Title

Sequential regulatory activity prediction across chromosomes with convolutional neural networks.

Authors

David R. Kelley, Yakir A. Reshef.

1. Calico Labs. South San Francisco, CA, USA.
2. Department of Computer Science. Harvard University. Cambridge, MA, USA.

Correspondence to: drk@calicolabs.com

Abstract

Functional genomics approaches to better model genotype-phenotype relationships have important applications toward understanding genomic function and improving human health. In particular, thousands of noncoding loci associated with diseases and physical traits lack mechanistic explanation. Here, we develop the first machine-learning system to predict cell type-specific epigenetic and transcriptional profiles in large mammalian genomes from DNA sequence alone. Using convolutional neural networks, this system identifies promoters and distal regulatory elements and synthesizes their content to make effective gene expression predictions. We show that model predictions for the influence of genomic variants on gene expression align well to causal variants underlying eQTLs in human populations and can be useful for generating mechanistic hypotheses to enable GWAS loci fine mapping.

Introduction

Although many studies show strong relationships between variation in genotype and phenotype across a range of human diseases and traits, the mechanisms through which this relationship operates remain incompletely understood. Noncoding variation has especially stifled progress; most genomic loci statistically associated with phenotypes via genome-wide association studies (GWAS) do not alter coding sequence, but mechanisms for only a rare few have been thoroughly described. Numerous lines of evidence suggest that many noncoding variants influence traits by changing gene expression [1-3]. In turn, gene expression determines the diversity of cell types and states in multi-cellular organisms [4]. Thus, gene expression offers a tractable intermediate phenotype for which improved modeling would have great value.

Large-scale consortia and many individual labs have mapped the epigenetic and transcriptional profiles of a wide variety of cells [4-6]. Further, it has recently become appreciated that many of these data can be accurately modeled as a function of underlying DNA sequence using machine learning. Successful predictive modeling of transcription factor (TF) binding, accessible chromatin, and histone modifications has provided mechanistic insight and useful interpretation of genomic variants [7-11]. In particular, the substantial training data available from the 3 billion-nucleotide human genome has enabled deep learning approaches to achieve significant gains [10,11].

Despite this progress, models to predict cell type-specific gene expression from DNA sequence have remained elusive in complex organisms. Existing models all use experimental annotations as input (e.g. peak calls for various known regulatory attributes), allowing them to shed light on the relationships between these annotations, but disabling their application to regulatory variant interpretation [12,13]. Even with intra-experiment training data to infer the relevant sequence motifs, the complexity of distal regulation across up to millions of nucleotides challenges the current generation of methods [14]. However, a solidifying base of gene regulation principles from inquiry into enhancer biology and 3D chromosomal contact domains has yet to be fully incorporated into expressive machine learning models [15]. Considering larger sequences and cues from diverse experimental data offers a path forward. More effective models would enable researchers to profile one instance of a tissue or cell type and project that profile to individuals with varying genomic sequence.

Here, we leverage thousands of epigenetic and transcriptional profiles from hundreds of human cell types and novel machine-learning algorithms to provide comprehensive models of transcription as a function solely of DNA sequence. From intra-experiment training data and DNA sequence, we can predict mRNA expression, as measured by capped analysis of gene expression (CAGE), with concordance matching the acquisition of a replicate experiment. Using the model, we predict the difference between the two alleles of genomic variants, focusing particularly on predicted changes to gene expression. These predictions align well with magnitudes of effect reported in expression QTL studies performed in human populations. We demonstrate the considerable potential value of this observation to identify likely causal variants and mechanisms within GWAS loci.

Basenji

In previous work, we introduced a deep convolutional neural network approach named Basset for modeling “peak”-based chromatin profiles, focusing particularly on DNase hypersensitivity [11]. That model makes a single binary prediction for a given input sequence of 500-1000 bp for each training dataset provided. Here, we modify the Basset architecture to (1) model distal regulatory interactions and (2) predict finer resolution, quantitative (as opposed to binary) genomic profiles that are more appropriate for the dynamic range of gene expression (Figure 1). As a related approach, but one begging for metaphor to a more nimble and far-sighted hound, we refer to this method as Basenji.

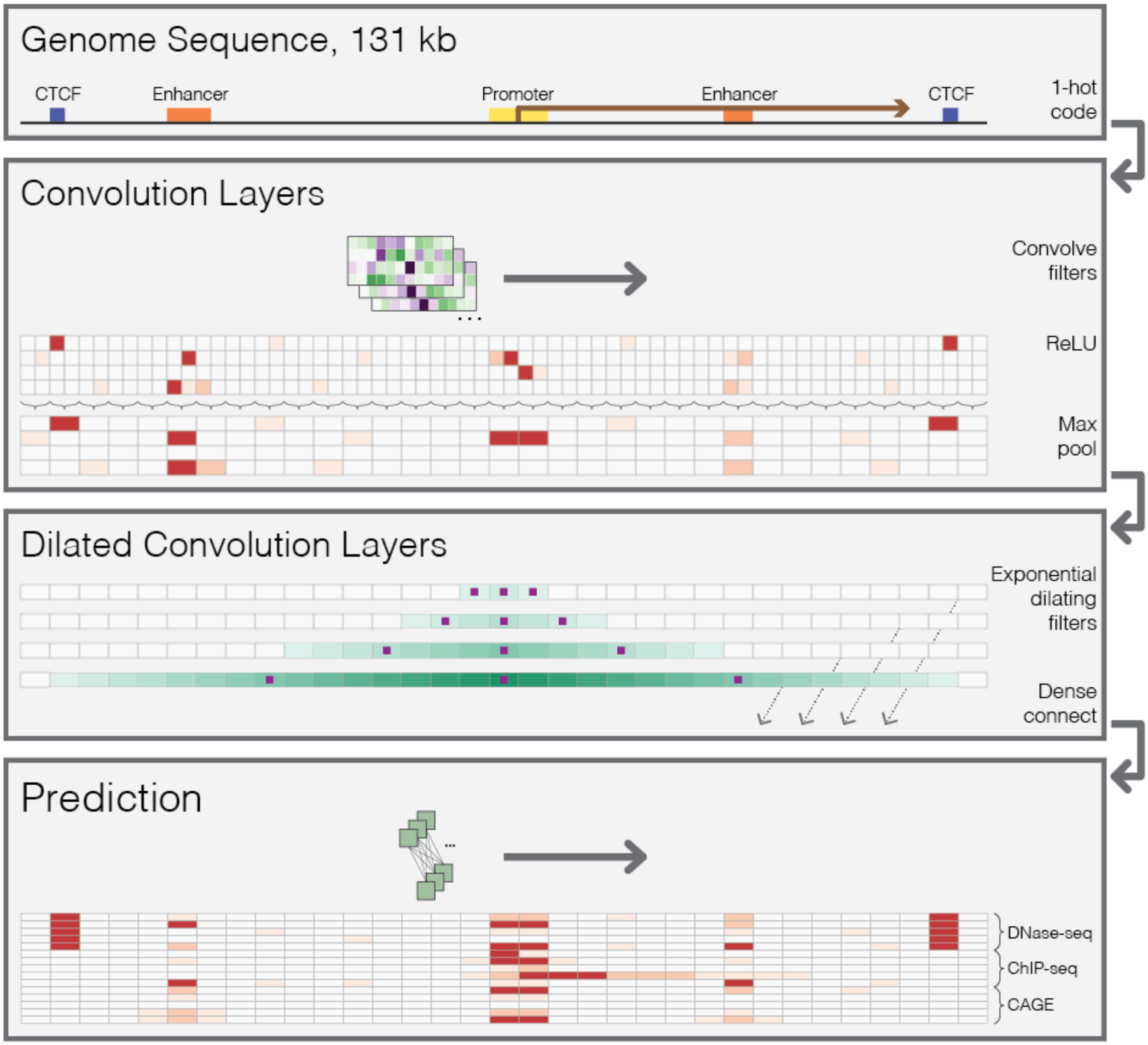


Figure 1 - Sequential regulatory activity prediction.

DNA sequences come in to the model one hot coded to four rows representing A, C, G, and T. The annotations are fabrications to help convey the reasons for the various elements of the architecture. We apply several layers of convolution and max pooling, similarly to previous methods [11], to obtain representations that describe 128 bp bins. To share information across large distances, we apply several layers of dilated convolutions. The purple squares indicate the columns that the convolution directly sees; the teal shade is drawn proportional to the number of operations performed on that column with respect to the center position. Dilated convolution layers are densely passed on to the final prediction layer, where a shared fully connected neural network makes predictions across the sequence. We compare these predictions to the experimental counts via a Poisson regression loss function and use stochastic gradient descent with back propagation to fit the model parameters.

The model accepts much larger (2^{17}=)131 kb regions as input and, similarly to Basset, performs multiple layers of convolution and pooling, to condense the DNA sequence to a sequence of vector representations for 128 bp regions. To share information across long distances, we then apply several layers of densely connected dilated convolutions (Methods). After these layers, each 128 bp region is represented by a vector that considers the detected regulatory elements across a large span of sequence. Finally, we apply a fully connected neural network layer to parameterize a multi-task Poisson regression on a normalized count of aligned reads to that region for every dataset provided [16]. That is, the model's ultimate goal is to predict read coverage in 128 bp bins across these sequences.

Modeling binned count data as opposed to peak data required careful preprocessing beyond that performed in the standard pipelines of genomics consortium projects. For example, processed consortium data discards multi-mapping read alignments, which leaves half the genome incompletely annotated, despite the substantial evidence that repetitive sequence is critical to gene regulation [17]. Thus, we downloaded the original sequencing reads for 593 ENCODE DNase-seq, 1704 ENCODE histone modification ChIP-seq, 356 Roadmap DNase-seq, 603 Roadmap histone modification ChIP-seq, and 973 FANTOM5 CAGE experiments. We processed these data with a pipeline that includes additional computation to make use of multi-mapping reads and normalize for GC bias (Methods). Though additional data modalities may require slight modification, this base pipeline will allow seamless addition of future data.

We trained to fit the model parameters on one set of genomic sequences annotated by all datasets and benchmarked predictions on those same datasets for a held-out set of sequences. We used a Basenji architecture with 4 standard convolution layers, pooling in between layers by 2, 4, 4, and 4 to a multiplicative total of 128, 7 dilated convolution layers, and a single fully connected layer to predict the 4229 coverage datasets. We optimized all additional hyper-parameters using Bayesian optimization (Methods).

Prediction accuracy

To assess how effectively the model predicts the signal in these datasets, we compared predictions to coverage signal in the 128 bp bins for a set of held-out test sequences (Figure 2A). As previously observed, accuracy varies by the type of data—punctate peak data tend to be more directly dependent on the underlying sequence, making for an easier prediction task (Figure 2B). Accordingly, Basenji predictions explained the most variance in DNase-seq and ChIP-seq for active regulatory regions. Lower accuracy for the broad chromatin domains marked by modifications like H3K79me2 and H3K9me3 is expected because they depend more on distant sequence signals and incompletely understood propagation mechanisms [18].

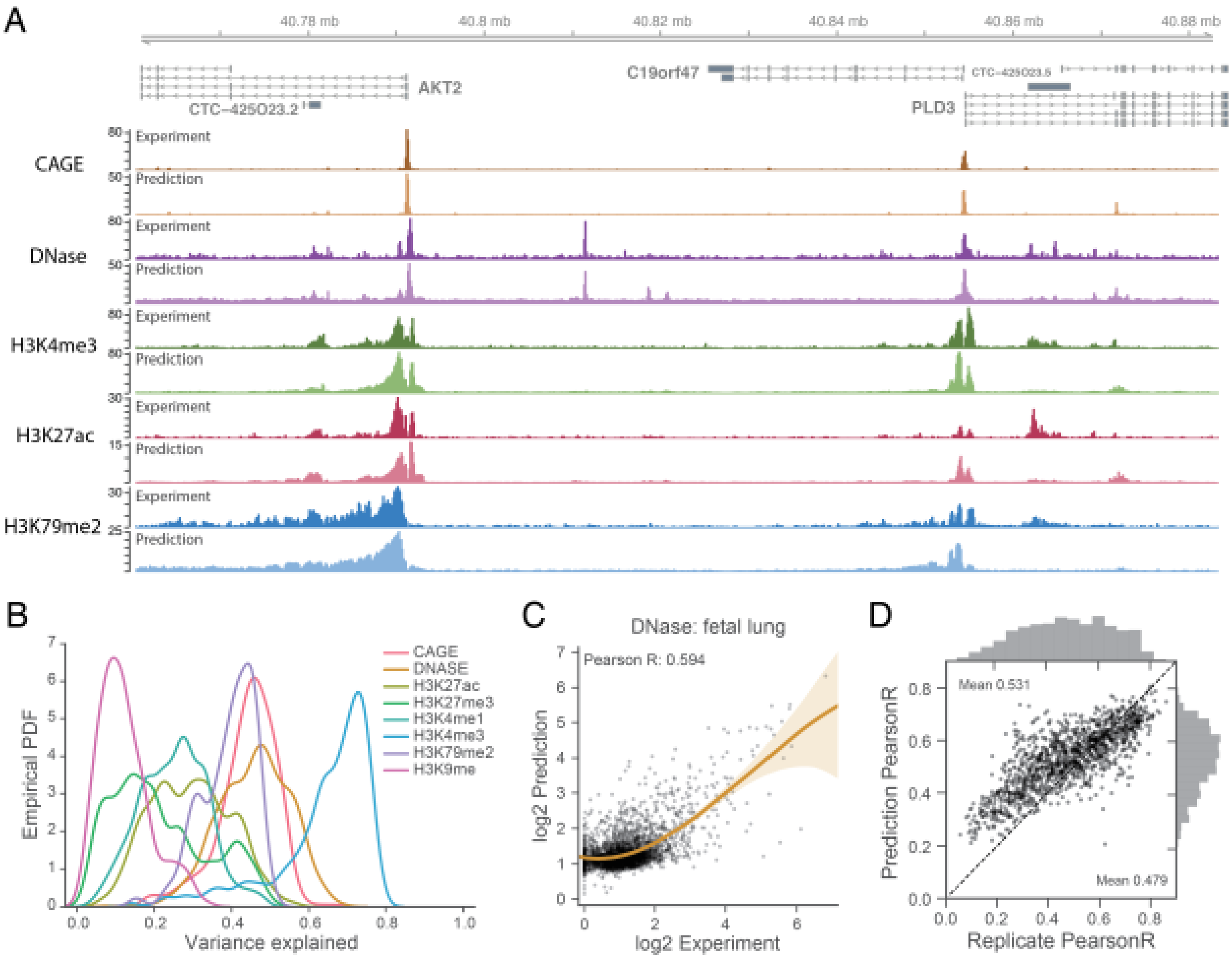


Figure 2 - Basenji predicts diverse epigenetic and transcriptional profiles from DNA sequence.

(A) The AKT2 locus exemplifies the genome-wide accuracy of Basenji predictions; gene promoters and the strongest distal regulatory elements are easily identified, with some false positive and negative predictions for weaker elements. For each track, the darker version on top represents the experimental coverage and the lighter version below represents Basenji predictions. (B) We computed the variance explained (R^2) for each experiment and plot here the distributions classified by dataset type. Basenji predicts punctate peak data, but broad chromatin marks remain challenging. (C) For the median accuracy DNase experiment, fetal lung, we plotted the log2 predictions versus log2 experiment coverage in 128 bp bins. (D) For all replicated experiments, we plotted log-log Pearson correlation between the replicate experiments versus the correlation between the experiment and prediction (averaged across replicates). Basenji predictions exceed the accuracy of a second replicate on average and are more accurate for experiments with less variability between replicates.

The quantitative signal prediction identified peaks called from the original data with a similar level of accuracy as previous approaches (Supplementary Figure 1). The mean AUPRC for DNase peaks was 0.44 across the genome, approaching the 0.56 previously reported for Basset on an easier dataset enriched for active regions and with smoothing across similar experiments [11].

1284 replicated experiments (mostly technical, rather than biological) allowed us to appreciate that model predictions in the 128 bp bins had greater correlation than the signal obtained from a second replicate on average (Figure 2D; Supplementary Figure 2). Even cross-replicate predictions (i.e. the prediction for replicate one versus the real data for replicate two) matched the correlation between replicates (Supplementary Figure 3). Basenji was more accurate for experiments with more correlated replicates, suggesting that higher quality data enables more effective modeling of the sequence dependence of the regulatory signal (Figure 2D). The silencing modifications H3K9me3 and H3K27me3 had low replicate consistency; improved data may lead to better modeling of repressive chromatin in the future (Supplementary Figure 2,3).

We hypothesized that the 7 dilated convolution layers enabled the model to capture the distal influences that are an established feature of human gene regulation. To isolate the influence of receptive field width, we trained similar models with 1-7 dilated layers. Test accuracy increased with increasing receptive field for all data types, confirming the value added by the additional dilated convolution layers of the final network (Supplementary Figure 4).

Cell type-specific gene expression

A driving goal of this research is to effectively model cell type-specific gene expression. CAGE quantifies gene expression by capturing and sequencing 5' capped mRNAs to measure activity from genes' various start sites [6]. To offer a gene-centric view from Basenji predictions, we focused on annotated TSSs and computed CAGE accuracy for those outside the training set. After log2 transform, the mean Pearson correlation of gene predictions with experimental measurements across cell types was 0.84 overall and 0.75 for nonzero genes (Figure 3A,B), which is on par with models that consider measurements of the relevant regulatory events with sequencing assays rather than predicting them from sequence [13]. Correlation is greater for CAGE datasets with more reads aligned to TSSs due to greater sequencing depth and signal to noise ratio (Figure 3A), suggesting that the lower half of the datasets are constrained more by sparse, noisy signal rather than algorithm learning capacity.

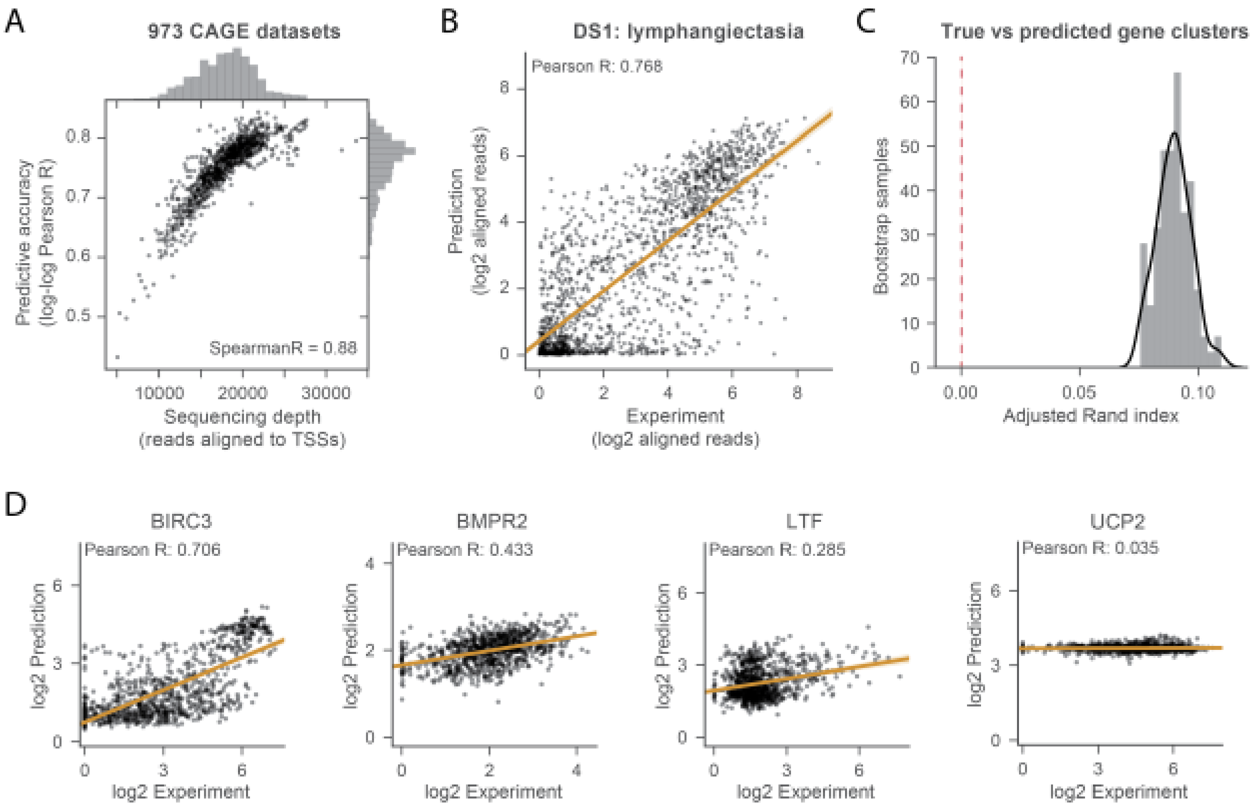


Figure 3 - Basenji predicts cell type-specific gene transcription.

(A) We computed Pearson correlation between the log2 prediction and experiment across all test set genes for each CAGE dataset. We plotted those correlations against the total number of reads aligned to test gene TSSs, which measures the relevant sequencing depth. (B) For the median accuracy cell, the DS1 lymphangiectasia cell line, we plotted the experiment coverage versus Basenji prediction. (C) For both the experimental measurement and Basenji prediction, the gene expression by CAGE dataset matrix displays clusters. We measured the similarity of those clusters between the real and predicted data by bootstrap sampling gene subsets, clustering both the real and predicted data, and computing the adjusted Rand index between the cluster sets (Methods). The adjusted Rand index is significantly greater than the null model value zero (p-value < 1e-26). (D) We plotted gene expression versus prediction after quantile normalization across cell types for the genes ranked in the 95th, 75th, 50th, and 25th percentiles by Pearson correlation.

Predictions varied across cell types, suggesting that the model learns cell type-specific transcriptional programs (Supplementary File 5). For these experiments, we normalized the predictions and real data across experiments with quantile normalization. Substituting the predictions into a hierarchical clustering of the true expression maintained apparent clusters. To quantify this concordance, we performed Gaussian mixture model clustering of bootstrap gene samples for the real and predicted expression profiles. The adjusted Rand index distribution (mean

0.087) indicated significant agreement between clusters (p-value < 1e-18).

To quantify the greater difficulty of predicting highly cell type-specific expression, we computed the mean squared prediction error for sets of genes binned into quartiles by their coefficient of variation across all CAGE experiments. The stable expression across cell types of housekeeping genes relies on a particular promoter architecture [19], which the model learns well; accordingly, predictions are closer to their measured values in the genes most stable across experiments (Supplementary Figure 6). Beyond this first quartile, increasing variability does not weaken prediction accuracy, further supporting the view that the model has learned cell type-specific regulatory programs.

We found it instructive to closely examine genes with variable expression patterns. We computed accuracy statistics independently for each gene on their vectors of quantile normalized predictions and experimental measurements across cell types. In Figure 3D, we display genes at the 95th, 75th, 50th, and 25th percentiles ranked by correlation. Instances of effective predictions across several orders of magnitude, such as for *BIRC3* with its greatest expression in the small intestine, stomach, and spleen, lend credence to the model. In many cases, Basenji has learned that the gene's expression varies across cells, but underestimates the dynamic range of the variance. For example, *BMPR2* and *LTF* predictions correlate with the experimental measurement, but compress the range between the most and least expressed cells. Some degree of variance reduction is warranted because the CAGE measurement includes stochastic noise that Basenji will implicitly smooth out. However, poor prediction of some genes, such as mitochondrial protein *UCP2* indicate that an inability to capture more complex regulation [20] likely also dampens prediction confidence.

Distal regulatory elements

To further explore the role of distal sequence in the model's predictions, we computed saliency maps for the regions surrounding TSSs to quantify the influence of genome segments. Briefly, the saliency scores depend on the magnitude of the gradient of the model's prediction at that TSS with respect to each of the 128 bp segments that arise after the convolutional layers and before the dilated convolutions share information across wider distances (Methods). Peaks in this saliency score detect distal regulatory elements, and its sign indicates enhancing versus silencing influence.

The region surrounding *PIM1* in the GM12878 cell line exemplifies this approach (Figure 4A). The promoter region is highly influential, including repressive segments; mutating the driver motifs in these regions would increase the prediction. More distant elements are also captured; we identified two enhancers annotated by ENCODE, one with a panoply of motifs highlighted by POU2F and another with two adjacent PU.1 motifs. We searched for perturbation data to support these regulatory interactions. In a collection of 59 siRNA TF knockdowns performed in a similar lymphoblastoid cell line GM19238, POU2F1 and POU2F2 knockdowns resulted in differential expression of PIM1 mRNA with p-values 0.026 and 0.066 respectively [21]. PU.1 could not be depleted sufficiently.

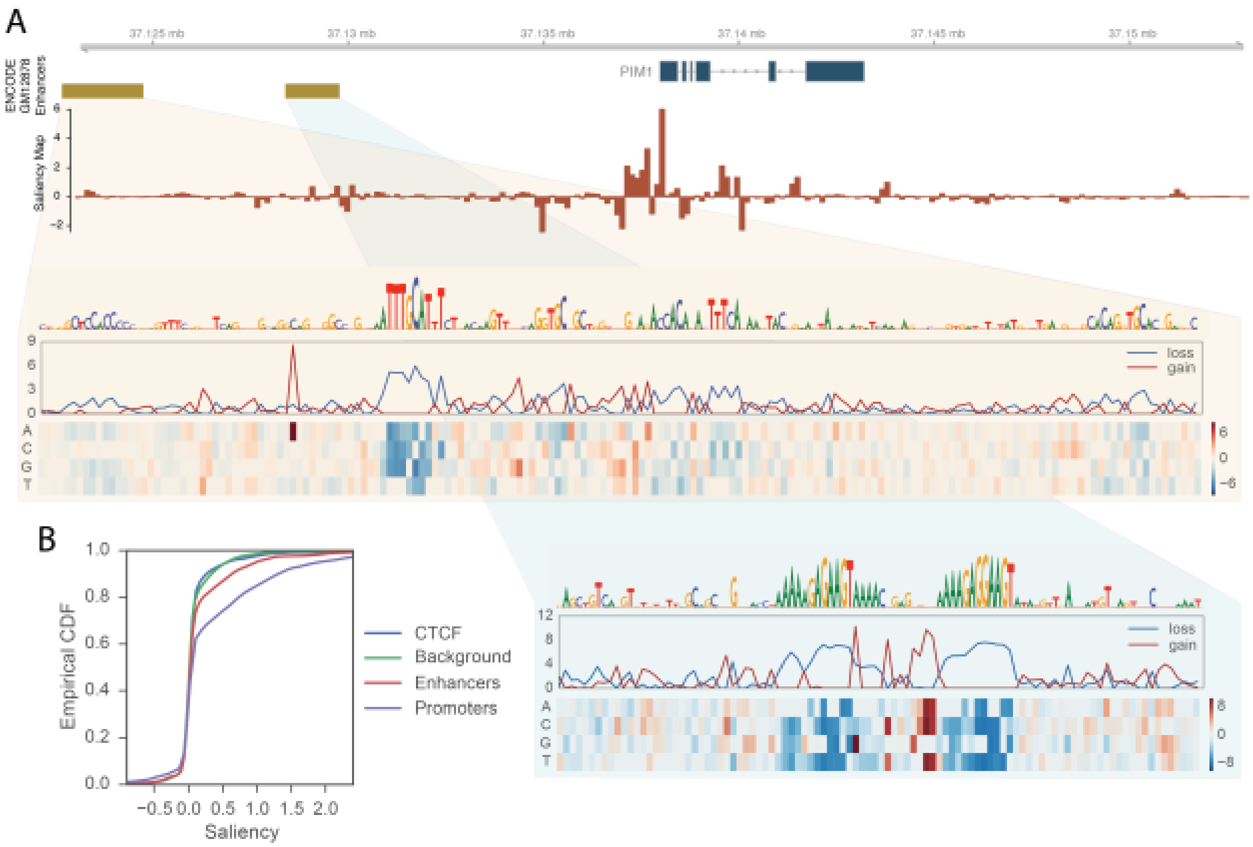


Figure 4 - Basenji identifies distal regulatory elements.

(A) ENCODE enhancer annotations for *PIM1* in GM12878 specify two downstream regulatory elements. Basenji saliency scores (see Methods) mark these elements, in addition to a variety of others that lack typical enhancer chromatin. In silico saturation mutagenesis of these elements with respect to Basenji's *PIM1* GM12878 CAGE prediction outline the driving motifs. The upstream cis-regulatory module most prominently features a POU2F factor motif, while the downstream element consists solely of two adjacent PU.1 motifs. (B) We plotted the cumulative distributions of the maximum saliency score for elements of various annotation classes in GM12878. Genome-wide, promoter and enhancer annotations have greater saliency scores than null sequence. CTCF binding sites outside of those regulatory elements appear to not be taken advantage of.

To assess this method's ability to detect such elements genome-wide, we downloaded several curated annotations from ENCODE for GM12878—promoters, enhancers (not overlapping promoters), and CTCF binding sites (not overlapping promoters or enhancers) [5]. We computed the maximum saliency value overlapping instances of these annotations and shuffled background sets. Saliencies for promoters and enhancers were significantly greater than those for the background set (Mann-Whitney U test p-values 9e-22 and 4e-7 respectively), but CTCF sites were not (Figure 4B). The established role of CTCF in distal regulation suggests significant potential in future work to more effectively model the contribution of these elements.

Expression QTLs

Functional profiling of genotyped individuals is widely used to detect influential genomic variation in populations. The Gene-Tissue Expression (GTEx) project offers one such data collection, having measured RNA abundance via RNA-seq in 44 separate human tissues post-mortem and searched for genomic variants significantly correlated with gene expression (eQTLs) [22]. Without observing such data, a trained Basenji model can be used to predict which single nucleotide polymorphisms (SNPs) are eQTLs by comparing model output for the different SNP alleles. To benchmark this approach, we downloaded the GTEx V6p release and focused on 19 tissues that were reasonable semantic matches for FANTOM5 CAGE profiles.

Given a SNP-gene pair, we define its SNP expression difference (SED) score as the difference between the predicted CAGE coverage at that gene's TSS (or summed across multiple alternative TSS) for the two alleles (Figure 5A). Linkage disequilibrium (LD) complicates the comparison to eQTL statistics; marginal associations and significance calls depend on correlated variants in addition to the measured variant, and association scans are better powered for variants that tag more genetic variation [1]. To put SED on a level plane with the eQTL statistics, we distributed the SED scores according to variant correlations to form a signed LD profile of our SED scores, here denoted SED-LD (Methods) [55].

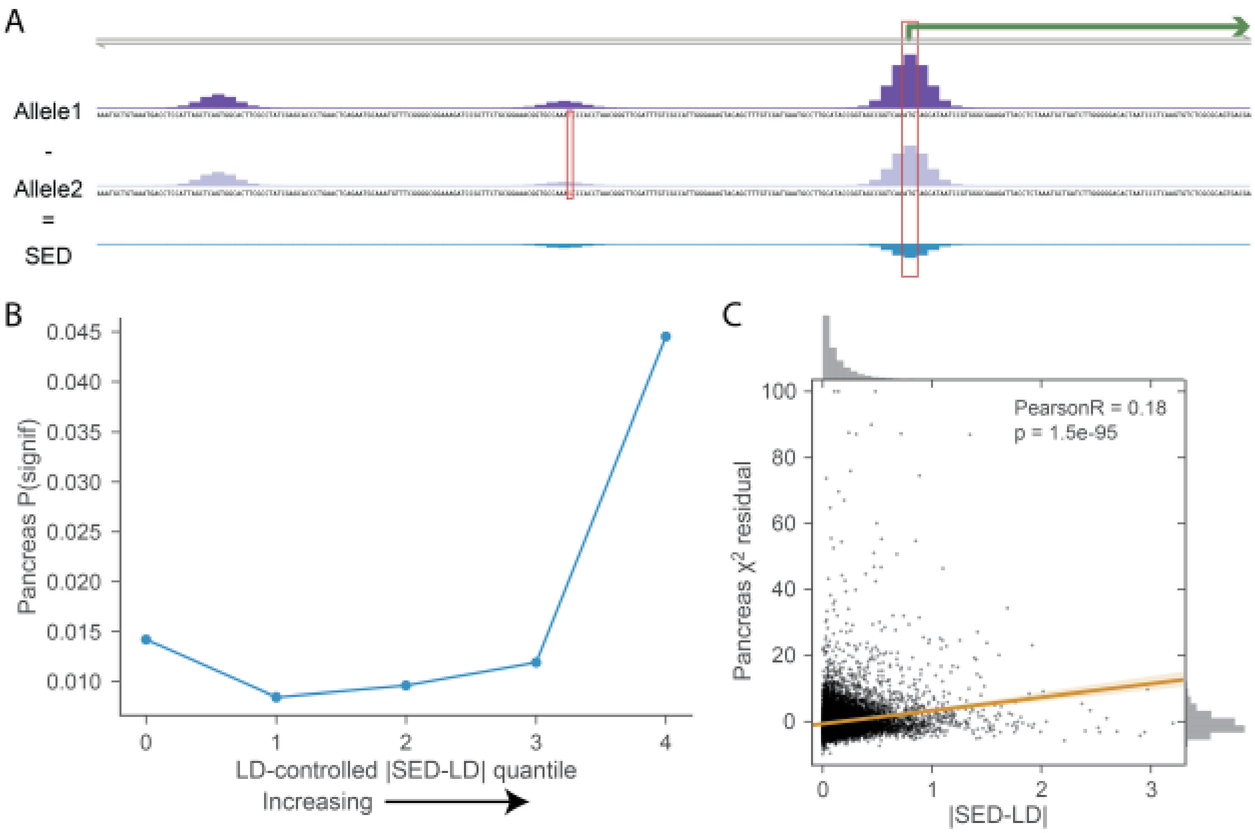


Figure 5 - Basenji gene-specific variant scores enrich for eQTLs.
(A) We defined SNP expression difference (SED) scores for each bi-allelic variant and gene combination as the difference between the model prediction for the two alleles at that gene's TSSs. (B) We computed the signed LD profile of the SED annotations (denoted by SED-LD) to more readily compare to eQTL measurements in human populations (Methods). |SED-LD| shows a strong relationship with eQTL statistics from GTEx. Here, we binned variants into five quantiles by the difference between their regression predictions including and excluding |SED-LD| and plotted the proportion of variants called significant eQTLs in pancreas. The proportion rises with greater |SED-LD| to 4.1x in the highest quantile over the average of the bottom three quantiles, which represented the median enrichment in a range of 3-7x across the 19 tissues. See Supplementary Figure 7 for all tissues and TSS-controlled analysis. (C) Plotting |SED-LD| versus the chi-squared statistics reveals a highly significant correlation.

We checked whether the absolute value of SED-LD correlated with eQTL chi-squared statistics after controlling for the total amount of genetic variation tagged by each SNP as measured by LD score [23] on a set of LD-pruned variants (Methods). Indeed, |SED-LD| significantly correlated with the adjusted eQTL statistics in all 19 tissues (p-values all <1e-54 using LD-pruned variants from chr1). To assess the quantitative extent of this enrichment, we ranked variants by the difference between their predictions from regression models including and excluding |SED-LD| for each tissue and binned into five quantiles. The proportion of variants called significant eQTLs was 3-7x greater in the top quantile relative to the average of the bottom three in all tissues (Figure 5B). This effect was robust to controlling for distance to TSS (Supplementary Figure 7). Thus, our analyses support a robust predictive relationship between Basenji scores and population measurements of RNA abundance, despite the additional layers of post-transcriptional regulation captured by the eQTL analysis and presently invisible to Basenji.

Disease-associated loci

Basenji's utility for analyzing human variation goes beyond intermediate molecular phenotypes like eQTLs to downstream ones like physical traits and disease. Basenji also offers substantial upside—eQTL analysis is highly informative for disease variant interpretation, but few cell types can plausibly be sampled to conduct such an investigation. With Basenji, a single experiment is sufficient to predict a genomic variant's influence on gene expression in that cell type. We hypothesized that a predictive view of the 973 human samples profiled by CAGE would offer a novel perspective on disease variants.

To test the utility of Basenji SNP scores for this application, we acquired a curated set of disease variants studied by the successful DeepSEA method to predict variant influence on TF binding and chromatin [10]. DeepSEA trained deep convolutional neural networks to predict ENCODE and Roadmap DNase and ChIP-seq peak calls. The set includes 12,296 bi-allelic SNPs taken from the NIH GWAS Catalog database [24] and a negative set with matched minor allele frequencies that we sampled down to the same size. We followed the DeepSEA authors' approach of ignoring linkage disequilibrium in order to compare fairly. For each SNP, we computed the log2 ratio between the predictions for the two alleles in each 128 bp bin across the surrounding region. We assigned the SNP its maximum absolute value of this ratio. 200 principal components were sufficient to represent the full profile well. A logistic regression model to predict GWAS catalog presence using the Basenji principal component features achieved 0.661 AUROC, slightly greater than the .658 achieved by DeepSEA using a more sophisticated model that also included conservation statistics (Supplementary Figure 8). Adding DeepSEA's predictions as a feature to our logistic regression model increased accuracy to 0.7045, confirming the value of our predictions.

Having established meaningful signal in the predictions, we analyzed a set of 1170 loci associated with immune phenotypes and processed using the linkage-based statistical fine mapping approach PICS [25]. 67 loci contained a variant predicted to alter a gene's transcription in one of the CAGE experiments by >10%, and an additional 73 contained a variant predicted to alter one of the chromatin profiles >10% at a gene's start sites. rs78461372, associated with multiple sclerosis via linkage with the lead variant rs74796499 [26], emerged from this analysis. Basenji predicts the C>G at rs78461372 to increase transcription of the nearby *GPR65* in many cells, most severely acute lymphoblastic leukemia cell lines, thyroid, insular cortex, and a variety of immune cells. *GPR65* is a receptor for the glycosphingolipid psychosine and may have a role in activation-induced cell death or differentiation of T-cells [27]. Without mention of *GPR65* in the literature, sphingolipid metabolism has emerged as a therapeutic target for MS via the drug fingolimod, a sphingosine analogue that alters immune cell trafficking and is now in clinical use [28]. The model also predicts a small increase for *GALC*, 12.9 kb away, in many immune cells. Both genes have been implicated in several immune diseases (inflammatory bowel disease, Crohn's disease, ulcerative colitis) via variants independent of this set [29], and both genes may propagate a downstream causal influence on the disease.

PICS fine mapping assigns rs78461372 a 5% probability of causal association with multiple sclerosis and the leading variant rs74796499 a 24% probability. Basenji predicts no effect for rs74796499 or any other variants in the PICS credible set. To validate the predicted stronger effect of rs78461372 on nearby transcription, we consulted the GTEx multiple tissue eQTL analysis [22]. GTEx supports Basenji's diagnosis, detecting significantly increased *GPR65* expression for individuals with the minor allele at rs78461372 in transformed fibroblasts (marginal beta 0.75; p-value 1.8e-9); the competing correlated variant rs74796499 has a smaller measured effect (marginal beta 0.66; p-value 4.9e-7).

To better understand the model prediction, we performed an in silico saturation mutagenesis [11,30] in several affected cell types. That is, we generated sequences that introduce every possible mutation at all sites in the region, predicted CAGE activity, and computed the difference from the reference prediction. The functional motifs that drive the model's prediction emerge as consecutive sites where mutations result in large differences. rs78461372 overlaps an ETS factor motif adjacent to a RUNX factor motif, where the G allele confers a stronger hit to the JASPAR database PWM for ETS, discovered using the motif search tool Tomtom [31,32] (Figure 6). Interestingly, Basenji predicted opposite effects for these motifs in different cell types. In immune cells, where *GPR65* and *GALC* are more active, disruption of the motifs would result in a decreased prediction. Alternatively, in e.g. insular cortex, disruption of the same motifs would increase the model's predictions. Altogether, Basenji predictions shed substantial light on this complex and influential locus, offering several promising directions for future work to unravel the causal mechanism.

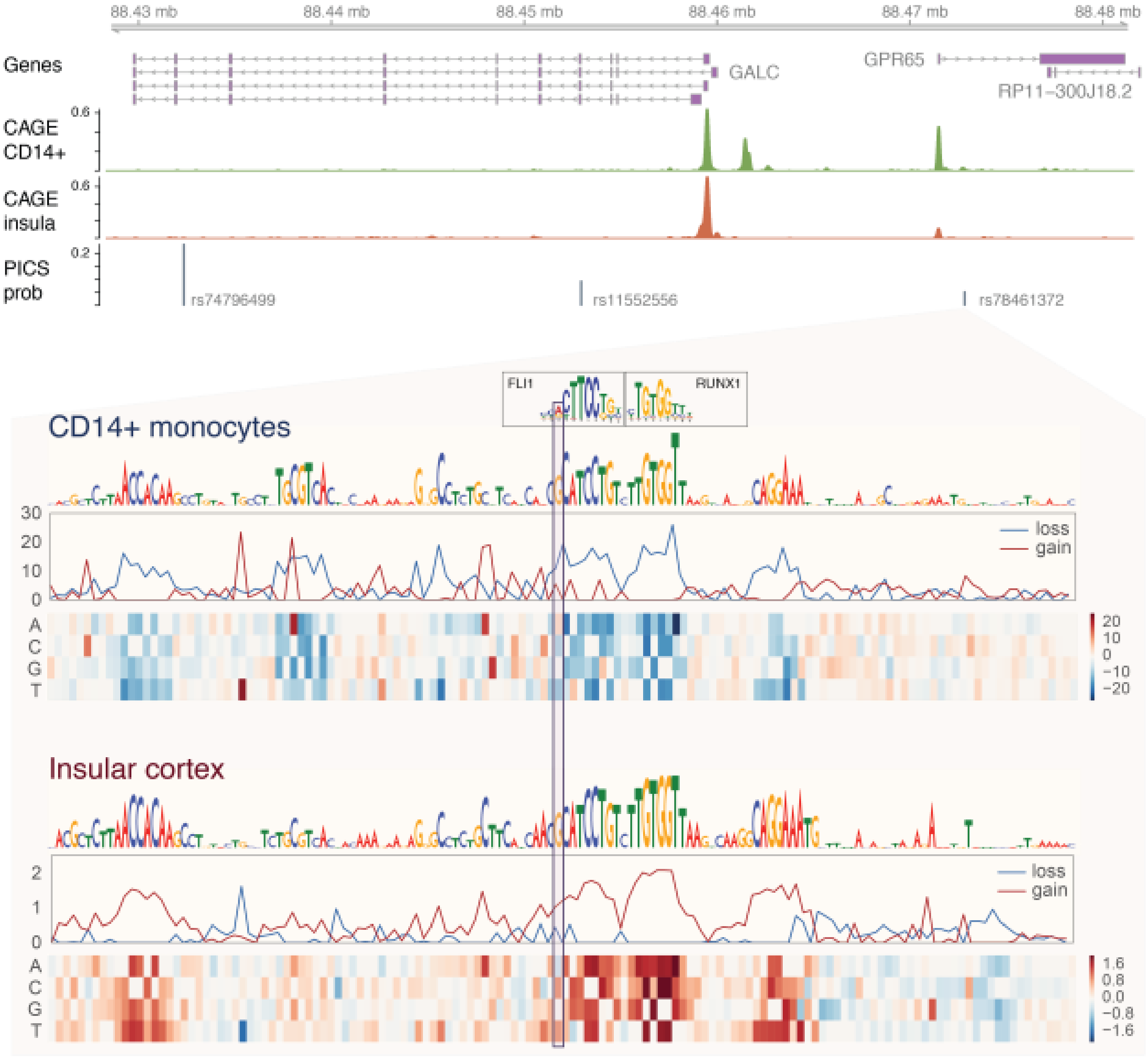


Figure 6 - Basenji gene-specific variant scores illuminate multiple sclerosis associated locus.
Lead variant rs74796499 is associated with multiple sclerosis [53]. Among the credible set of linked variants, Basenji predicts that rs78461372 would alter transcription of the nearby genes GPR65 and GALC. In immune cells, such as treated CD14+ cells depicted here, both genes are transcribed and the C>G introduces an ETS factor motif that enhances transcription. In contrast, in other cell types, e.g. insular cortex, where GPR65 is far less transcribed, Basenji predicts the same motifs play a role in repressing the gene.

Discussion

Transcriptional regulation is the primary driver of gene expression specificity across cell types and states. The genome research community needs more effective models of how sequence determines transcription in large mammalian genomes in order to understand how genomic alterations influence the downstream physical output of those genomes. Here, we introduced a comprehensive model to predict epigenetic and transcriptional profiles from DNA sequence. A deep convolutional neural network, trained on >4000 datasets, shares information across large distances with dilated layers to make sequential predictions along the chromosomes. The model explains considerable variance in these data, including cell type-specific activity. Predictions for sequences containing different versions of variant alleles agree with measurements made in human populations and subjected to eQTL analysis.

Although we demonstrated the present utility of this approach, there are several indications that we may be scratching the surface of what will be possible in this space. The datasets analyzed vary in quality, both by signal-to-noise ratio and technical variance from under-sampling with limited sequencing. We observed increasing predictive accuracy for experiments with greater sequencing depth and greater consistency between replicates. Thus, sequence-based modeling will benefit from improved experimental protocols, which are an area of active research (e.g., CUT&RUN in place of ChIP [33]). Furthermore, most of the samples either describe cell lines or heterogeneous tissue samples. Pending efforts to profile more pure, specific cell types and states will enhance our ability to thoroughly detect all regulatory elements and offer precise predictions of when and where regulatory activity will occur [34].

Dilated convolutions extended the reach of our model to view distal regulatory elements at distances beyond previous models to 32 kb receptive field width. This distance contains many, but certainly not all regulatory elements [35,36], and extending the model's vision and insulator-awareness will be an active area of future research, with considerable potential to improve predictive accuracy and variant interpretation.

Despite focusing only on transcription without considering post-transcriptional regulation and interaction across regulatory layers [37], we found our predictions highly informative of which genomic variants would be highlighted as eQTLs in RNA-seq population studies measuring RNA abundance levels. We foresee considerable potential in further integrating regulatory activity models trained on functional genomics profiles with population genotyping and phenotyping. These orthogonal approaches both offer views into how the noncoding genome works, and their joint consideration ought to sharpen those views. We envision Basenji as an important step forward in this direction.

Methods

Data preprocessing

Finer resolution analysis of broad regions exposes expressive machine learning models more to biases in functional genomics sequencing experiments (e.g. fragment GC%) [38,39] and repetitive DNA. Processed data available for download by the consortiums disposes multi-mapping reads and largely ignores these biases. Thus, we carried out our own processing of this data, with greater care taken to account for how these factors would influence the downstream training algorithms.

We downloaded FASTQ files for 973 CAGE experiments performed by FANTOM5 [6], 593 DNase and 1704 histone modification ChIP-seq performed by ENCODE [5], and 356 DNase and 603 histone modification ChIP-seq performed by the Epigenomics Roadmap [4]. We aligned the reads with Bowtie2, requesting the program return a maximum of 10 multi-mapping alignments [40]. We proportioned these multi-reads among those 10 positions using an EM algorithm that leverages an assumption that coverage will vary smoothly [41]. In the algorithm, we alternate between estimating expected coverage across the genome using a Gaussian filter with standard deviation 32 and re-allocating multi-read weight proportionally to those coverage estimates. We normalized for GC% bias using a procedure that incorporates several established ideas [39,42]. We assigned each position an estimated relevant GC% value using a Gaussian filter (to assign greater weight to nearby nucleotides more likely to have been part of a fragment relevant to that genomic position). Then we fit a third degree polynomial regression to the log2 coverage estimates. Finally, we reconfigured the coverage estimates to highlight the residual coverage unexplained by the GC% model. A python script implementing these procedures to transform a BAM file of alignments to a BigWig file of inferred coverage values is available in the Basenji tool suite.

Avoiding assembly gaps and unmappable regions >1 kb, we extracted (2^{17}=)131 kb non-overlapping sequences across the chromosomes. We discarded sequences with >35% unmappable sequence, leaving 14,533 sequences. We separated 5% for a validation set, 5% for a test set, and the remaining 90% for training. Within each sequence, we summed coverage estimates in 128 bins to serve as the signal for the model to predict.

Model architecture and training

We implemented a deep convolutional neural network to predict the coverage values as a function of the underlying DNA sequence. The high-level structure of the network consisted of convolution layers, followed by dilated convolution layers, and a final fully connected layer (Figure 1). All layers applied batch normalization, rectified linear units, and dropout. Standard convolution layers applied max pooling in windows of 2, 4, 4, and 4 to reach the 128 bp bin size. We compared the predicted and measured values via a Poisson regression log-likelihood function.

Dilated convolutions are convolution filters with gaps whose size increases by a factor of two in each layer, enabling the receptive field width to increase exponentially [43]. Dense connection of these layers means that each layer takes all previous layers as input, as opposed to taking only the preceding layer [44]. This architecture allows for far fewer filters per layer because the incoming representation from the standard convolutional module and the subsequent refinements of the dilated layers are all passed on; this allows each layer to focus on modeling the residual variance not yet captured [45]. We applied seven dilated convolution layers in order to reach a receptive field width of ~32 kb. This width will capture only a subset of possible distal regulatory interactions [35,36], and we intend to engineer methods to increase it. Nevertheless, it captures substantially more relevant regulatory sequence than previous models, and evidence that variant effect magnitude decreases with distance suggests there will be diminishing returns to extension [22].

We optimized the loss function via stochastic gradient descent with learning rates adapted via ADAM [46]. Our TensorFlow implementation leverages automatic differentiation and the chain rule to compute the gradient of the loss function with respect to each parameter to step towards a local optimum [47]. We used Bayesian optimization via the GPyOpt package to search for effective hyper-parameters throughout the model, including the convolution widths, convolution filter numbers, fully connected unit numbers, dropout rates, learning rate, and momentum parameters [48] [https://github.com/SheffieldML/GPyOpt].

Gene expression cluster comparison

To measure Basenji's ability to recapitulate gene expression clusters from the real data, we focused on the 2000 most variable genes and sampled sets of 1000 using a bootstrap procedure. For each sample, we performed Gaussian mixture model clustering with 10 clusters on the real and predicted gene expression matrixes across cell types. We quantified the similarity of the clusters with the adjusted Rand index statistic. The distribution of the statistic was approximately normal; thus, we estimated the mean and variance of the distribution to compute a p-value that the distribution was greater than zero.

Regulatory element saliency maps

We desired a computationally efficient measurement of the influence of distal sequence on predictions at gene TSSs. Deep learning research has suggested several effective schemes for extracting this information. Guided by the insights of prior work [49,50], we computed experiment-specific saliency maps as the dot product of the 128 bp bin representations and their gradients with respect to the model prediction for that experiment. The rectified linear unit nonlinearity guarantees that all representation values will be positive. Thus, positive gradients indicate that stronger recognition of whatever triggered the unit would increase the prediction (and weaker recognition would decrease it); negative gradients indicate the opposite. Taking the dot product with the sequence's representation amplifies the signal and sums across the vector, aggregating the effect into one signed value. Positive values identify regions where activating elements were recognized, and negative values identify repressor elements.

GTEX eQTL analysis

We downloaded the eQTL analysis in the GTEx V6p release and primarily studied the chi-squared statistics and significance calls [22]. Nearby variants in the population data can have highly correlated statistics due to linkage disequilibrium. In contrast, Basenji can isolate the contribution of individual variants. To place SED scores and eQTL statistics on a level plane, we computed their signed LD profile [55] using LD computed from 1000 Genomes Phase 3 Europeans [51]. The signed LD profile of a signed genomic annotation gives the expected marginal correlation of each SNP to a hypothetical phenotype for which the true causal effect of each SNP is the value of the annotation at that SNP.

We included several covariates that are known to influence eQTL chi-squared statistics [52]. LD score measures the amount of variation tagged by an individual variant [23]. We found that LD score correlated with the chi-squared statistics. Thus, we downloaded pre-computed scores for the European 1000 Genomes from the LDSC package [23] and included them in the analyses. We also found that distance to the nearest TSS correlated with the chi-squared statistics; variants closer to TSSs are more likely to influence gene expression. To control for this effect, we annotated SNPs with indicator variables classifying TSS distance as < 500 bp, 500-2000 bp, 2000-8000 bp, or 8000-32000 bp and computed LD scores to each annotation using LD information from 1000 Genome Phase 3 Europeans, as in [1].

Finally, we focused on chromosome 1 for computational efficiency and pruned the set of variants down to exclude variants with LD > 0.5. In a first analysis, we fit regression models individually for each tissue with LD score and |SED-LD| to the chi-squared statistics, and considered the significance of coefficients assigned to |SED-LD| across all variants. In a second analysis, we added the TSS-LD variables to the regression.

Software availability

The code to preprocess data, train models, and perform the analyses described is available from https://www.github.com/calico/basenji

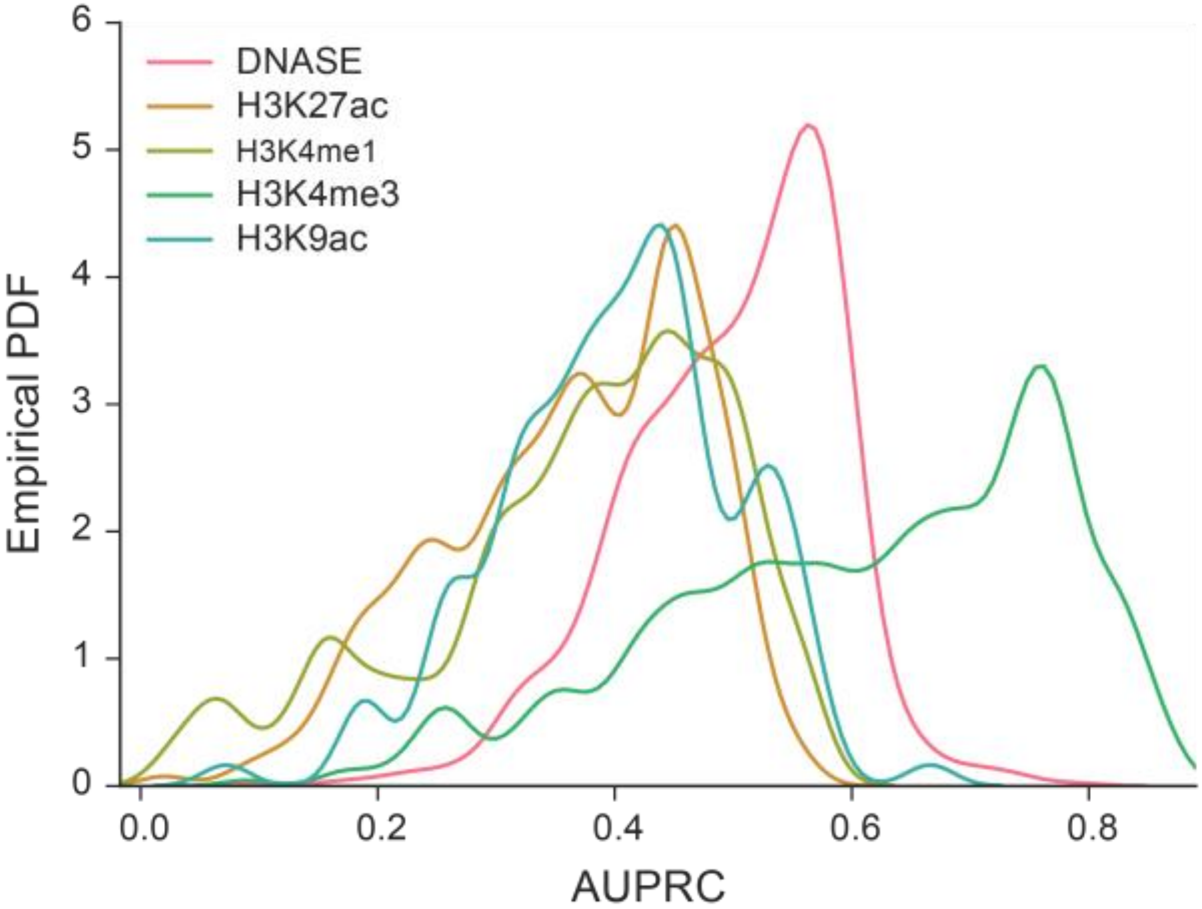
Acknowledgments

We thank David Hendrickson, Daphne Koller, Jasper Snoek, Geoffrey Fudenberg, and Marta Mele for feedback on the manuscript. We thank Ben Passarelli for invaluable computing hardware support.

Competing financial interests

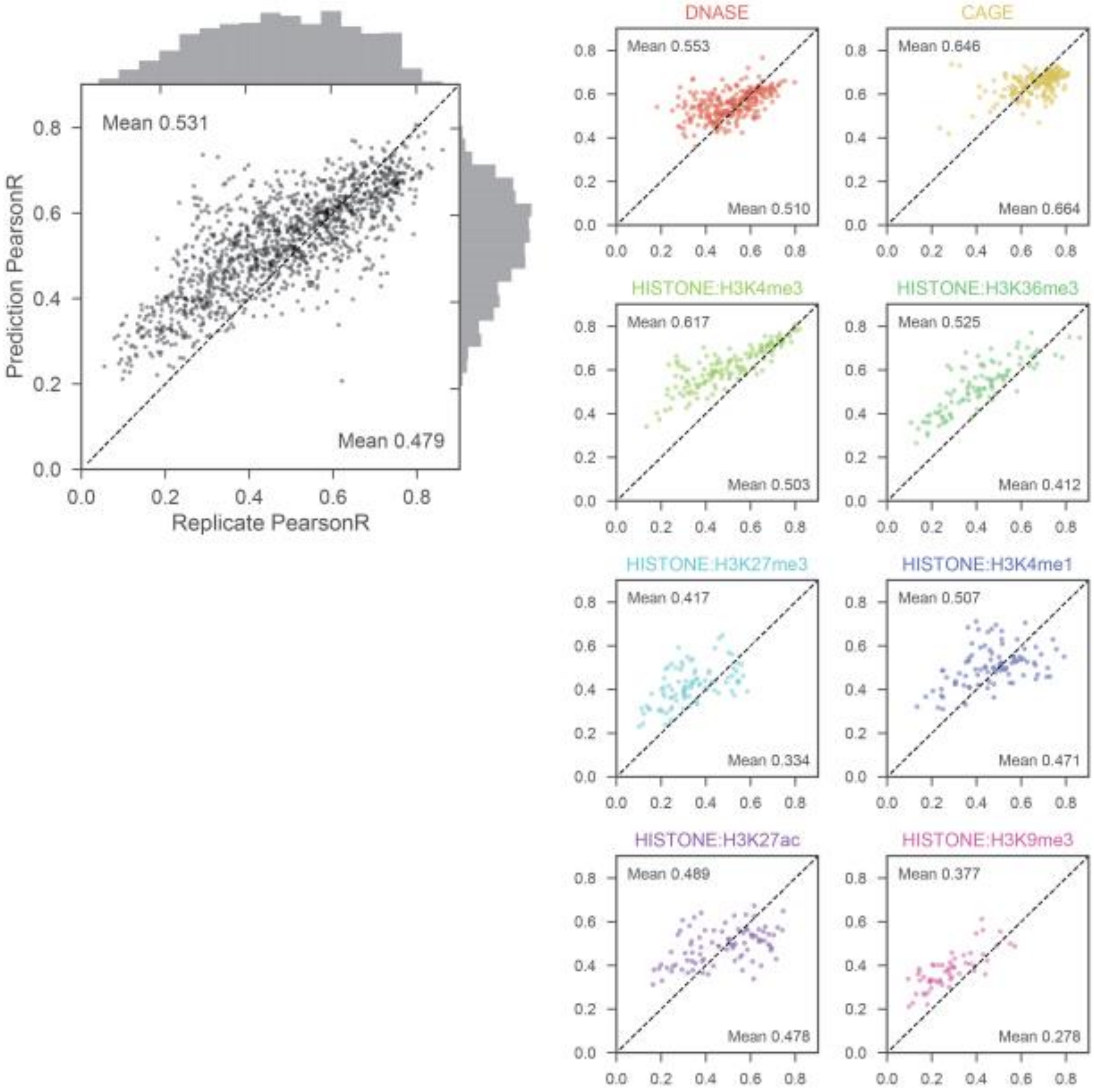
DRK is employed by Calico LLC.

Supplementary Figures

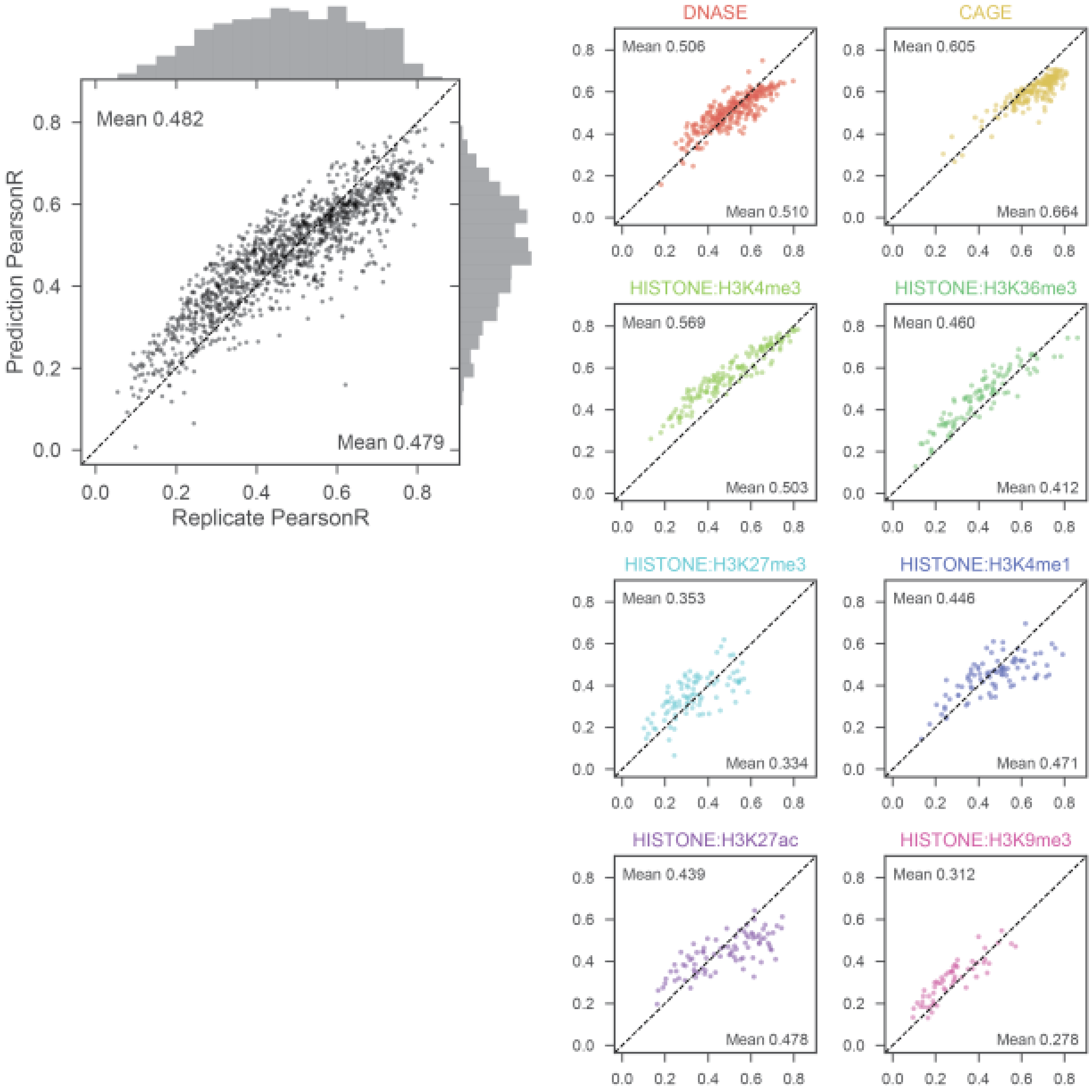


Supplementary Figure 1 - Basenji accurately predicts peaks.

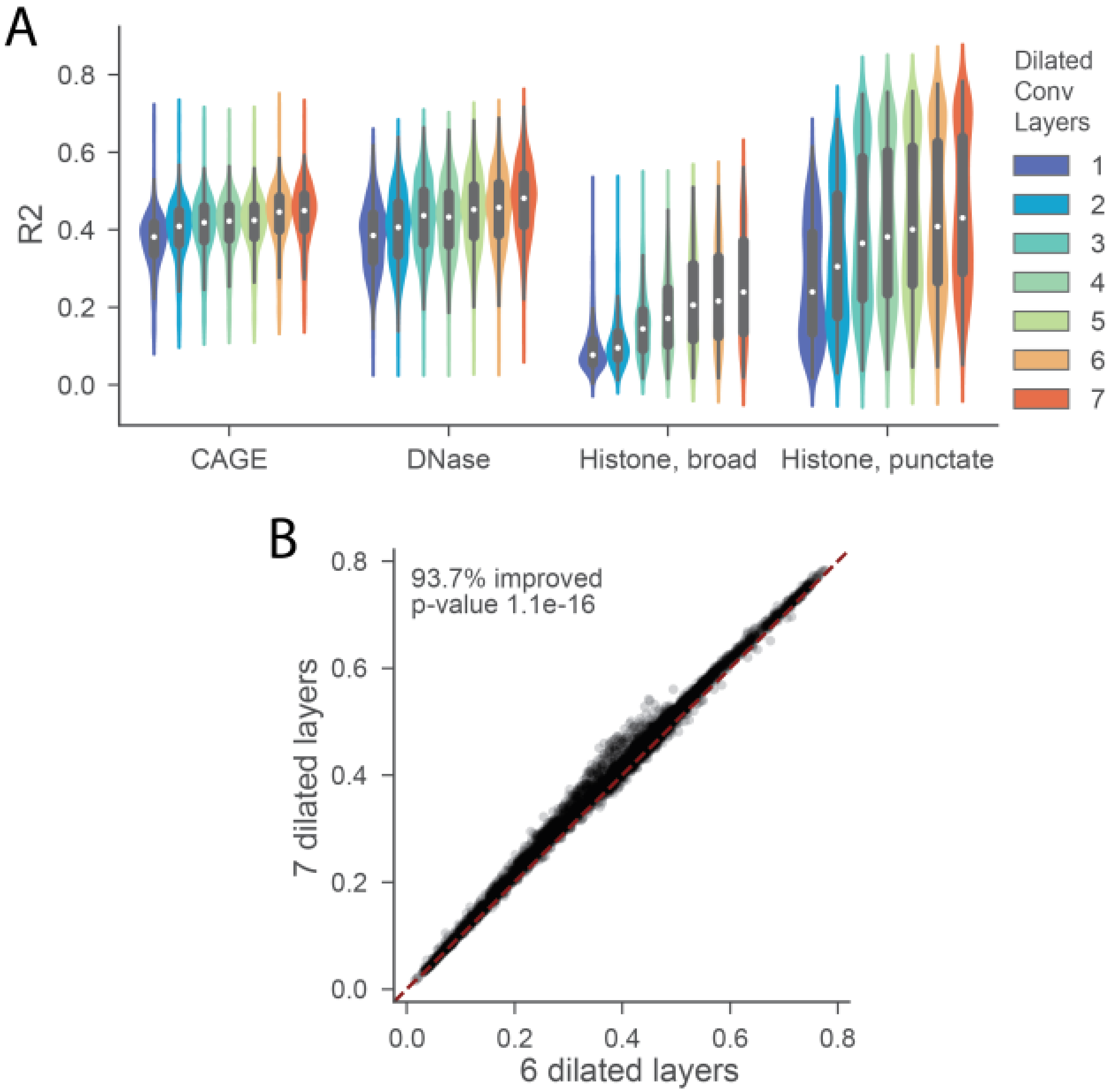
For each dataset, we called peaks on the smoothed count data within the 128 bp bins using a Poisson model similar to the MACS2 approach and applied a .01 FDR cutoff [54]. We computed the area under the precision-recall curve (AUPRC) for Basenji predictions of each experiment and plot the distributions for the various types of punctate dataset above.



Supplementary Figure 2 – Basenji predictions within-replicate match replicate concordance. For all replicated experiments, we plotted log-log Pearson correlation between the replicate experiments versus the correlation between the experiment and prediction (averaged across replicates). On the right, we make the same plots, faceted by experiment type.

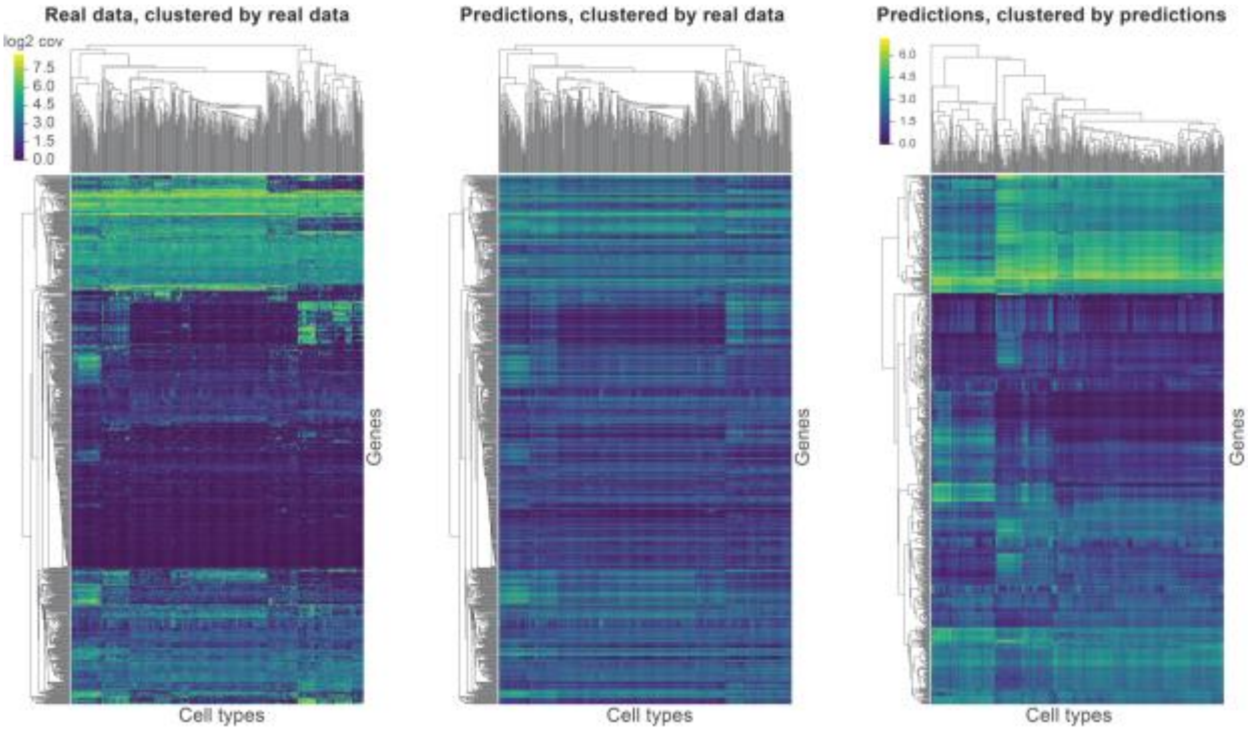


Supplementary Figure 3 – Basenji predictions cross-replicate match replicate concordance. For all replicated experiments, we plotted log-log Pearson correlation between the replicate experiments versus the correlation between the experiment and its replicate's prediction (averaged across replicates). On the right, we make the same plots, faceted by experiment type.

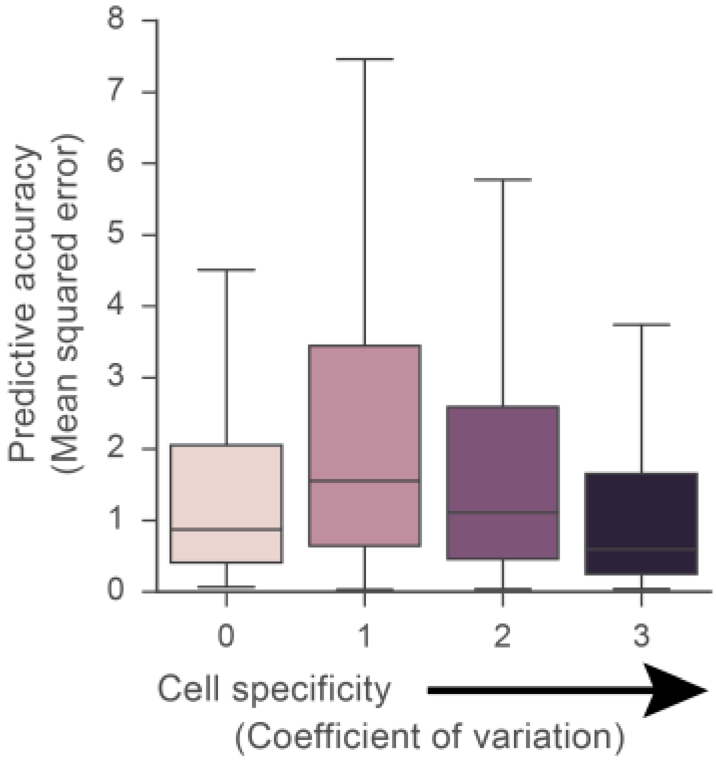


Supplementary Figure 4 - Dilated layers improve predictive accuracy.

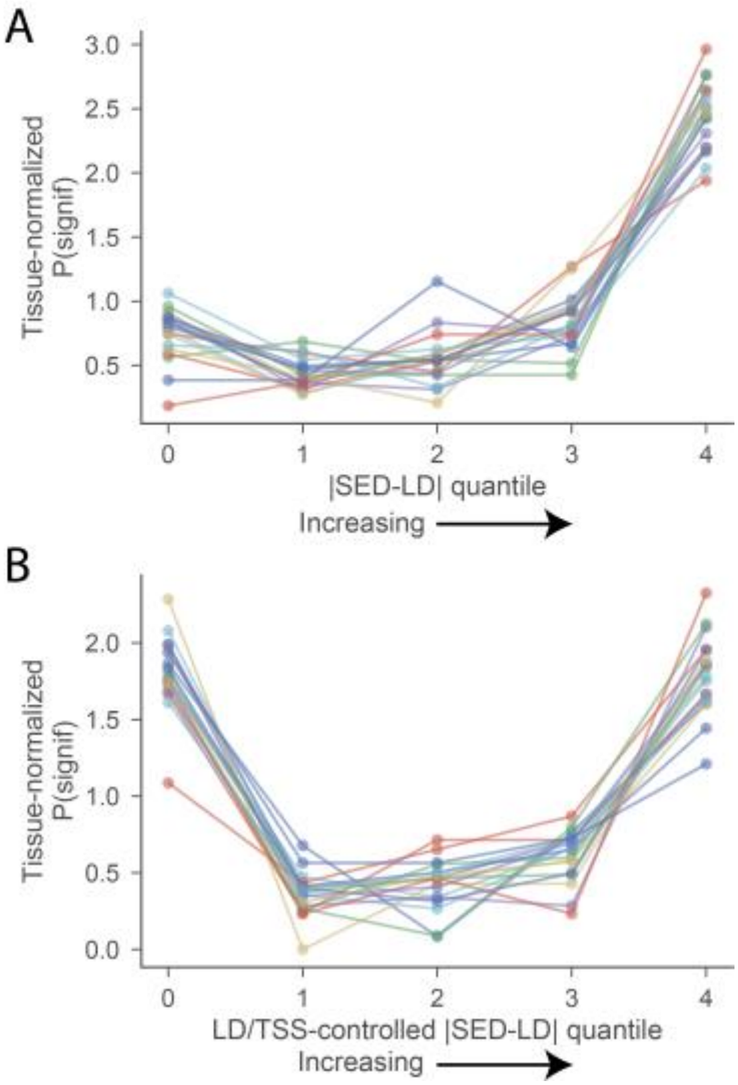
We trained models for a range of dilated convolution layer number. (A) We plotted the distribution of test R2 for each experiment, by data type. Test accuracy increases with each additional layer for all data types. (B) We plotted the test R2 of each experiment for the 6 layer versus 7 layer model. Adding the 7th layer improves test accuracy for 93.7% of the datasets.



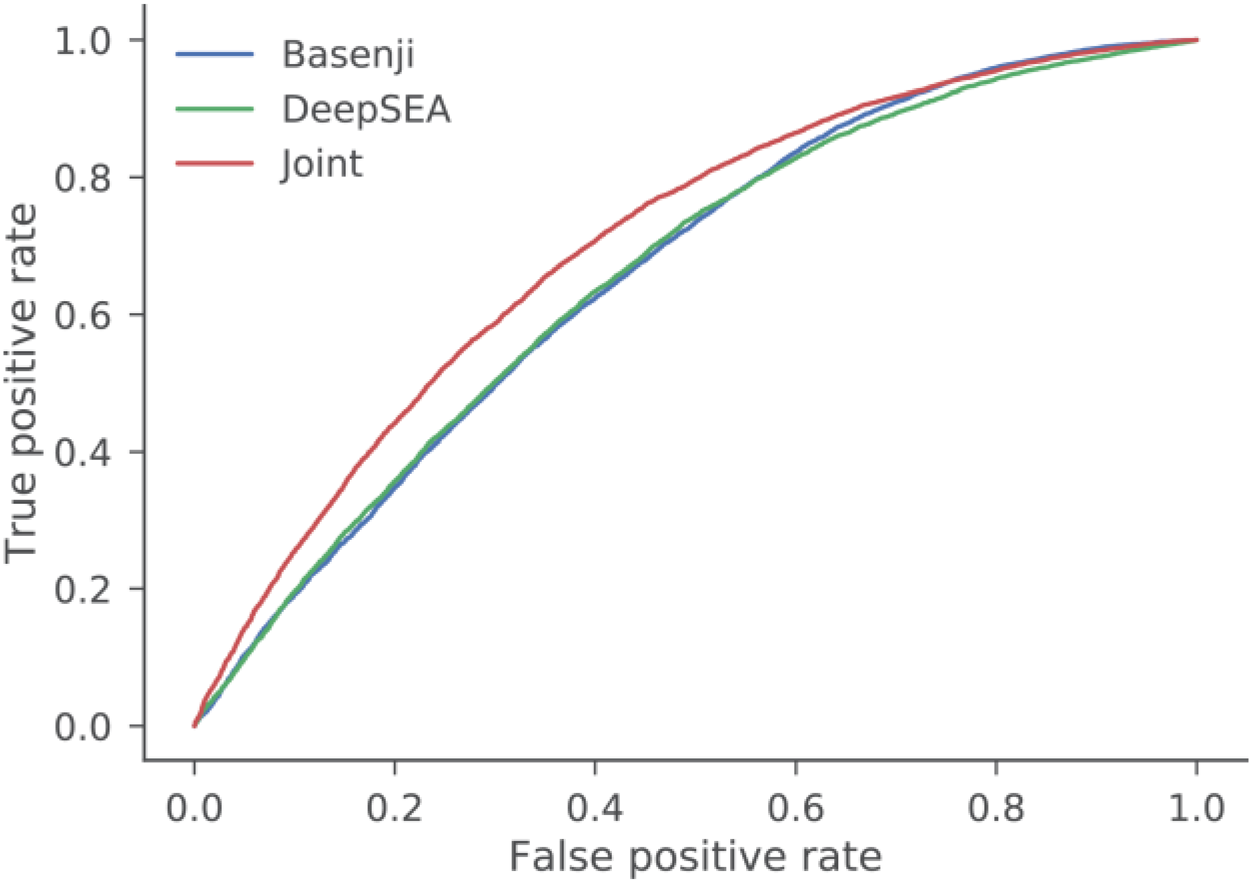
Supplementary Figure 5 - Predictions recapitulate cell type-specific expression clusters.
On the far left, we clustered and plotted as a heat map the real gene expression matrix across cell types after quantile normalization. On the far right, we similarly plotted the Basenji gene predictions matrix. In the center, we substituted Basenji predictions into the real data clustered heat map. Although the sharp definitions smear, the clusters remain visible. We used Euclidean distance and average linkage in the hierarchical clustering.



Supplementary Figure 6 – Accuracy is maintained for genes with cell type-specific expression. After removing genes expressed < 0.1 fragments per 128 bp bin, we ranked genes by their coefficient of variation across cell types to establish four quartile sets. We computed mean squared prediction error (MSE) across cell types for each gene and represent the distributions of MSE as a boxplot for each quartile set. Box represents the interquartile range (IQR), and whiskers represent 1.5*IQR. We predict the first quartile of genes with ubiquitous, steady expression the most accurately. Beyond that, MSE levels off, and accuracy is maintained with increasing cell type specificity.



Supplementary Figure 7 – SNP expression difference predictions relate to GTEx eQTL statistics. We distributed SED by the LD correlation matrix to more readily compare to eQTL measurements in human populations (Methods). |SED-LD| shows a strong relationship with eQTL statistics from GTEx. (A) For each tissue, we ranked the variants by the difference between their regression predictions including and excluding |SED-LD| and formed five quantiles. We computed the proportion of significant eQTLs in each quantile and divided by the proportion of all variants called eQTLs in that tissue to normalize the tissues to a level plane. The line plots show those normalized significance proportions in each quantile, which rise to 3-7x over the average of the bottom three quantiles in all 19 tissues. (B) We observed that TSS distance also related to variant eQTL statistics and recomputed the regression-based ranking and quantiles including TSS distance covariates (Methods). The highest SED-LD quantile remains highly enriched for eQTLs. Enrichment of the lowest quantile may be attributable to variants that influence gene expression via mechanisms beyond the transcriptional regulation that Basenji focuses on [22]. Variants that affect post-transcriptional mechanisms such as splicing would collect in the lowest quantile where the SNPs tag substantial variation near the gene, but have low |SED-LD| predictions.



Supplementary Figure 8 – Basenji predictions exceed previous methods for GWAS classification. We computed SNP expression difference scores for a dataset containing 12,296 bi-allelic SNPs taken from the NIH GWAS Catalog database [24] and a negative set with matched minor allele frequency. We computed log2 fold changes between the predictions for the two alleles at each position in the surrounding region. We let the score for each SNP be the maximum of the absolute value of that fold change across the sequence. Finally, we reduced the dimensionality of the feature set to 200 with PCA and trained a logistic regression classifier to predict presence in the GWAS catalog. The DeepSEA authors previously computed predictions for this data using their method and conservation statistics in a more sophisticated model. Basenji-based scores match DeepSEA, and a joint model using both exceeds either one.

References

1. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet (2015). doi:10.1038/ng.3404
2. O'Connor, L. J. et al. Estimating the proportion of disease heritability mediated by gene expression levels. bioRxiv 118018 (2017). doi:10.1101/118018
3. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics 16, 197–212 (2015).
4. Ernst, J. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
5. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
6. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014).
7. Whitaker, J. W., Chen, Z. & Wang, W. Predicting the human epigenome from DNA motifs. Nat Methods 12, 265–272 (2015).
8. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput Biol 10, e1003711 (2014).
9. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotech 33, 831–838 (2015).
10. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods (2015). doi:10.1038/nmeth.3547
11. 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).
12. González, A. J., Setty, M. & Leslie, C. S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. Nat Genet (2015). doi:10.1038/ng.3402
13. Cheng, C. et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 22, 1658–1667 (2012).
14. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167, 1170–1187 (2016).
15. Dekker, J. & Mirny, L. A. The 3D Genome as Moderator of Chromosomal Communication. Cell 164, 1110–1121 (2016).
16. Hashimoto, T. et al. A synergistic DNA logic predicts genome-wide chromatin accessibility. Genome Res. (2016). doi:10.1101/gr.199778.115
17. Feschotte, C. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics 9, 397–405 (2008).
18. Hathaway, N. A. et al. Dynamics and Memory of Heterochromatin in Living Cells. 149, 1447–1460 (2012).
19. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics 13, 233–245 (2012).
20. Donadelli, M., Dando, I., Fiorini, C. & Palmieri, M. UCP2, a mitochondrial protein regulated at multiple levels. Cell. Mol. Life Sci. 71, 1171–1190 (2013).
21. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The Functional Consequences of Variation in Transcription Factor Binding. PLoS Genet 10, e1004226 (2014).
22. Aguet, F. et al. Local genetic effects on gene expression across 44 human tissues. bioRxiv 074450 (2016). doi:10.1101/074450
23. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in

genome-wide association studies. Nat Genet 47, 291–295 (2015).
24. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45, D896–D901 (2017).
25. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).
26. Lambert, J.-C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452–1458 (2013).
27. Wang, J.-Q. et al. TDAG8 is a proton-sensing and psychosine-sensitive G-protein-coupled receptor. J. Biol. Chem. 279, 45626–45633 (2004).
28. Brinkmann, V. et al. Fingolimod (FTY720): discovery and development of an oral drug to treat multiple sclerosis. Nat Rev Drug Discov 9, 883–897 (2010).
29. Koscielny, G. et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 45, D985–D994 (2017).
30. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47, 955–961 (2015).
31. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. Genome Biol 8, R24 (2007).
32. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44, D110–D115 (2016).
33. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife Sciences 6, e21856 (2017).
34. Regev, A. et al. The Human Cell Atlas. bioRxiv 121202 (2017). doi:10.1101/121202
35. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 47, 598–606 (2015).
36. Javierre, B.-M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369–1384.e19 (2016).
37. Skalska, L., Beltran-Nebot, M., Ule, J. & Jenner, R. G. Regulatory feedback from nascent RNA to chromatin and transcription. Nature Reviews Molecular Cell Biology 18, 331–337 (2017).
38. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Reviews Genetics 15, 709–721 (2014).
39. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40, e72–e72 (2012).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).
41. Zhang, Q. & Keleş, S. CNV-guided multi-read allocation for ChIP-seq. Bioinformatics 30, 2860–2867 (2014).
42. Teng, M. & Irizarry, R. A. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-Seq peak callers. bioRxiv 090704 (2016). doi:10.1101/090704
43. Yu, F. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv:1511.07122 (2015).
44. Huang, G., Liu, Z., Weinberger, K. Q. & van der Maaten, L. Densely Connected Convolutional Networks. arXiv:1608.06993 (2016).
45. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. arXiv:1512.03385 (2015).
46. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014).
47. Abadi, M. et al. TensorFlow: A system for large-scale machine learning. arXiv:1605.08695

(2016).

48. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. Advances in Neural Information Processing Systems, 2951–2959 (2012).
49. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. arXiv:1704.02685 (2017).
50. Bach, S. et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10, e0130140 (2015).
51. Consortium, T. 1. G. P. A global reference for human genetic variation. Nature 526, 68–74 (2015).
52. Liu, X. et al. Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. The American Journal of Human Genetics 0, (2017).
53. IMSGC, I. M. S. G. C. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 45, 1353–1360 (2013).
54. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008).
55. Reshef, Y. et al. Quantifying directional effects of transcription factor binding on polygenic disease risk using GWAS summary statistics; (Abstract 376/W). Presented at the 66th Annual Meeting of the American Society of Human Genetics, October 20, 2016, Vancouver, Canada.