@stanford.edu

Supplementary Figures

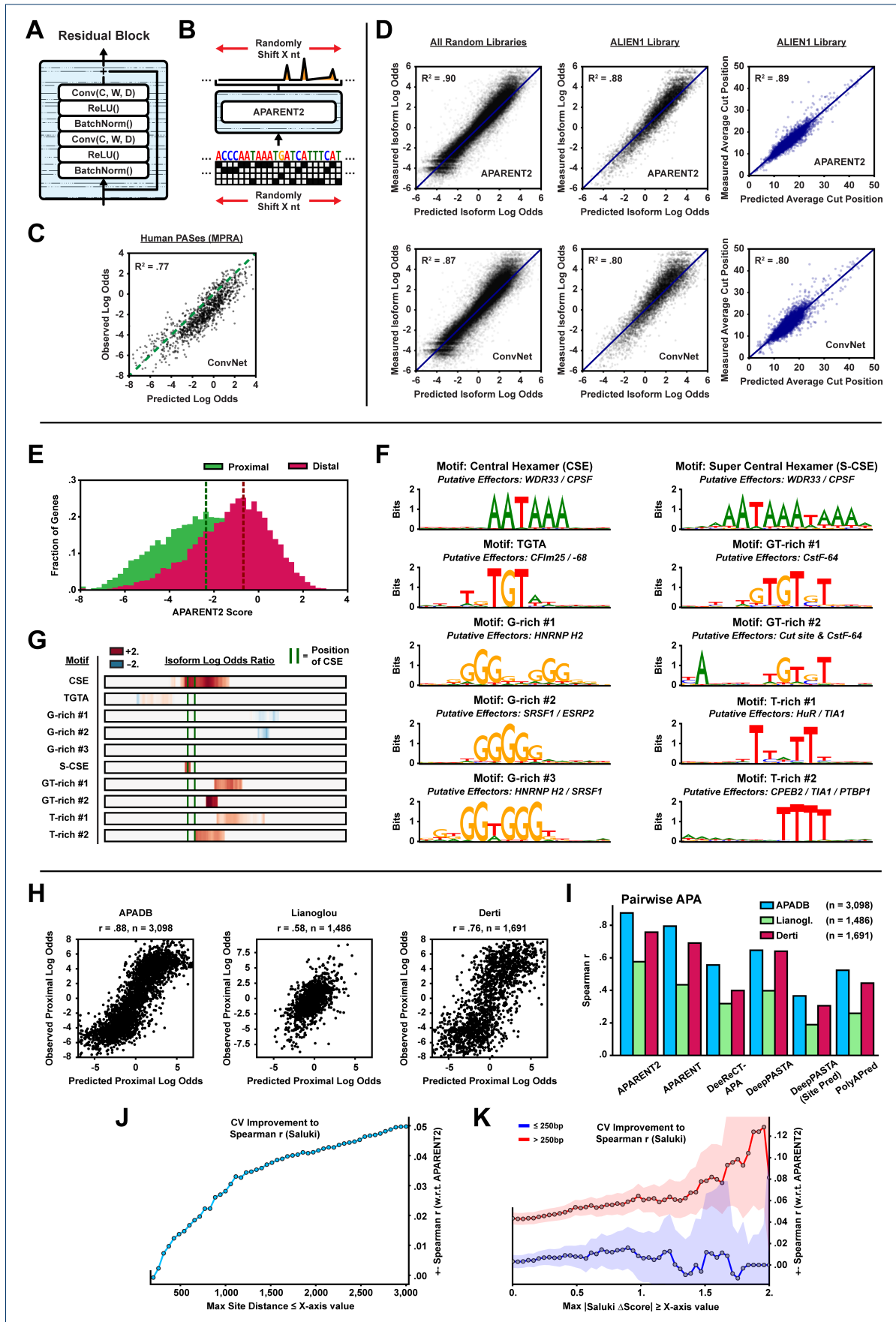


Figure S1: **A** The internal architecture of a residual block. C = # of channels, W = filter width, D = dilation rate. **B** During training, the input sequences and their target distributions are randomly shifted by a number of nucleotides. **C** Predicted vs measured isoform log odds of held-out human PASs measured in the MPRA of Bogard *et al.* (2019) for the baseline ConvNet model ($n = 1,085$). **D** Predicted vs measured proximal isoform log odds of held-out test sequences from all MPRA libraries (left) ($n = 60,198$) or the ALIEN1 library only (middle) ($n = 7,755$), and predicted vs measured average cut position downstream of the CSE of ALIEN1 test sequences (right) ($n = 6,203$). **E** Distribution of predicted isoform log odds using APARENT2 for the proximal-most and distal-most PASs in the 3' UTR of $n = 12,503$ genes (PolyADB V3). **F** Selection of RNA binding protein (RBP) motifs generated by TF-MoDISco. Each motif is represented as a position weight matrix (PWM). The motifs were generated from the SHAP scores of $n = 20,000$ randomly sampled PAS sequences from PolyADB V3. **G** Estimated isoform log odds ratios in the presence of a given motif at a specific position in the sequence (i.e. the increase, *or* decrease, in predicted isoform log odds when a given motif is present). **H** Predicted vs observed isoform log odds between pairs of adjacent human 3' UTR PASs, as measured in three separate native transcriptomic datasets. Predictions are made by linear regression of the APARENT2 scores of the proximal and distal signals and their log-distance as features. Read counts were pooled across tissues. A minimum read count of 500 was used as cutoff for all three data sources. **I** Comparison of correlation between predicted and measured isoform log odds for pairs of adjacent human 3' UTR PASs. Each model predicts logit scores which are used to fit a pairwise APA regressor (20-fold cross-validation). **J** Improvement in spearman r when using Saluki scores in addition to APARENT2 as input, as a function of the maximum distance between any adjacent pair of PASs in a given gene (data source = APADB). **K** Improvement in spearman r as a function of the minimum difference in Saluki score between at least one pair of PASs (blue / red = genes with PAS distances ≤ 250 bp / > 250 bp; shaded area = 90% confidence interval estimated by 10,000-fold bootstrapping; median improvement across data sources).

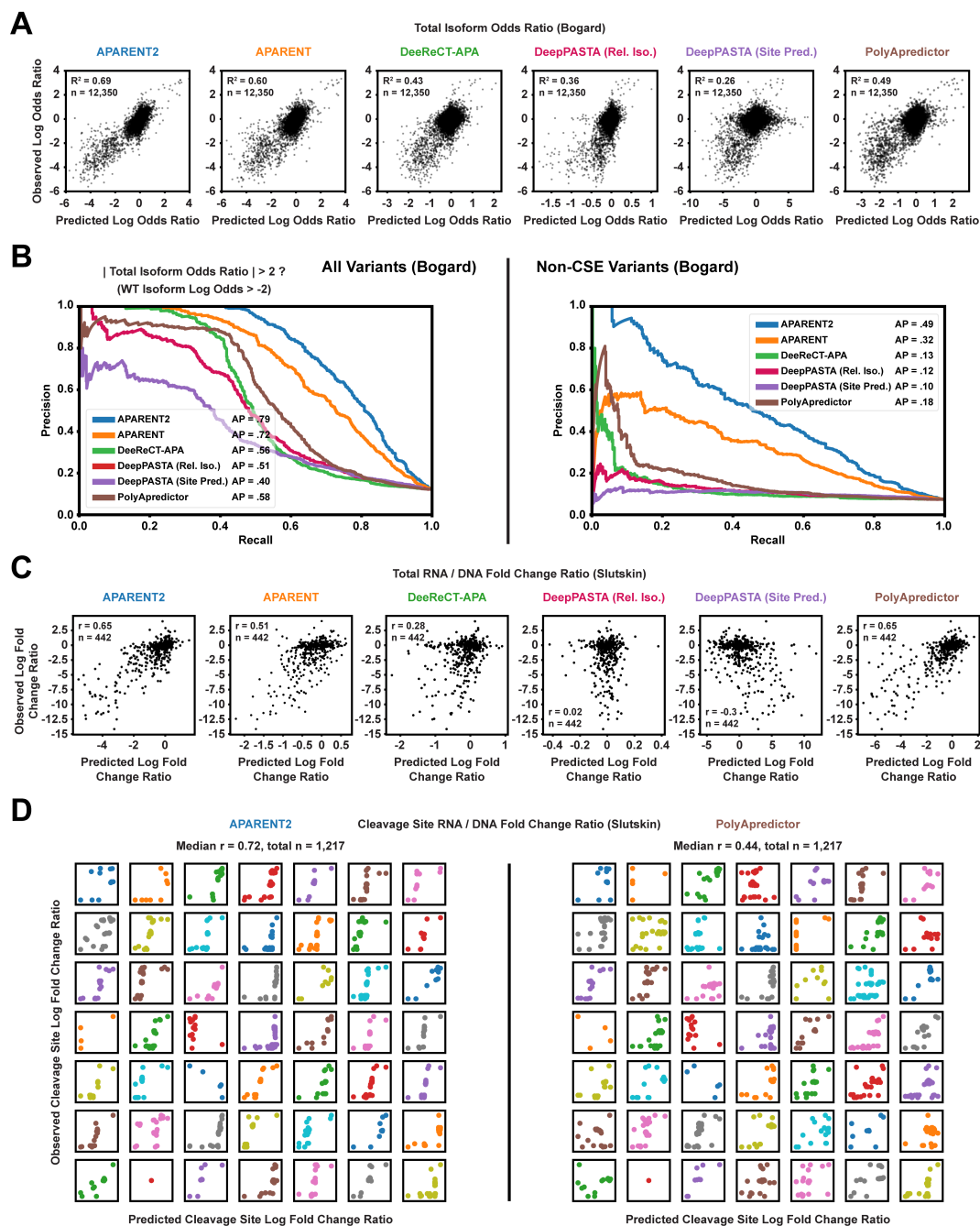
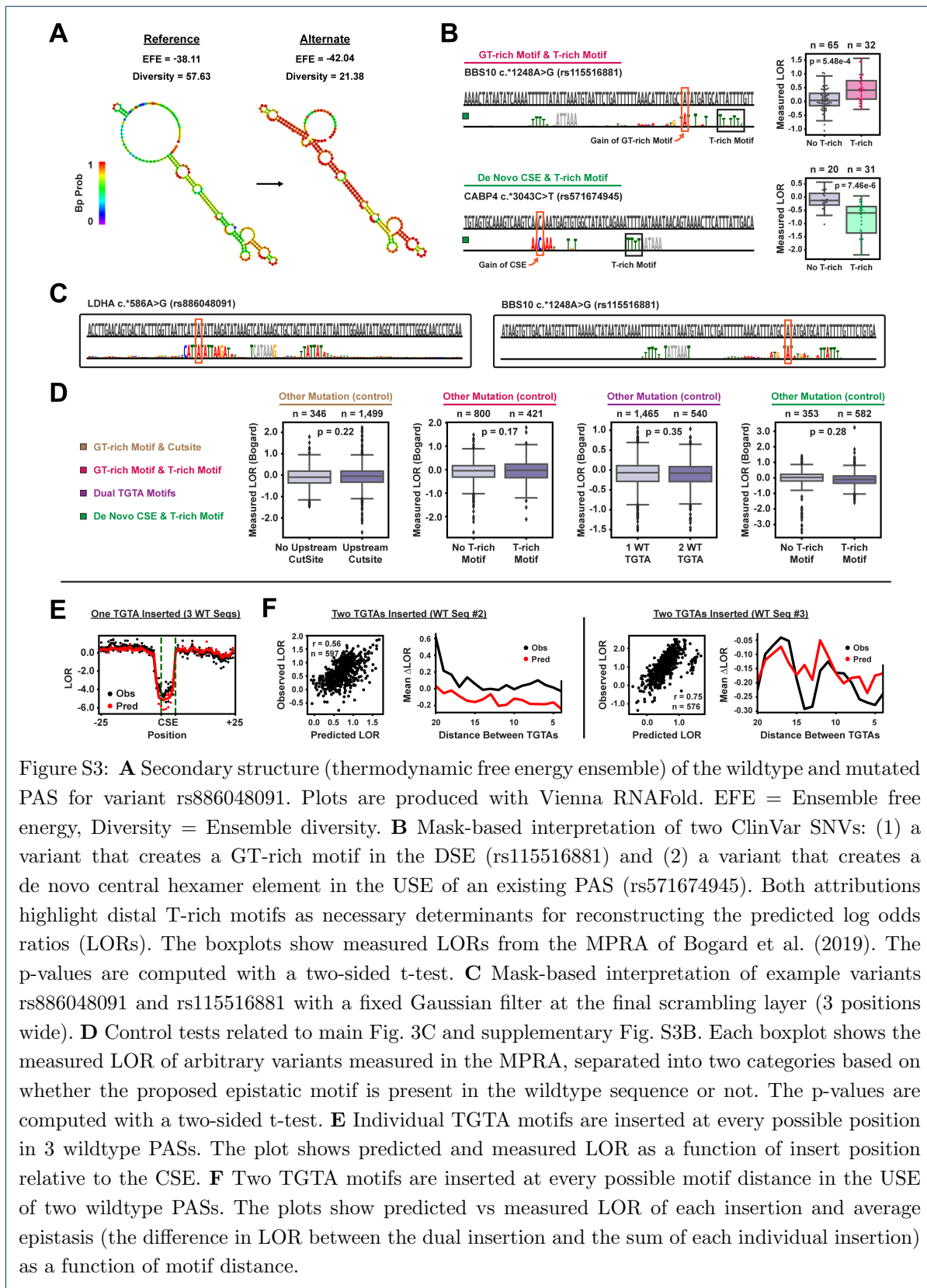


Figure S2: **A** Comparison of predicted vs measured variant isoform log odds ratios on the MPRA data from Bogard et al. (2019) ($n = 12,350$). Individual scatter plots are shown for each tested model. **B** Comparison of precision-recall curves when tasking each model with classifying disruptive APA variants on the data from Bogard et al. (2019). Left: All variants. Right: Non-CSE variants only. The data only includes variants with a wildtype isoform log odds > -2. **C** Comparison of predicted vs measured RNA/DNA log fold change ratios on the data from Slutskin et al. (2019). Individual scatter plots are shown for each tested model. Only variants from the PolyApredictor test set are included ($n = 442$). **D** Comparison of predicted vs measured cleavage site RNA/DNA log fold change ratios on the scanning mutagenesis data from Slutskin et al. (2019). Individual scatter plots are shown for each wildtype UTR. Each data point corresponds to one cleavage position. Total $n = 1,217$.



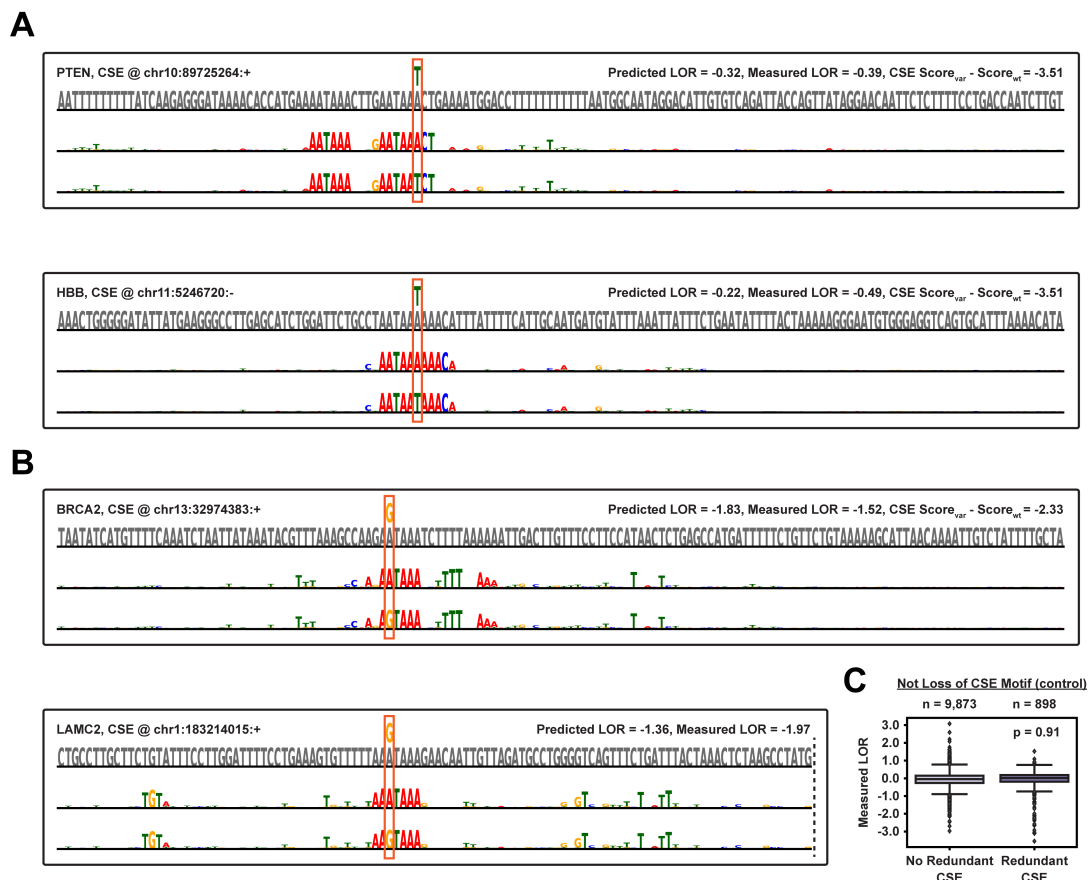


Figure S4: **A** Additional mask-based interpretations of functionally silent CSE mutations in the PTEN- and HBB genes. Annotated on the right are predicted and observed isoform log odds ratios (as measured in an MPRA), as well as CSE hexamer regression scores. **B** Additional interpretations of variants with dampened effect sizes (with respect to a linear hexamer regression model) in the BRCA2- and LAMC2 genes. Predicted and measured log odds ratios on the right. **C** Boxplot showing measured log odds ratios (LORs) from the MPRA of Bogard et al. (2019), for arbitrary mutations that occur in PASs with a single CSE hexamer or in PASs with additional redundant CSE hexamers. The p-value is computed with a two-sided t-test.

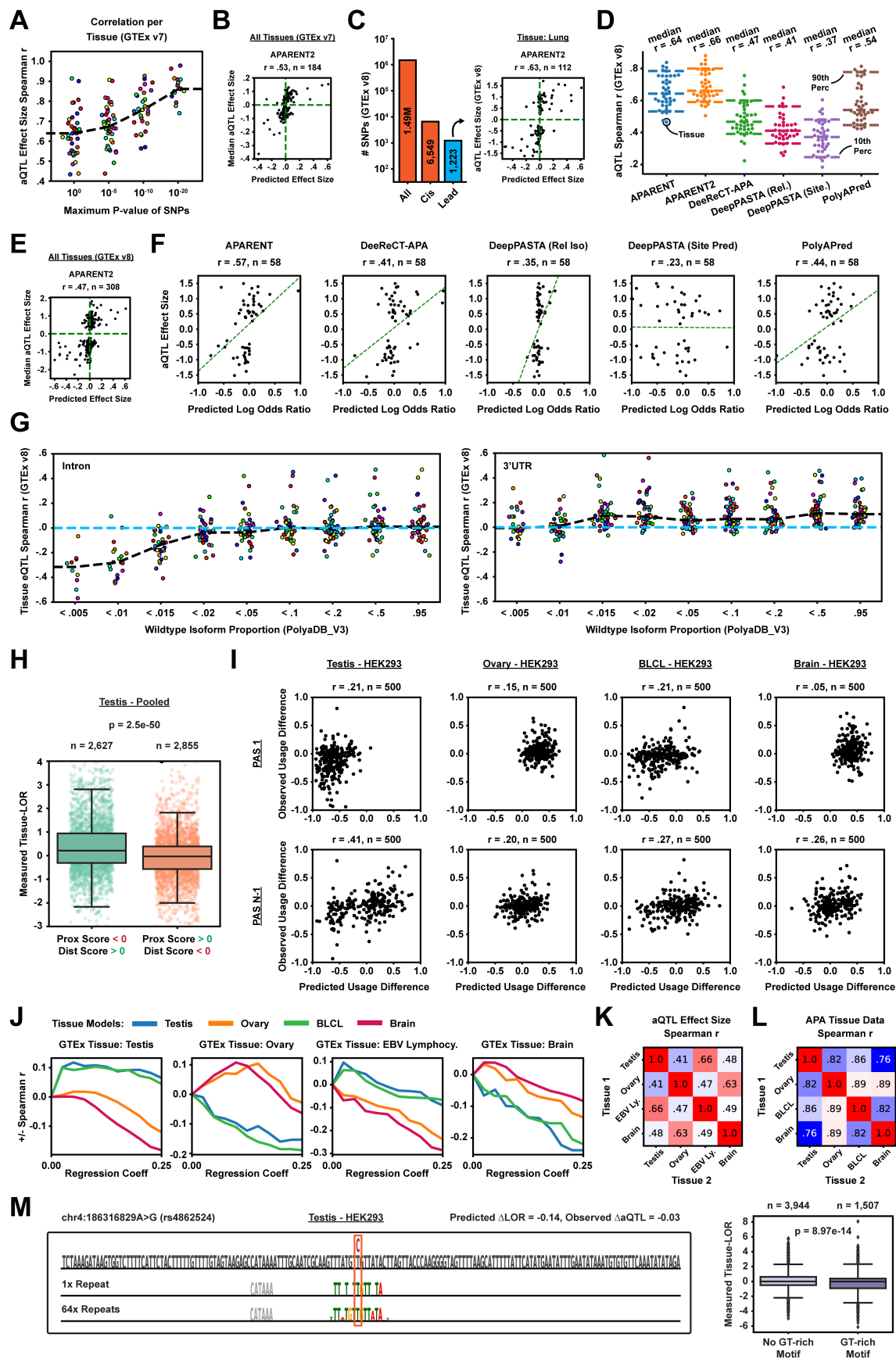
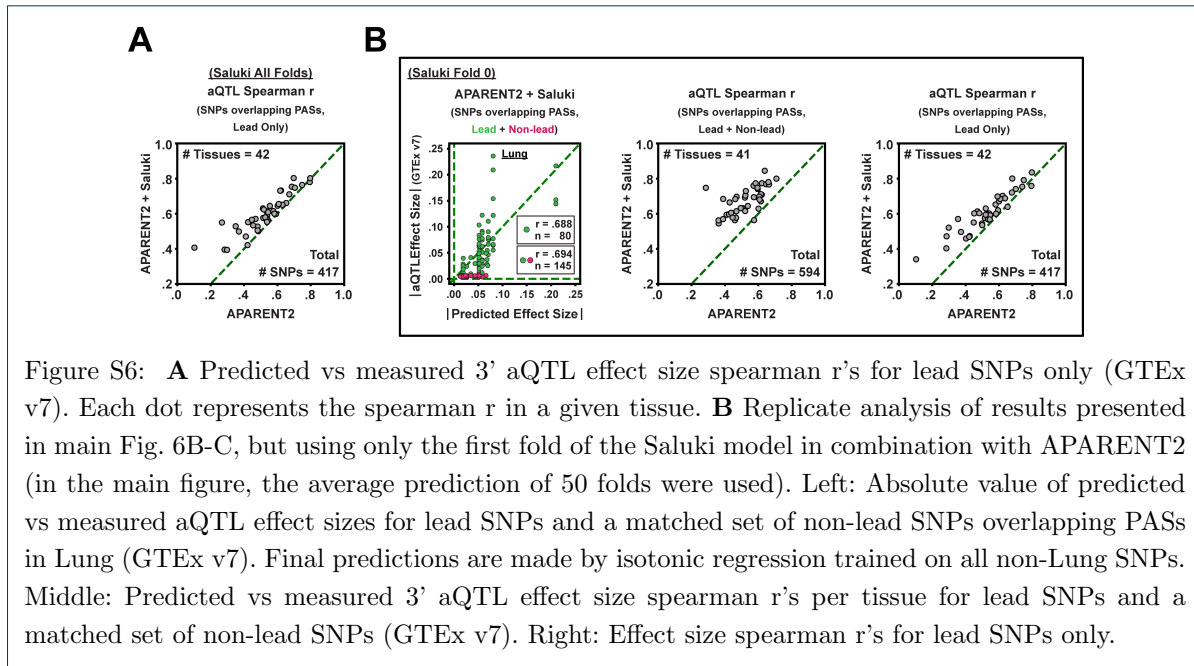


Figure S5: **A** Correlation between predicted isoform log odds ratios (LORs; using APARENT2) and estimated 3' aQTL effect sizes from the GTEx v7 atlas. Each dot corresponds to the spearman r correlation for a particular tissue, after having filtered the set of SNPs to those with a p-value below the cutoff specified by the x-axis. Number of unique lead cis-aQTLs across all tissues = 366. **B** Predicted effect sizes (using APARENT2) vs median 3' aQTL effect sizes taken across all tissues of the GTEx v7 atlas. Only lead SNPs with support in more than 1 tissue are included. **C** Left: Number of 3' aQTLs, cis-acting 3' aQTLs and lead 3' aQTLs respectively in the newer GTEx v8 atlas. Right: Predicted vs measured 3' aQTL effect sizes in Lung (GTEx v8). **D** Predicted vs measured 3' aQTL effect size spearman r 's (GTEx v8). Each dot corresponds to the spearman r for a given APA model in a particular tissue type. The dashed horizontal lines denote the 10th, 50th and 90th percentiles of each distribution. **E** Predicted effect sizes (using APARENT2) vs median 3' aQTL effect sizes taken across all tissues of the GTEx v8 atlas. Only lead SNPs with support in more than 1 tissue are included. **F** Predicted isoform log odds ratio of each model vs estimated aQTL effect sizes of the data from Mittleman et al. (2020) ($n = 58$). **G** Correlation between APARENT2 isoform log odds ratios and estimated eQTL effect sizes for 1,007 intronic GTEx eQTLs and 2,225 3' UTR eQTLs, as a function of wildtype PAS usage as measured in tissue-pooled data from PolyADB v3. **H** Measured difference in isoform log odds between testis and tissue-pooled data from Lianoglou et al. (2013). The left distribution is the subset of proximal PASs with APARENT2 scores <0 and distal APARENT2 scores (in the same gene) >0 . Inversely, the right distribution is the subset of PASs with proximal APARENT2 scores >0 and distal scores <0 . The p-value is computed with a two-sided t-test. **I** Predicted vs measured tissue-specific difference in isoform proportion between HEK293 and the target tissue, on a held-out test set of $n = 500$ genes. Results are shown for either the most proximal, or next-to-last, PAS of each gene. **J** $+/-$ Spearman r correlation w.r.t baseline APARENT2 predictions of GTEx aQTL effect sizes, separated by GTEx tissue type, as a function of regression coefficient γ (see Methods) for each tissue-specific model. **K** Correlation of 3' aQTL effect sizes between tissues (GTEx v7). **L** Correlation (spearman r) of PAS usage between tissues in data from Lianoglou et al. (2013) ($n = 6,440$ genes). **M** Reconstructive interpretation mask for a GTEx SNP (rs4862524) with a weak tissue-specific effect in Testis. Boxplot shows differential PAS usage (difference in isoform log odds). The p-value is computed with a two-sided t-test.



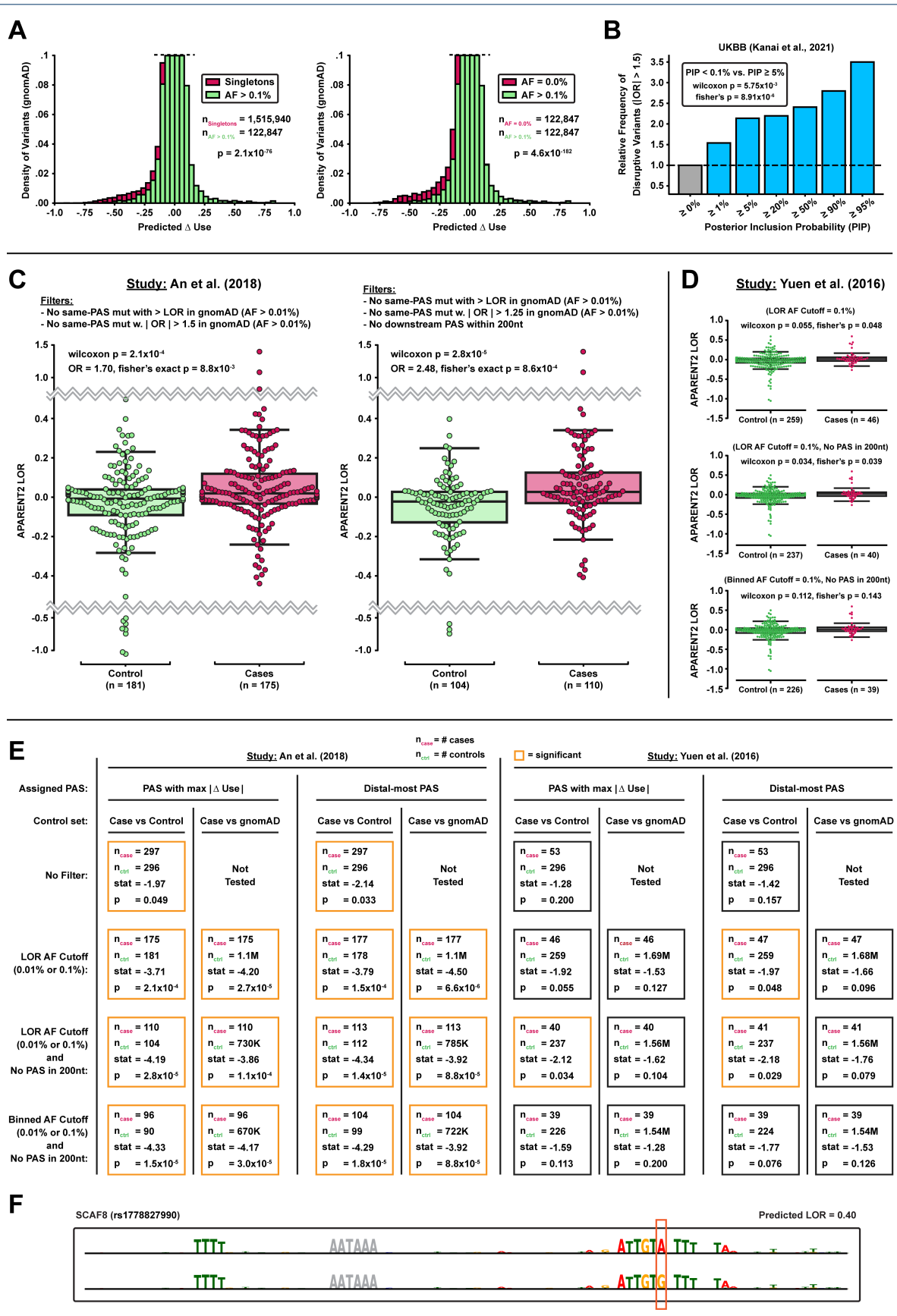


Figure S7: **A** Left: Predicted Δ isoform proportions for singletons ($n = 1,515,940$) and common variants ($AF > 0.1\%$; $n = 122,847$) from gnomAD v3 that overlap annotated PASs in PolyADB V3. Right: Comparison of predictions for a matched set of unobserved PAS variants ($AF = 0.0\%$; $n = 122,847$) and common variants ($AF > 0.1\%$) from gnomAD. **B** Relative frequency (enrichment) of fine-mapped SNPs from UK Biobank that overlap annotated PASs ($n = 90,356$) with a predicted absolute-valued isoform odds ratio > 1.5 , as a function of more stringent posterior inclusion probabilities (PIP). **C** Distribution of predicted isoform log odds ratios among ASD cases and controls from the WGS study of An et al. (2018). Left: Case- and control variants are removed if they occur in PASs that have common variants in gnomAD ($AF > 0.01\%$) with larger effect size (log odds ratio) than the investigated variant or common variants that have an absolute odds ratio larger than 1.5 ($n_{\text{control}} = 181$, $n_{\text{cases}} = 175$). Right: Additional removal of variants that occur in PASs with a downstream PAS within 200nt in PolyADB V3 ($n_{\text{control}} = 104$, $n_{\text{cases}} = 110$). A more stringent odds ratio threshold of 1.25 was used. This is the same filtering procedure as in main Fig. 7G, but here the allele count of variants within the same PAS with similar effect sizes have not been aggregated prior to filtering against gnomAD (see Methods for details on the variant filtering procedure). **D** Replicate analysis where the case variants are from the WGS study of 200 families from Yuen et al. (2016) (the control variants come from An et al., 2018). The three filtering steps used in supplementary Fig. S7C and main Fig. 7G are also used here: (1) filtering variants with neighboring common mutations in gnomAD ($AF > 0.1\%$; $n_{\text{control}} = 259$, $n_{\text{cases}} = 46$), (2) additionally removing variants with downstream protective PASs within 200nt ($n_{\text{control}} = 237$, $n_{\text{cases}} = 40$), and (3) same filtering criteria as (2) but gnomAD AFs are re-calculated by aggregating allele counts of similar predicted effect size in the same PAS ($n_{\text{control}} = 226$, $n_{\text{cases}} = 39$). **E** Summary of statistical tests performed on the cohort data from An et al. (2018) (left table) and Yuen et al. (2016) (right table). P-values were calculated using Wilcoxon rank-sum tests. The rows denote different filtering criteria and the columns denote the control group used (controls either come from An et al. or from an identically filtered view of gnomAD). Two different methods were tested for assigning variants to PASs (when variants overlap two nearby PASs): (1) the PAS resulting in largest predicted variant effect size or (2) the distal-most PAS. Significant tests are marked with an orange border. **F** Mask-based interpretation of an ASD-associated PAS mutation rs1778827990 in SCAF8.

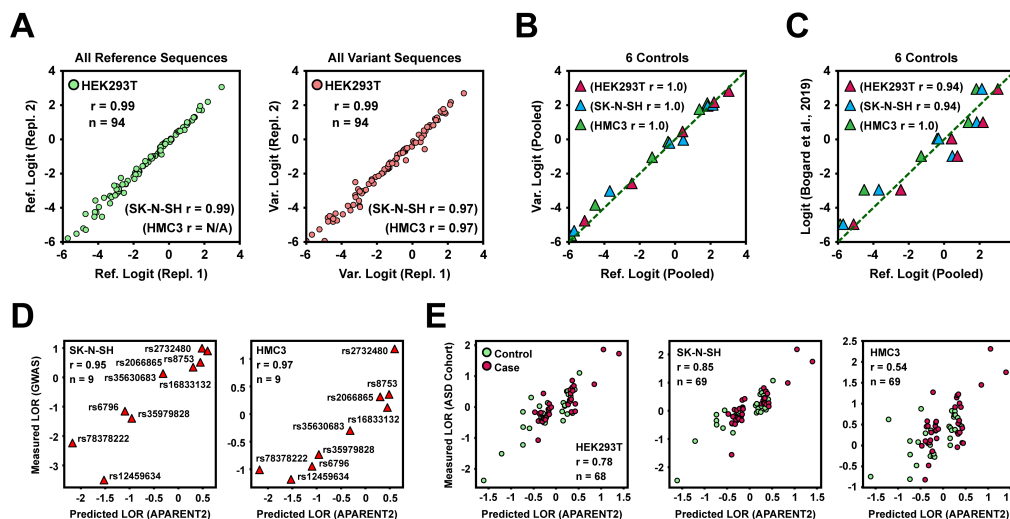


Figure S8: **A** Replicate correlation of proximal isoform logits as measured in the plasmid reporter MPRA for the subset of the 100 assayed PASs that had a minimum of 5 supporting reads. Left: Measured logits of reference (wildtype) PASs. Right: Measured logits of variant (alternate) PASs. The scatter plots show the results for HEK293T, while summary correlation metrics for SK-N-SH and HMC3 are annotated in the plots. HMC3 had only one biological replicate with sufficient read depth for the reference library. **B** Proximal isoform logits of 6 control PASs that were assayed (without any designed mutations) in both the reference and variant libraries. The isoform logits were estimated by pooling the counts across replicates. **C** Correlation between the 6 control PASs of the reference library and their estimates from the MPRA of Bogard et al. (2019) where they had previously been measured. **D** Predicted vs measured variant effect sizes (log odds ratios, *LORs*) of 9 assayed GWAS SNPs. Results are shown for SK-N-SH and HMC3. **E** Predicted vs measured variant *LORs* of the subset of assayed Autism variants with a minimum of 50 reads (pooled replicates). Green = control variants, Magenta = case variants.

