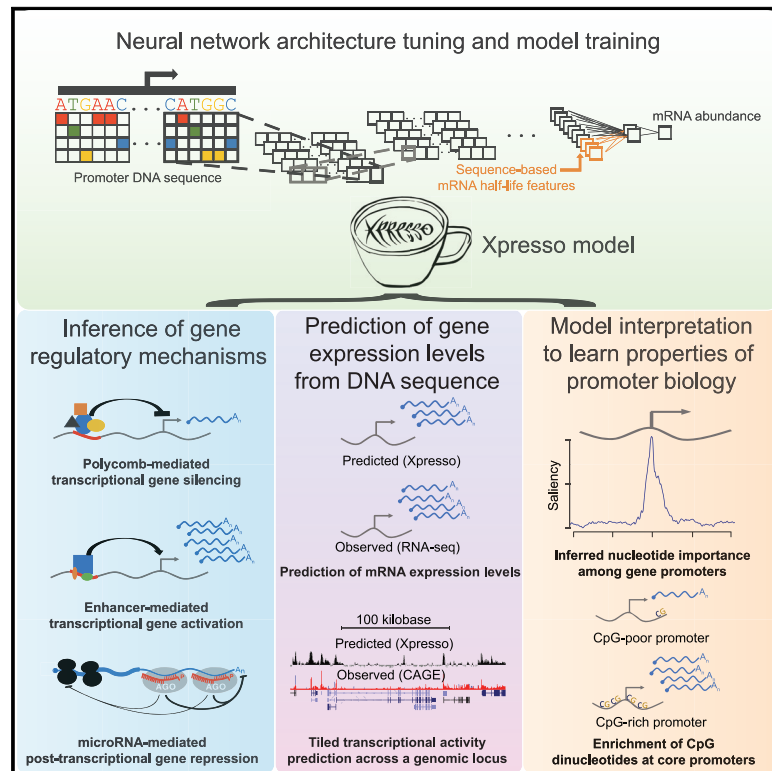


# Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks

## Graphical Abstract



## Authors

Vikram Agarwal, Jay Shendure

## Correspondence

vagar@calicolabs.com (V.A.),  
shendure@uw.edu (J.S.)

## In Brief

Agarwal and Shendure show that deep neural networks can strongly predict mRNA abundance solely from promoter sequence. The residuals of these predictions facilitate inferences about the regulatory influence of enhancers, heterochromatic domains, and microRNAs. Model interpretation reveals that CpG dinucleotide content at core promoters is associated with transcriptional activity.

## Highlights

- Deep neural networks strongly predict gene expression levels solely from DNA sequence
- Models help infer transcriptional and post-transcriptional gene regulatory mechanisms
- Predictive power is competitive with models using thousands of biochemical datasets
- CpG dinucleotide content at core promoters is strongly predictive of mRNA abundance



## Resource

# Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks

Vikram Agarwal<sup>1,2,\*</sup> and Jay Shendure<sup>1,3,4,5,\*</sup><sup>1</sup>Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA<sup>2</sup>Calico Life Sciences LLC, South San Francisco, CA 94080, USA<sup>3</sup>Howard Hughes Medical Institute, Seattle, WA 98195, USA<sup>4</sup>Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA, USA<sup>5</sup>Lead Contact\*Correspondence: [vagar@calicolabs.com](mailto:vagar@calicolabs.com) (V.A.), [shendure@uw.edu](mailto:shendure@uw.edu) (J.S.)<https://doi.org/10.1016/j.celrep.2020.107663>

## SUMMARY

Algorithms that accurately predict gene structure from primary sequence alone were transformative for annotating the human genome. Can we also predict the expression levels of genes based solely on genome sequence? Here, we sought to apply deep convolutional neural networks toward that goal. Surprisingly, a model that includes only promoter sequences and features associated with mRNA stability explains 59% and 71% of variation in steady-state mRNA levels in human and mouse, respectively. This model, termed Xpresso, more than doubles the accuracy of alternative sequence-based models and isolates rules as predictive as models relying on chromatic immunoprecipitation sequencing (ChIP-seq) data. Xpresso recapitulates genome-wide patterns of transcriptional activity, and its residuals can be used to quantify the influence of enhancers, heterochromatic domains, and microRNAs. Model interpretation reveals that promoter-proximal CpG dinucleotides strongly predict transcriptional activity. Looking forward, we propose cell-type-specific gene-expression predictions based solely on primary sequences as a grand challenge for the field.

## INTRODUCTION

Cellular function is governed in large part by the repertoire of proteins present and their relative abundances. Initial attempts to model the gene regulatory forces specifying the mammalian proteome posited a major role for translational regulation, implying that mRNA levels might be more poorly predictive of protein abundance than is often assumed (Schwanhäusser et al., 2011). However, subsequent reanalyses of those data have shown that as much as 84% of variation in protein levels can be explained by mRNA levels, with transcription rates contributing 73%, and mRNA degradation rates contributing 11% (Li et al., 2014). This work reinforces the view that steady-state protein abundances are highly predictable as a function of mRNA levels (Vogel et al., 2010).

Although quantitative models that predict protein levels from mRNA levels are available (Edfors et al., 2016), we lack models that can accurately predict mRNA levels. Steady-state mRNA abundance is governed by the rates of transcription and mRNA decay. For each gene, a multitude of regulatory mechanisms are carefully integrated to tune these rates and thus specify the concentrations of the corresponding mRNAs that cells of each type will produce. Key mechanisms include (1) the recruitment of an assortment of transcription factors (TFs) to a gene's promoter region; (2) epigenetic silencing, as frequently demarcated

by Polycomb-repressed domains associated with H3K27me3 histone marks (Cao et al., 2002); (3) the activation of genes by enhancers, stretch enhancers (Parker et al., 2013), and super-enhancers (Whyte et al., 2013); and (4) the degradation of mRNA through microRNA-mediated targeting (Agarwal et al., 2015). Jointly modeling these diverse aspects of gene regulation within a quantitative framework has the potential to shed light on their relative importance, to elucidate their mechanistic underpinnings, and to uncover new modes of gene regulation.

Previous attempts to model transcription and/or mRNA decay can be broadly split into those based on correlative biochemical measurements and those based on primary sequence. In the former category, there have been several attempts to model the relationship between TF binding, histone marks, and/or chromatin accessibility and gene expression (e.g., predicting the expression levels of genes based on chromatin immunoprecipitation sequencing [ChIP-seq] and/or DNase I hypersensitivity data) (Cheng et al., 2011, 2012; Dong et al., 2012; Karlič et al., 2010; McLeay et al., 2012; Ouyang et al., 2009; Schmidt et al., 2017). Although such models can clarify the relationships between heterogeneous, experimentally derived biochemical marks and transcription rates, their ability to deliver mechanistic insights is limited. For example, for models relying on histone marks, the temporal deposition of such marks might follow, rather than precede, the events initiating transcription, in which



case, the histone marks reinforce or maintain, rather than specify, a transcriptional program. For models relying on measurements of TF binding, a substantial fraction of ChIP-seq peaks lack the expected DNA binding motif and potentially reflect artifactual binding signals originating from the predisposition of ChIP to pull down highly transcribed regions or regions of open chromatin (Jain et al., 2015; Krebs et al., 2014; Teytelman et al., 2013).

In the latter category, there have also been a few attempts to model transcript levels or mRNA decay rates based solely on primary sequence. For example, a model of the spatial positioning of *in silico* predicted TF binding sites relative to transcriptional start sites (TSSs) was able to explain 8%–28% of variability in gene expression (McLeay et al., 2012). Models based on simple features, including the guanine-cytosine (GC) content and lengths of different functional regions (e.g., the 5' UTR, open reading frame [ORF], introns, and 3' UTR) and ORF exon junction density (Sharova et al., 2009; Spies et al., 2013) explain as much as 40% of the variability in mRNA half-lives, which are, in turn, estimated to explain 6%–15% of the variability in steady-state mRNA levels in mammalian cells (Li et al., 2014; Schwanhäusser et al., 2011; Spies et al., 2013). However, most variation in steady-state mRNA levels has yet to be explained by sequence-based models.

To what extent is gene expression predictable directly from the genome sequence? Relevant to this question, a study relying on a massively parallel reporter assay (MPRA) demonstrated that the transcriptional activities associated with isolated promoters can explain a majority (~54%) of endogenous promoter activity (van Arensbergen et al., 2017). This result establishes a clear mechanistic link between the primary sequence of promoters and variability in gene expression levels. This, in turn, implies that there may exist a mathematical function which, if properly parameterized, could accurately predict mRNA expression levels based upon nothing more than genomic sequence. However, it remains unknown whether such a function is “learnable” given limited training data and highly incomplete domain-specific knowledge of the parameters governing gene regulation (e.g., biochemical parameters describing the affinity of TFs to their cognate motifs [ $K_d$ ], constants describing the rates of TF binding and unbinding [ $K_{on}$  and  $K_{off}$ , respectively], potential cooperative effects from the combinatorial binding of TFs (as measured by Hill coefficients), the distance dependencies between TF binding relative to the TSS and RNA polymerase II recruitment, and competition for binding between TFs and histones [Segal and Widom, 2009], etc.).

Methods based upon deep learning are providing unprecedented opportunities to automatically learn relationships among heterogeneous data types in the context of incomplete biological knowledge (Angermueller et al., 2016). Such methods often employ deep neural networks; in which, multiple layers are employed hierarchically to parameterize a model, which transforms a given input into a specified output. For example, deep convolutional neural networks have been used to predict the binding preferences of RNA and DNA binding proteins (Alipanahi et al., 2015), the impact of noncoding variants on the chromatin landscape and mRNA levels (Zhou and Troyanskaya, 2015; Zhou et al., 2018), the chromatin accessibility of a cell type from a

DNA sequence (Kelley et al., 2016), and genome-wide epigenetic measurements of a cell type from a DNA sequence (Kelley, 2019; Kelley et al., 2018).

The application of deep learning to model the various regulatory processes governing gene expression in a unified framework has great potential and could enable the discovery of fundamental relationships between primary DNA sequence and steady-state mRNA levels that have, heretofore, remained elusive. To that end, we introduce Xpresso, a deep convolutional neural network that jointly models promoter sequences and features associated with mRNA stability to predict steady-state mRNA levels.

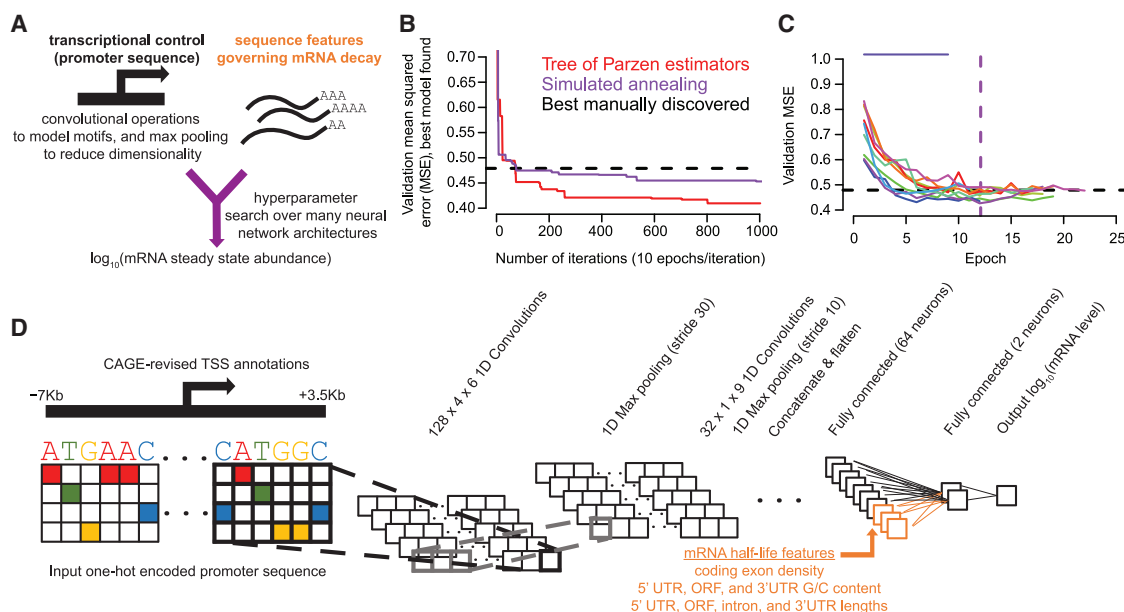
## RESULTS

### An Optimized Deep-Learning Model to Predict mRNA Expression Levels

We aspired to train a quantitative model using nothing more than a genomic sequence to predict mRNA expression levels. To simplify the prediction problem, we first evaluated the correlation structure of 56 human cell types in which mRNA expression levels had been collected and normalized by the Roadmap Epigenomics Consortium (Roadmap Epigenomics Consortium et al., 2015). An evaluation of the pairwise Spearman correlations of mRNA expression levels among cell types revealed that most cell types were highly correlated, exhibiting an average correlation of ~0.78 between any pair of cell types (Figure S1). This justified the initial development of a cell-type-agnostic model for predicting median mRNA expression levels. We observed that median mRNA levels for chrY genes were highly variable because of the sex chromosome differences among cell types. Histone mRNAs were also undersampled and measured inaccurately because of the dependency of the underlying RNA sequencing (RNA-seq) protocols on poly(A)-tails, which histones lack. We, therefore, excluded chrY and histone genes from our analyses.

We initialized a search of hyperparameters defining the architecture of a neural network that could more optimally predict gene expression levels while jointly modeling both promoter sequences and sequence-based features correlated with mRNA decay (Figure 1A). During this search, we varied several key hyperparameters defining the deep neural network (Table 1). The mRNA decay features, which included the GC content and lengths of different functional regions (e.g., the 5' UTR, ORF, introns, and 3' UTR) and ORF exon junction density (Sharova et al., 2009; Spies et al., 2013) were not varied.

We applied two optimization strategies that have shown promise in the context of hyperparameter searches: the simulated annealing (SA) and the Tree of Parzen estimators (TPE) (Bergstra et al., 2011). We compared their performance to the best manually defined deep-learning architecture, which was guided by prior knowledge that information governing transcription rate is most likely localized to sequence elements within  $\pm 1,500$  bp promoter around a TSS and further inspired by an existing deep-learning architecture previously used to predict regions of chromatin accessibility from DNA sequence (Table 1) (Kelley et al., 2016). We observed that each optimization strategy progressively discovered better sets of hyperparameters, with the



**Figure 1. Overview of Optimization Strategy for Deep Learning-Based Model Training Scheme**

(A) Overview of a predictive model of mRNA steady-state abundance that integrates information from sequences at the promoter region of a gene and annotation-based sequence features associated with mRNA decay.

(B) Validation error associated with the best model found at each iteration during the search for an optimal set of hyperparameters to predict median mRNA abundance across tissues.

(C) Performance of 10 independent trials given the optimal architecture discovered in (B). Nine of 10 trials achieve convergence, with the vertical purple dashed line indicating the best model achieved was at the twelfth epoch, and the horizontal dashed line indicating the best manually discovered model as shown in (B).

(D) Best deep-learning architecture discovered during the hyperparameter search in (B) corresponds to an architecture with two sequential convolutional and max pooling layers followed by two sequential fully connected layers.

TPE method achieving the best validation mean squared error (MSE) of 0.401 (Figure 1B).

Given the stochastic nature of training deep-learning models, we devised a strategy to train 10 independent trials using the best deep-learning architecture specified by the hyperparameters discovered using the TPE method. We observed that nine of 10 trials converged to similar MSE values (Figure 1C). For our final model, we selected the parameters derived from the specific trial and epoch that minimized the validation MSE. All the following results of this study report the performance derived from the best of 10 trained models.

Our final model, which considered the region 7 kb upstream to 3.5 kb downstream of the TSS, was comprised of two sequential convolutional and max-pooling layers, followed by two fully connected layers preceding the output neuron (Table 1; Figure 1D), and consisted of 112,485 parameters in total (Figure S2A). An evaluation of suboptimal hyperparameters suggested this 10.5-kb sequence-window region was not critical for good performance because an alternative model spanning the region 1.5 kb upstream to 7.5 kb downstream of the TSS obtained a similar validation MSE (Figure S2B). Furthermore, our manually defined architecture, spanning the region  $\pm 1.5$  kb of the TSS achieved an  $r^2$  of 0.53, only 6% worse than the model discovered by TPE, indicating that a localized region around the core promoter region captured most learnable information, with only a modest additional contribution gained from the consideration of surrounding regions.

To evaluate the relationship between the number of genes in the training set and the performance of the model, we subsampled the training set and evaluated the MSE and  $r^2$  on the validation and test set, respectively. We found that the greatest gain in performance occurred between 4,000 and 6,000 training examples (Figure 2A).

### Performance of Predictive Models in Human and Mouse

We next sought to compare the generality and performance of our method across mammalian species. We focused on 18,377 and 21,856 genes in human and mouse, respectively, for which we could match promoter sequences and gene expression levels, and held out 1,000 genes in each species as a test set. The best human model achieved an  $r^2$  of 0.59 (Figure 2B). Although we achieved this result using cap analysis gene expression (CAGE) data to refine the TSS annotations (Figure 1C), the use of CAGE was optional because the performance of the model using only Ensembl TSS annotations was only modestly worse ( $r^2$  of 0.54). Using an identical model architecture and training scheme (Figure 1C), the best mouse model in the mouse achieved a substantially higher  $r^2$  of 0.71 (Figure 2C).

To test whether the regulatory rules learned by each model could generalize across species, we re-trained human- and mouse-specific models using training and validation sets that were matched to have the same group of one-to-one orthologs. We then tested the performance of these models on a held-out group of one-to-one orthologs in either the same or opposite

**Table 1. Search Space and Hyperparameters Discovered**

Hyperparameter	Range	Step Size (if Discrete)	Best Identified Manually	Best Identified from Search
Batch Size	2 <sup>[5,7]</sup>	1	64	128
Upstream Distance from TSS	[−10,000, 0]	500	−1,500	−7,000
Downstream Distance from TSS	[0, 10000]	500	1,500	3,500
[Convolutional → Max Pooling] Layer(s)	[One, Two, Three, Four]		Two	Two
***Number of convolutional filters	2 <sup>[4, 7]</sup>	1	64/64	128/32
***Convolution filter length	[1, 10]	1	5/5	6/9
***Convolution dilation rate	[1, 4]	1	1/1	1/1
***Max pooling pool size/stride	[5, 100]	5	10/20	30/10
Densely Connected Layer(s)	[One, two]		One	Two
***Number of Neurons in Layer	2 <sup>[1, 8]</sup>	1	100	64/2
***Dropout Probability	[0, 1]		0.5	0.00099/ 0.01546

Within brackets, the upper and lower ranges are listed if the variable is either discrete or continuous, and all possible values are listed if the variable is categorical. If the variable is discrete, the step size is provided. Beneath each hyperparameter are the associated nested hyperparameters (indicated by \*\*\*) that are searched. For example, if three [Convolutional → Max Pooling] layers are selected, each of these three layers possesses four additional types of hyperparameters to search among. The values for consecutive layers that were ultimately selected in the final models are separated by a slash (/).

species. The models trained on each species achieved similar performances on the test set of the opposite species as that of the same species (Figure 2D), demonstrating that the learned regulatory principles generalize across the mammalian phylogeny. The similar  $r^2$  values between human and mouse obtained when restricting the analyses to one-to-one orthologs suggested that the 12% greater  $r^2$  previously observed in the mouse relative to human (Figures 2B and 2C) was due to differences in the distributions of gene expression levels between the mouse and human. Indeed, although more than 20% of all mouse genes were non-expressed (Figure 2E, displayed on the x-axis as −1 because of the addition of a pseudocount of 0.1 before log-transforming the reads per kilobase of transcript per million mapped reads [RPKM] values), fewer than 10% of all human genes were non-expressed. In contrast, the subset of 15,348 one-to-one orthologs in human and mouse revealed a similar proportion of non-expressed genes (Figure 2E). To directly test the effect of these differences, we downsampled the proportion of non-ex-

pressed mouse genes to match that of the human and re-trained a mouse-specific model. The  $r^2$  decreased only modestly to 0.65 on a held-out test set, suggesting that the differential proportion of non-expressed genes only partially explains the improved performance in the mouse.

Finally, we were interested in assessing to what extent species-specific gene expression differences can be explained by differences in promoter sequence. Toward that goal, we evaluated the relationship between expression levels and one-to-one orthologs. We found that the two species showed strikingly concordant median expression levels, with only a subset of 584 genes enriched by at least 10-fold in one species (Figure 2F). A binary classifier based upon the difference in predictions from models trained in each species could discriminate these species-specific mRNAs (area under the curve [AUC] = 0.78; Figure 2G).

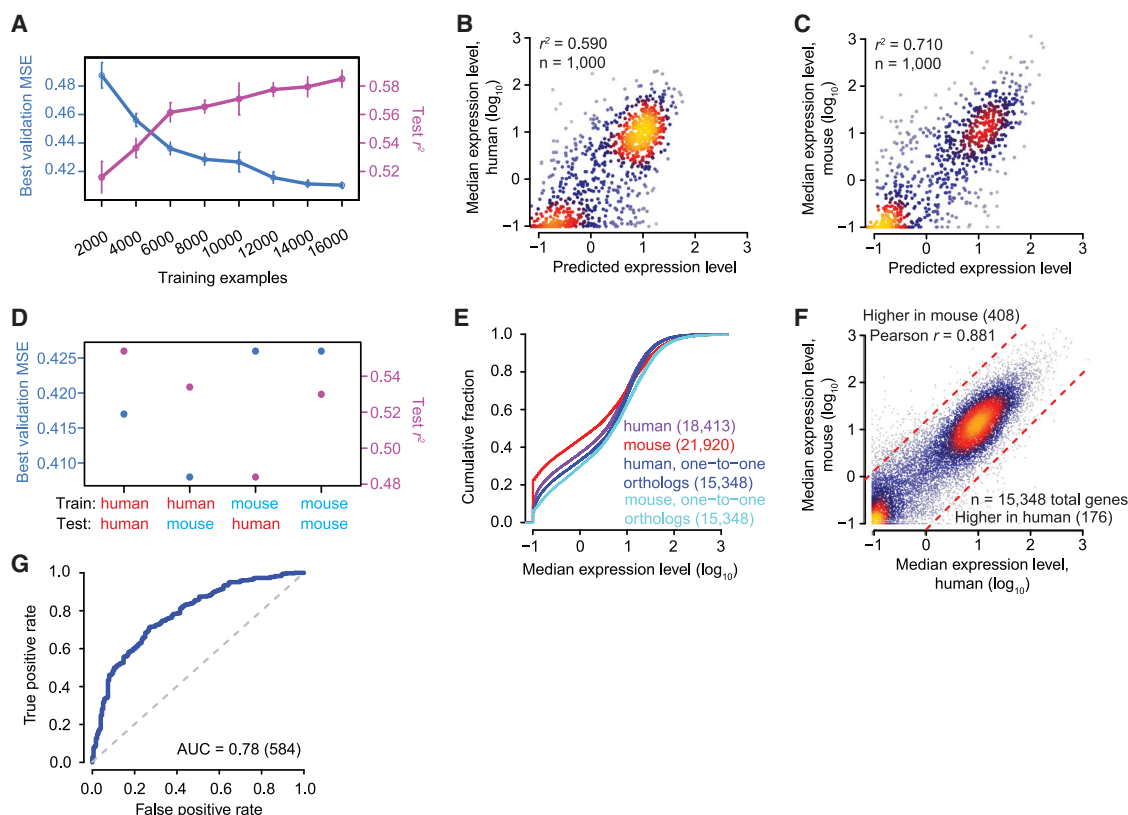
### Cell-Type-Specific Models Implicate a Diversity of Gene Regulatory Mechanisms

Given the generality of our deep-learning framework, we next sought to build cell-type-specific models. With the same hyperparameters, we trained new models to predict the expression levels of all protein-coding genes for human myelogenous leukemia cells (K562), human lymphoblastoid cells (GM12878), and mouse embryonic stem cells (mESCs) (Table S1). To avoid overfitting, we developed a nested 10-fold cross-validation-based procedure to ensure that the prediction for any given gene resulted from its being part of a held-out sample. From the residuals, we investigated whether we could observe the influence of additional gene regulatory mechanisms that were not initially considered, or incompletely accounted for, in the Xpresso model.

We first evaluated K562 cells, finding that our cell-type-specific predictions correlated with observed K562 expression levels with an  $r^2$  of 0.51 (Figure 3A). We hypothesized that genes predicted inaccurately might be under the control of gene regulatory mechanisms not considered by our model. One likely mechanism involves enhancers, *cis*-acting regulatory elements that may be located hundreds of kilobases away from a TSS. For example, in K562 cells, distal enhancers have been implicated as modulating the expression of most genes of the  $\alpha$ -globin and  $\beta$ -globin loci, GATA1, MYC, and others (Fulco et al., 2016; Klann et al., 2017; Xie et al., 2017). Reasoning that such genes should be consistently underestimated by our predictions, we plotted the distribution of their residuals (Figure S3A). Indeed, all of these genes were expressed much more highly than our predictions in K562 cells, reinforcing the notion that such genes are activated by regulatory mechanisms beyond promoters. Among the biggest outliers were the  $\beta$ -globin genes, which, in some cases, were expressed more than four orders of magnitude more highly than predicted (Figure 3A).

To more systematically identify genes activated or silenced by non-promoter mechanisms, we developed a method to predict them genome-wide. Given that enhancers are frequently associated with large domains of H3K27Ac activity and that silenced genes are frequently associated with heterochromatic domains marked by H3K27me3, we examined genome-wide chromatin-state annotations based upon the hierarchical hidden Markov





**Figure 2. Predictive Models Explain Variability in Gene Expression in the Human and Mouse**

(A) Effect of downsampling the training set size on the validation error and performance of the test set. To generate the mean and 95% confidence intervals displayed, the training set was downsampled (without replacement) into 10 batches. For each batch, the performance for the best of 10 models, defined as the model with the minimal validation error, was computed.

(B and C) Performance of human (B) and mouse (C) models on held-out test sets. Regions are colored according to the density of data from light blue (low density) to yellow (high density).

(D) Effect of training and testing model performance within and across mammalian species using a test set matched for the same set of one-to-one orthologs. Shown is the performance for the best of 10 models acquired for the full training set derived from each species.

(E) Cumulative distributions of median mRNA expression levels among tissues for all annotated human and mouse genes as well as those corresponding to the subset of one-to-one human-to-mouse orthologous genes.

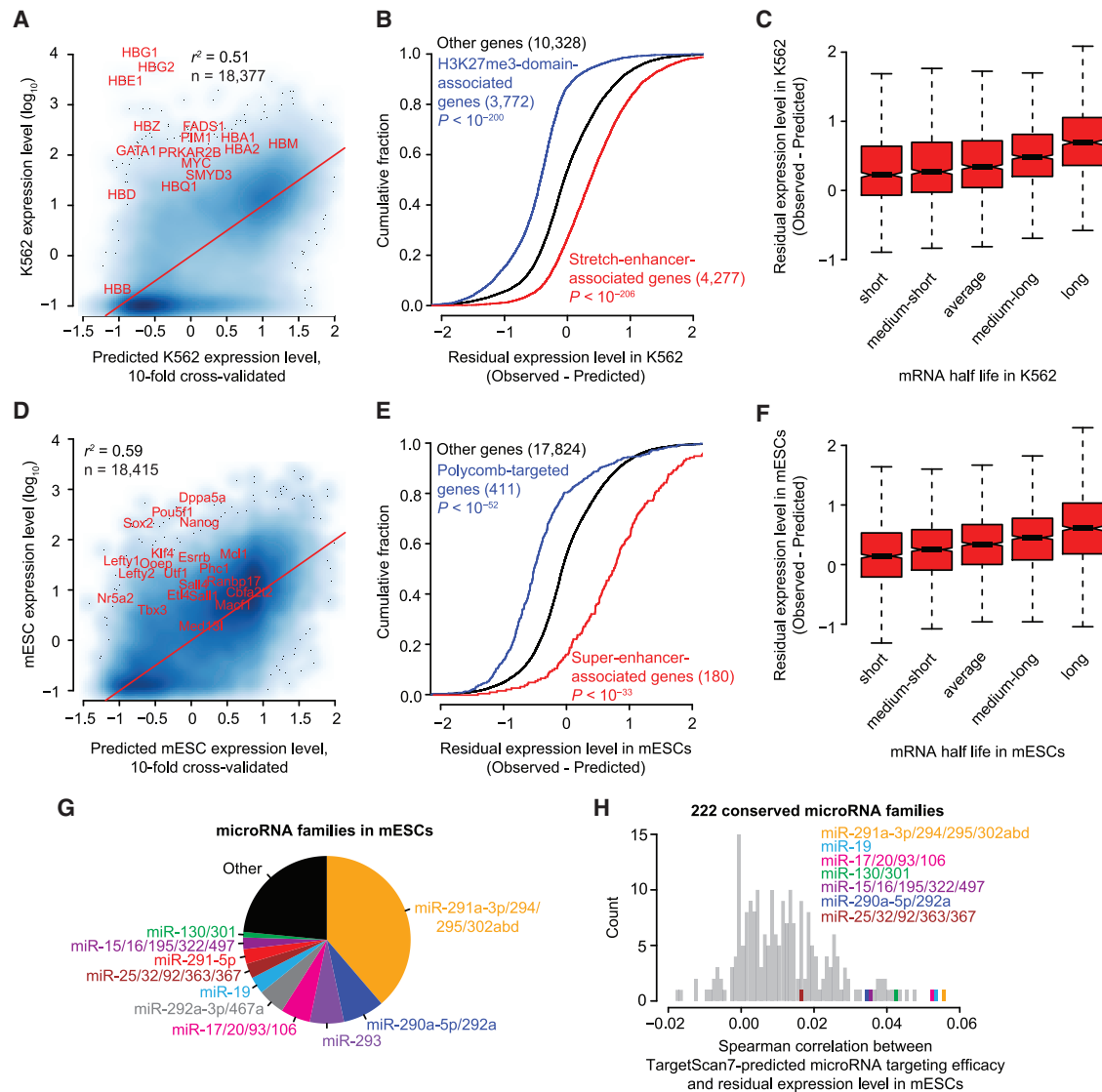
(F) Relationship between median mRNA expression levels in human and mouse for one-to-one orthologs. The dotted red lines correspond to the threshold used to call species-specific genes, corresponding to a 10-fold change in expression in one species relative to the other. The number of species-specific genes surpassing that threshold is indicated in parentheses.

(G) Performance of a classifier that uses the difference between nested, cross-validated, species-specific predictions to distinguish mRNAs whose expression is strongly enriched in the mouse or human.

model (diHMM), a method that annotates such domains genome-wide in K562 and GM12878 cells using histone marks (Marco et al., 2017). Although a subset of H3K27Ac-associated domains were originally called “super-enhancers” (Marco et al., 2017), we find it more appropriate to refer to them as “stretch enhancers,” which are more loosely defined as clusters of enhancers spanning  $\geq 3$  kb (Parker et al., 2013). In total, we identified 4,277 genes that overlap stretch enhancers, and 3,772 genes that overlap with H3K27me3 domains, ignoring those that happen to overlap with both. Consistent with our expectation, these collections of genes were significantly associated with predominantly positive and negative residuals, respectively, relative to the background distribution of other genes (Figure 3B). The same was true for GM12878 cells (Fig-

ure S3B), reinforcing the generality of this finding across cell types.

We next examined whether our residuals could be further explained by post-transcriptional gene regulatory mechanisms. Highly reproducible mRNA half-life estimates for 5,007 genes in K562 cells were previously measured using TimeLapse-seq (Schofield et al., 2018). We observed a positive correlation between mRNA half-lives (which were log transformed) and our residuals (Pearson correlation = 0.28;  $p < 10^{-15}$ ). We visualized this trend by splitting the half-lives into five equally sized bins (Figure 3C). These results show that, although we included sequence-based features associated with mRNA decay rates in the model, these were insufficient to capture the full contribution of mRNA decay rates to steady-state mRNA levels.



**Figure 3. A Diversity of Gene Regulatory Mechanisms Are Associated with the Residuals of Cell-Type-Specific Models**

(A) Relationship between nested 10-fold cross-validated predictions and actual mRNA expression levels in K562 cells. Regions are colored according to the density of data from light blue (low density) to dark blue (high density). Labeled in red are the globin genes alongside others implicated as genes activated by strong enhancers in K562 cells. Gene names were moved slightly to enhance readability.

(B) Cumulative distributions of residuals corresponding to all stretch-enhancer-associated genes, H3K27me3-domain-associated genes, and control genes not associated with either. Similarity of the distributions to that of the set of controls was tested (one-sided Kolmogorov-Smirnov [K-S] test,  $p$ -value); the number of mRNAs analyzed in each category is listed in parentheses.

(C) Relationship between mRNA half-life and residuals in K562 cells; indicated is the median residual value (bar), 25th and 75th percentiles (box), and the minimum of either 1.5 times the interquartile range or the most extreme data point (whiskers). Half-life was measured using TimeLapse-seq (Schofield et al., 2018) in K562 cells ( $n = 5,007$  genes), and partitioned into five equally sized bins spanning the range of half-life values.

(D) This panel mirrors that shown in (A), except that it highlights genes associated with known enhancers and super-enhancers in mouse embryonic stem cells (mESCs).

(E) This panel mirrors that shown in (B), except that it compares genes associated with super-enhancers and Polycomb-repressed domains in mESCs.

(F) This panel mirrors that shown in (C), except that half-life measurements were measured using SLAM-seq (Herzog et al., 2017) in mESCs ( $n = 6,266$  genes).

(G) Pie chart indicating the relative proportions of microRNA families expressed in mESCs. Colored are the top 10 most abundant miRNA families.

(H) Histogram of the Spearman correlation values between TargetScan7-predicted microRNA targeting efficacy (Agarwal et al., 2015) and residual expression level in mESCs, for 222 miRNA families conserved among mammals. Highlighted are the subset of seven highly abundant miRNA families colored in panel (G) that are also conserved among mammals.

We next turned our attention to mESCs. Key stem cell identity genes include *Pou5f1* (also known as *Oct4*), *Sox2*, *Nanog*, and a host of others whose dependence upon enhancers and super-enhancers has been experimentally validated with CRISPR-deletion experiments (Moorthy et al., 2017; Whyte et al., 2013). Similar to the other cell types, although a mESC-specific model could strongly predict mRNA expression levels in mESCs ( $r^2 = 0.59$ ), key stem cell identity genes harbored residuals that were strongly biased toward positive values (Figures 3D and S3C), confirming that their promoter sequences and mRNA sequence features could not adequately explain their high abundance. Extending this question more systematically to 180 protein-coding genes thought to be governed by super-enhancers in mESCs (Whyte et al., 2013), we observed a strong enrichment for highly positive residuals in these genes relative to all other genes (Figure 3E).

In mESCs, genes associated with the Polycomb repressive complex (PRC), as delineated by binding to both PRC1 and PRC2 and frequently marked with H3K27me3, are thought to be associated with key developmental regulators, many of which are silenced but poised to be activated upon differentiation (Boyer et al., 2006). We observed that this group of genes, in contrast to the super-enhancer-associated set, exhibited a strong enrichment in highly negative residuals relative to all other genes (Figure 3E), consistent with a model in which PRC-targeted genes are actively silenced.

Mirroring our analysis from human cells, we next evaluated the relationship between mRNA half-lives in mESCs and our residuals. We obtained reproducible mRNA half-life estimates for 6,266 genes in mESCs measured using SLAM-seq (Herzog et al., 2017). Similar to human cell types, residuals were positively correlated with mRNA half-lives (Pearson correlation = 0.24;  $p < 10^{-15}$ ; Figure 3F).

A distinguishing feature of mESCs relative to K562 cells is that more is known about the post-transcriptional regulatory mechanisms governing mRNA half-life. In particular, microRNAs (miRNAs) serve as strong candidates for further inquiry as they are guided by their sequence to bind and repress dozens to hundreds of mRNAs, mediating transcript degradation and thereby shortening an mRNA's half-life. The miR-290-295 locus, essential for embryonic survival, encompasses the most highly abundant microRNAs in mESCs. Under the control of a super-enhancer, the members of this miRNA cluster are expressed in a highly cell-type-specific manner and are thought to operate as key post-transcriptional regulators in ESCs (Whyte et al., 2013). We asked whether we could use our residuals to infer the endogenous regulatory roles of abundant miRNAs in mESCs. Defining a miRNA family as any miRNA sharing an identical seed sequence (as indicated by positions 2–8 relative to the miRNA 5' end [Agarwal et al., 2015]), we used existing small RNA sequencing data from mESCs (Denzler et al., 2016) (Gene Expression Omnibus: GSE76288) to quantify miRNA family abundances for the top 10 miRNA families. In addition to the miR-290-295 family, we detected other highly abundant families, including miR-17/20/93/106, miR-19, miR-25/32/92/363/367, miR-15/16/195/332/497, and miR-130/301; these miRNA families collectively comprised more than 75% of the total miRNA pool in mESCs (Figure 3G).

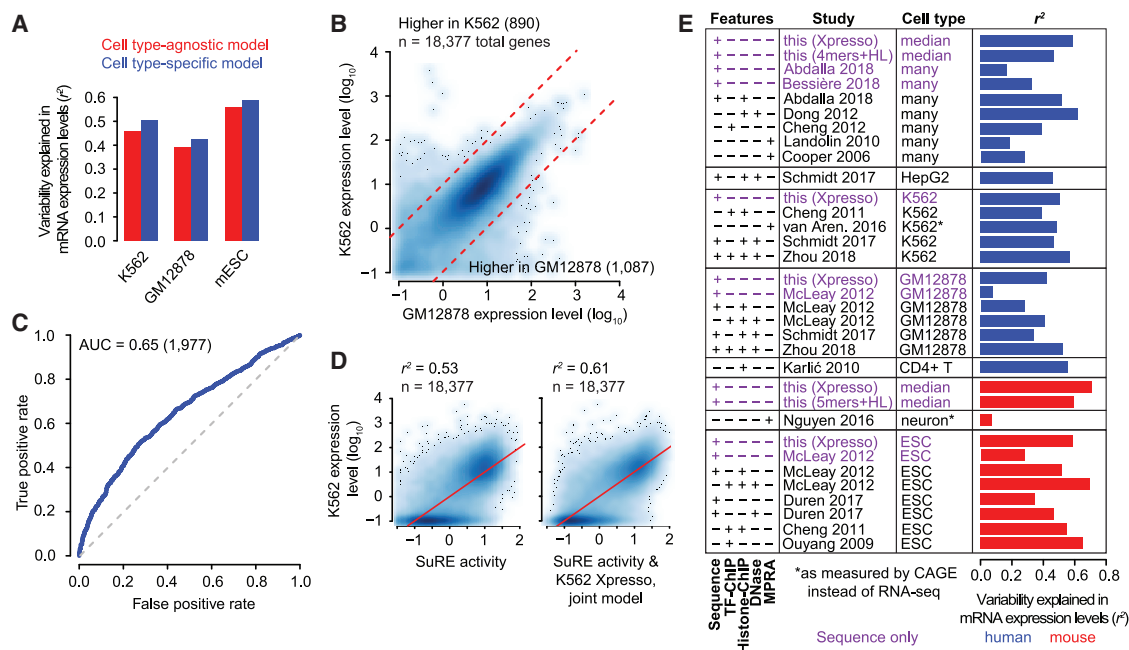
For each of the 222 miRNA families conserved across the mammalian phylogeny, which includes 7 of the 10 top miRNA families in mESCs, we assessed whether the predicted repression of targets correlated to our residuals. We used the TargetScan7 cumulative weighted context+ score (CWCS) (Agarwal et al., 2015) to rank predicted, conserved targets for the subset of mRNAs expressed in mESCs, assigning a CWCS of zero for non-targets. The miRNA family with the most highly ranked Spearman correlation corresponded to that of the miR-291a-3p/294/295/302abd family (Figure 3H), which was also the most highly expressed miRNA family in ES cells. The sign of this correlation was consistent with expectation because targets with more highly negative CWCSs (corresponding to predicted targets with greater confidence) had negatively shifted residual values. More generally, the 22 miRNA families comprising the highest 10% of Spearman correlations were strongly enriched for the seven miRNA families highly abundant in mESCs ( $p < 10^{-5}$ , Fisher's exact test; Figure 3H). Our results thus reinforce the finding that highly abundant miRNAs mediate target suppression (Mullokandov et al., 2012) and provide an alternative, fully computational method to infer highly active endogenous miRNA families in specific cell types solely from primary sequence and gene expression data. Collectively, we anticipate that cell-type-specific quantitative models for any arbitrary cell type can serve as a useful hypothesis generation engine for the characterization of active regulatory regions in the genome and key regulators such as miRNAs, including for cell types in which histone ChIP and small RNA sequencing data are limited or unavailable.

### Performance of Cell-Type-Specific Xpresso Models

To further characterize the ability of Xpresso to learn cell-type-specific expression patterns, we evaluated the relative performance of cell-type-agnostic and cell-type-specific models in predicting cell-type-specific mRNA levels. In all three cases considered, models trained on the cell type of origin outperformed those trained on median expression levels by about 3%–5% (Figure 4A). We next compared our K562 and GM12878 Xpresso models to evaluate how well these models could discriminate cell-type-specific mRNAs. We identified a cohort of 1,977 mRNAs enriched by at least 10-fold in one of these cell types relative to the other (Figure 4B). A binary classifier based upon the difference in predictions from each cell type could discriminate these cell-type-specific mRNAs modestly better than chance expectation (AUC = 0.65; Figure 4C).

Next, we sought to estimate the maximum possible performance for predicting gene expression from promoter sequences alone. A genome-wide MPRA measuring autonomous promoter activity in K562 cells, called Survey of Regulatory Elements (SuRE), linked 200 bp to 2-kb regions of the genome to an episomally encoded reporter to measure the transcriptional potential of regulatory sequences (van Arensbergen et al., 2017). SuRE, therefore, provides a means of assessing the regulatory information held in promoters that is independent of the influence of genomic context and distal regulatory elements. We observed that SuRE activity in the  $\pm 500$ -bp promoter region around a TSS was highly correlated to mRNA expression levels in K562 cells ( $r^2 = 0.53$ ; Figure 4D).





**Figure 4. Cell-Type-Specific Models Are Competitive with Methods Based upon Experimental Data**

(A) Variance explained in mRNA expression levels in each of three cell types (human K562 and GM12878 cells as well as mouse embryonic stem cells), using models trained on median expression levels (i.e., cell-type-agnostic model) or on the matched cell type (i.e., cell-type-specific model). The  $r^2$  shown is derived from the entire dataset, using the nested cross-validated predictions of each strategy.

(B) Relationship between mRNA expression levels in GM12878 and K562 cells. The dotted red lines correspond to the threshold used to call cell-type-specific genes, corresponding to a 10-fold change in expression in one cell type relative to the other. The number of cell-type-specific genes surpassing that threshold is indicated in parentheses.

(C) Performance of a classifier that uses the difference between cell-type-specific predictions to distinguish mRNAs whose expression is strongly enriched in either GM12878 or K562 cells.

(D) Shown in the left panel is the relationship between SuRE (van Arensbergen et al., 2017), a massively parallel reporter assay (MPRA) to measure autonomous promoter activity, and mRNA expression levels in K562 cells. Shown in the right panel is the relationship between a joint SuRE and Xpresso model and mRNA expression levels in K562 cells.

(E) Comparison of our sequence-only models to those derived from alternative strategies reported in the literature, often trained using a variety of cell-type-matched experimental datasets (i.e., as input or during intermediate training stages), such as those based upon ChIP of transcription factors or specific histone marks, DNase hypersensitivity measurements, and MPRA. Methods using nothing more than genomic sequence to predict expression are highlighted in purple, and results in human and mouse are shown in blue and red, respectively.

The comparable  $r^2$  achieved by SuRE measurements and the K562-specific Xpresso model in predicting K562 expression levels ( $r^2 = 0.53$  and  $0.51$ , respectively) provided a unique opportunity to evaluate how well our model captured the experimentally measurable information regarding a promoter's transcriptional activity. To assess the level of information shared between SuRE measurements and Xpresso predictions, we built a joint model to predict K562 levels ( $r^2 = 0.61$ ; Figure 4D). This modest 10% increase in performance over Xpresso alone indicates that Xpresso was able to learn the major sources of sequence-encoded information that explain mRNA expression levels.

### Predictive Models Perform Competitively with Models Using Experimental Data

We next evaluated the performance of Xpresso relative to an assortment of baseline and pre-existing models that attempted to predict mRNA levels, both with and without the consideration of mRNA half-life features. For the baseline models, we attempted to predict median expression level using simple  $k$ -mer counts in

the  $\pm 1,500$ -bp promoter region, the presence of predicted TF binding sites given known motifs available in the JASPAR database, or joint models considering both (Figure S4A). These models were trained using simple multiple linear regression and evaluated on the same test set as that used in Figure 2. Varying the  $k$ -mer size from  $k = 1$  to 6, we found the greatest gain in performance occurring between  $k = 1$  and 2 in both species. Consideration of known TF binding sites in a joint model at best only marginally improved performance, although a model considering these binding sites alone performed as well as a model based on 2-mers.

All of the models benefitted from the additional consideration of half-life features. We evaluated the coefficients associated with a model considering only half-life features to assess the relative contribution of individual features. The features most strongly associated with increased steady-state mRNA abundance in both the human and mouse corresponded to ORF exon density and 5' UTR GC content, followed by weaker associations to 5' UTR length and ORF length. In contrast, intron length was negatively associated with mRNA abundance (Figure S4B).

Overall, although our baseline models demonstrated that models built upon simple features could perform surprisingly well, our hyperparameter-tuned Xpresso model improved upon those models by 11.2% and 11.7% in human and mouse, respectively (Figure S4A). Our nested 10-fold cross-validation results further verified that Xpresso performed significantly better than the best alternative *k*-mer-based approach in both the human and mouse (Figure S4C).

Next, we compared our best baseline and Xpresso models with the reported results from existing models described in the literature, delineating five categories based upon the types of features used either as input data or as intermediate training stages: (1) those using nothing more than sequence features, which included our method and three others (Abdalla et al., 2018; Bessière et al., 2018; McLeay et al., 2012); (2) those using MPRA to measure promoter activity (van Arensbergen et al., 2017; Cooper et al., 2006; Landolin et al., 2010; Nguyen et al., 2016); (3) those using the binding signal of TFs at promoter regions, as measured by ChIP (Cheng et al., 2011, 2012; McLeay et al., 2012; Ouyang et al., 2009; Zhou et al., 2018); (4) those using the signal of histone marks, such as H3K4me1, H3K4me3, H3K9me3, H3K27Ac, H3K27me3, and H3K36me3 at promoters and gene bodies, as measured by ChIP (Abdalla et al., 2018; Cheng et al., 2011; Dong et al., 2012; Karić et al., 2010; McLeay et al., 2012; Schmidt et al., 2017; Zhou et al., 2018); and (5) those using the DNase hypersensitivity signal at promoters and nearby enhancers (Dong et al., 2012; Duren et al., 2017; McLeay et al., 2012; Schmidt et al., 2017; Zhou et al., 2018) (Figure 4E). Many of these models were trained and tested on cell lines, such as K562, GM12878, and mESCs, for which ChIP data are available for a multitude of histone marks and TFs. Thus, we were also able to compare the relative performance of our cell-type-specific models for these same cell lines. When matching for cell type, our Xpresso models nearly doubled the performance of alternative sequence-only models (Figure 4E). MPRA-based models exhibited a wide diversity of  $r^2$  values, although the genome-wide MPRA performed in K562 cells (van Arensbergen et al., 2017) performed comparably to Xpresso. Among all models examined, those using multiple forms of experimental data, such as TF ChIP, histone ChIP, and DNase, achieved the best  $r^2$  values of 0.62 (Dong et al., 2012) and 0.70 (McLeay et al., 2012) in human and mouse, respectively.

Overall, models using nothing more than genomic sequence are capable of explaining mRNA expression levels with as much predictive power as—and, often, more than—analogous models trained on abundant experimental data (Figure 4E). Our models have the advantage that they are simple to train on any arbitrary cell type, including those lacking experimental data, such as ChIP and DNase. Furthermore, sequence-only models can further augment the performance of existing models that predict mRNA levels in cell types for which experimental data are already available (Figure 4D).

### Xpresso Predicts Genome-Wide Patterns of Transcriptional Activity

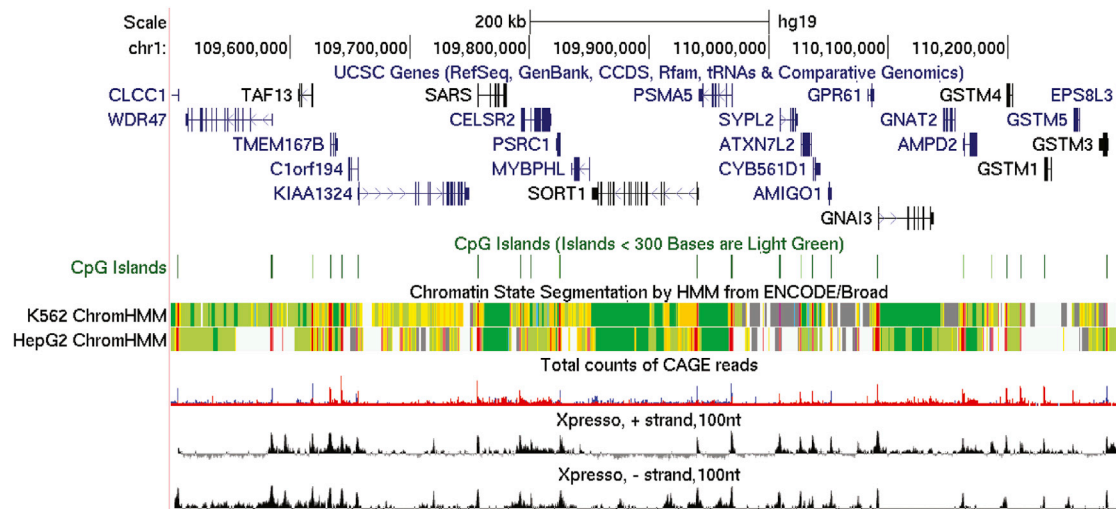
Convolutional neural networks have successfully been used to predict patterns of CAGE activity and histone ChIP signal throughout the genome. To accomplish that feat, a deep convolutional neural network was trained on genome-wide information

across entire chromosomes, using 131-kb windows that collectively encompassed ~60% of the human and/or mouse genomes (Kelley, 2019; Kelley et al., 2018). This led us to ask whether it was possible to predict a genome-wide CAGE signal using our cell-type-agnostic Xpresso model, which was trained, in contrast, on 10.5-kb windows comprising only ~5% of the human genome. As a proof of concept, we fixed the half-life features to equal that of the average gene and generated promoter-activity predictions in 100-bp increments along a randomly selected 800-kb region of the human genome that encompassed 20 genes, visualizing whether Xpresso could recapitulate the average pattern of CAGE activity across cell types (Figure 5). We observed that the Xpresso predictions faithfully reproduced the pattern of CAGE activity (FANTOM Consortium et al., 2014; Lizio et al., 2015) in this region. Peaks of high predicted transcriptional activity frequently corresponded to CpG islands (Gardiner-Garden and Frommer, 1987) and promoter regions across multiple cell types, as predicted by ChromHMM (Ernst and Kellis, 2012). Xpresso predicted similar expression signatures for both positive and negative DNA strands. Confirming that our results generalize across species, we observed consistent results on the 700-kb syntenic locus of the mouse genome (Figure S5).

### Xpresso Highlights the Importance of the Core Promoter Sequence in Predicting Expression Levels

Interested in ascertaining how our deep learning models could predict cell-type-agnostic gene expression levels with high accuracy, we developed a procedure to interpret the dominant features learned. Specifically, we tested four strategies intended to map the regions of the input space of deep learning models with the greatest contribution to the final prediction: (1) gradient \* input, (2) integrated gradients, (3) DeepLIFT (with rescale rules), and (4)  $\epsilon$ -LRP (Ancona et al., 2018). Each of these methods computed “saliency scores,” which represent a decomposition of the final prediction values into their constituent individual feature importance scores for each nucleotide in the input promoter sequences. We partitioned genes into four groups, including those predicted to be approximately non-expressed and three additional terciles (predicted low, medium, or high expression). We computed mean saliency scores for each of those groups and then computed the difference in mean scores relative to predicted non-expressed genes for each nucleotide position in the entire input window (i.e., 7 kb upstream of the TSS to 3.5 kb downstream). Averaging our results across our nested 10-fold cross-validated models, we discovered that the models had automatically learned to consistently rely upon local sequences from the 1-kb sequence centered upon the TSS to predict gene expression (Figures 6A and S6). Sequences in the core promoter (i.e., within 100 bp upstream of the TSS) best-discerned genes that were predicted to have high expression from those predicted to have medium expression (Figure 6A). Thus, from only genomic sequence and expression data, the model automatically learned spatial relationships that are consistent with experimental measurements (van Arensbergen et al., 2017).

Of note, because we used rectified linear units (ReLUs) in our networks, the  $\epsilon$ -LRP method resulted in identical results as the gradient \* input method (Ancona et al., 2018) (gradient \* input:



**Figure 5. Xpresso Predictions Recapitulate Transcriptional Activity across a 800-kb Region of Human Chromosome 1**

Shown are the Xpresso-predicted expression levels in 100-nt increments from the plus and minus strands of genomic sequence of the human hg19 genome assembly. Also shown are gene annotations, CpG island calls, ChromHMM genomic state segmentation calls among two cell types (with predicted promoter hidden states colored in red), and genome-wide CAGE signal aggregated among many cell types (with red indicating signal from the positive strand and blue indicating signal from the negative strand).

Figure S6A; e-LRP: data not shown). We also observed that the DeepLIFT method led to nearly identical results as the integrated gradients technique (integrated gradients: Figure S6B; DeepLIFT: data not shown), as observed previously in other contexts (Ancona et al., 2018). The results in the mouse model also emulated those of the human (Figures S6C and S6D), indicating that they generalize across species and do not depend upon the specific saliency scoring method used.

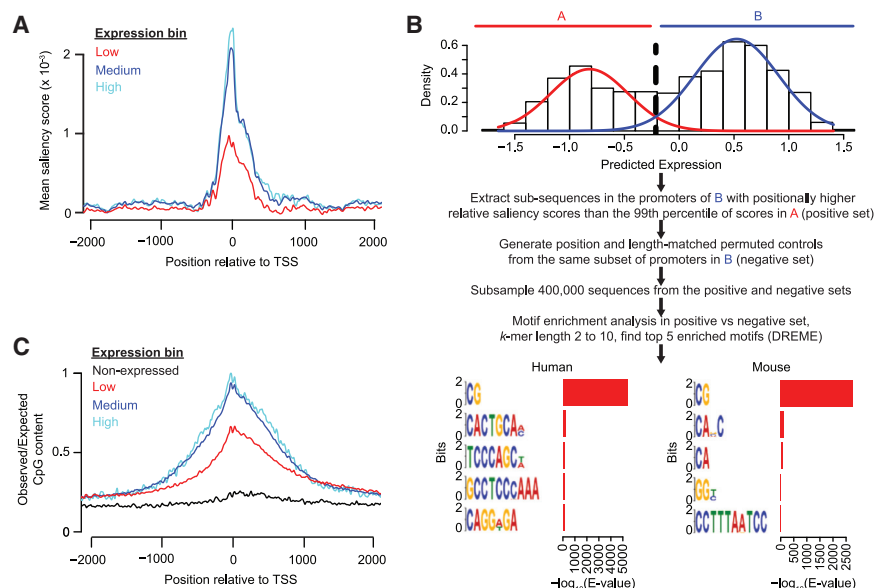
Although, in principle, it would be valuable to evaluate the saliency scores of individual promoters to dissect their regulatory grammar, we caution that the underlying methods are insufficiently robust to allow for mechanistic inferences at that resolution. Specifically, we have observed that, when we retrain convolutional architectures (specified by a fixed set of hyperparameters) on a fixed training set, the learned models can converge upon different sets of parameters, many of which can produce similar final predictions. However, the saliency scores from each learned model can differ substantially, indicating a lack of robustness; this problem and others are known limitations of neural network interpretation schemes (Ghorbani et al., 2017).

### CpG Dinucleotides Are the Dominant Signal Explaining Expression Levels

The ability of the model to automatically identify and heavily weight the core promoter in the expression prediction task naturally led to the question of which sequence motifs within this region were responsible for quantitatively defining expression level. We, therefore, devised a strategy to identify *k*-mers enriched in the genes predicted to be highly expressed (Figure 6B). The distribution of predicted gene expression levels was largely bimodal, allowing the partitioning of genes broadly into genes predicted to have low expression (class A) and high expression

(class B). We extracted sub-sequences from the promoters of class B whose saliency scores were higher than the 99th percentile of those observed at the same positions in the promoters of class A. To identify enriched *k*-mers in this set of sub-sequences, we devised an equivalently sized negative set of sub-sequences by permuting the extracted positions in class B to control for positional sequence biases. We then used DREME (Bailey, 2011) to identify *k*-mers enriched in our positive set relative to our negative set. Evaluating the E-values of the top five significantly enriched *k*-mers, we observed the dinucleotide CpG as enriched by orders of magnitude over the second best *k*-mer (Figure 6B), implicating it as a dominant factor discriminating highly expressed genes from lowly expressed ones. We repeated our procedure on genes in the top half of class B relative to its lower half, and identified an even stronger enrichment of CpG dinucleotides (data not shown). These observations suggest that both the mouse and human models predominantly use the spatial distribution of CpG dinucleotides surrounding the core promoter to predict the entire continuum of gene expression levels.

The model, therefore, arrives at a specific prediction: CpG dinucleotides are more enriched in the core promoters of highly expressed genes, relative to lowly expressed genes. To test that hypothesis, we evaluated the positional enrichment of all 16 possible dinucleotides around the TSS of genes in different gene expression bins, relative to chance expectation. Although CpGs are globally depleted, genes in higher gene expression bins preserved a greater fraction of CpGs closer to the TSS (Figure 6C). This property was true to a much lesser extent for other dinucleotides, with only AA/TT, CA/TG, and CC/GG dinucleotides being able to discriminate between highly and lowly expressed genes in both human and mouse genomes (Figure S7).



**Figure 6. The Spatial Distribution of CpG Dinucleotides in the Core Promoter Predicts Gene Expression Levels**

(A) Mean saliency scores (computed using the “gradient \* input” approach), relative to predicted non-expressed genes, for genes in three terciles of expression bins. Data have been loess-smoothed at the resolution of 100 bp.

(B) Strategy to ascertain enriched  $k$ -mers in genes predicted to be highly expressed. Shown are the top five significantly enriched  $k$ -mers retrieved from the human and mouse models.

(C) Positional enrichment of CpG dinucleotides relative to chance expectation in human promoters for different gene expression level bins. The expected frequency of CpG dinucleotides is computed as the joint probability of the composite C and G co-occurring based upon their frequency in the corresponding position and gene expression bin. Data have been loess-smoothed at the resolution of 100 bp.

## DISCUSSION

In this study, we demonstrate that a substantial proportion of variability across genes with respect to their steady-state mRNA expression levels is predictable from features derived solely from genomic sequence. In doing so, our work illustrates—as is the case for gene prediction—that the mathematical function linking genomic sequence to mRNA abundance is, in a large part, learnable *without* the use of additional sources of experimental data, such as those derived from DNase hypersensitivity, TF ChIP, histone ChIP, or MPRAs. Consistent with dogma and experiments (van Arensbergen et al., 2017), we find that the instructions governing transcriptional output are heavily enriched in a gene’s core promoter (more specifically, the  $\pm 500$  bp around the TSS). We establish Xpresso as an early initial attempt to confront the problem of gene-expression prediction from genomic sequence alone and anticipate that future algorithms can use our effort as a baseline model to improve upon this prediction task.

Our study provides a theoretical framework to further understand the fundamental question of how different modes of gene regulation contribute to steady-state abundance of mRNA. Querying the performance of the model while considering subsets of features associated with various mechanisms of gene regulation (e.g., mRNA decay and transcription rate) helped dissect their relative contributions to steady-state mRNA levels. Based on the proportion of variance explained by our model, thus far, we estimate that promoter sequences alone explain  $\sim 50\%$  of gene expression variability in humans. Collectively, these results reveal that *in silico* strategies to estimate the relative influence of various modes of gene regulation can approximate those more directly relying on experimental measurement.

Although our model makes substantial headway into predicting expression levels, between 40%–60% of variability still re-

mains unexplained, depending upon the cell type and species considered. We propose that the limitations of our model are also interesting in that they have the potential to inform and resolve the many layers of gene regulation that the model fails to capture. A residual analysis of highly expressed genes that the model underpredicts confirms that enhancers and super-enhancers have a measurably significant role in governing transcriptional programs. Incorporation of the effects of enhancers in the model is complicated by the difficulty of predicting which promoter(s) any given enhancer influences because these can be positioned hundreds of kilobases away and skip over genes, as well as the extent to which parameters such as distance modulate the level of enhancer-mediated activation. Such long-range dependencies are poorly modeled by convolutional neural networks. Although the incorporation of distal enhancers into the model has proved to be evasively difficult, the model can be used as a hypothesis generation engine to uncover additional gene regulatory mechanisms that further explain outliers. For example, our model provides a natural strategy to quantitatively rank candidate silenced and activated genes in different cell types in a way that prioritizes those that most heavily deviate from its predictions. We propose these rankings as a foundation to guide experimentalists interested in dissecting the layers of gene regulation that operate in their cell type of interest. In particular, such rankings could inform the selection of candidate regulatory elements to test by CRISPR-based functional assays (Gasparini et al., 2019; Klein et al., 2018) or MPRAs (van Arensbergen et al., 2017; Inoue and Ahituv, 2015).

We anticipate other potential applications for our model, such as in the realms of synthetic biology and medical genetics. For the former, models such as Xpresso could be used to design synthetic promoters with tunable levels of transcriptional activity. For the latter, methods such as Xpresso could be used to interpret the functional consequences of genetic mutations or indels within promoters on gene expression levels, e.g., to inform the



interpretation of genome-wide association studies and expression quantitative trait loci (eQTL) analysis. Although there is already promise for methods development in that direction (Zhou et al., 2018), it is currently premature to deploy such systems in the clinical realm because this requires much higher levels of predictive accuracy than are currently achieved by us or others.

Scanning large regions of the genome with our pipeline revealed a striking association between regions of high predicted transcriptional activity and CpG islands. Although CpG islands are a well-established feature of mammalian genomes that frequently demarcate promoter sequences (Antequera, 2003; Gardiner-Garden and Frommer, 1987), our results support the idea that the spatial positioning of CpGs around the core promoter is intimately associated with gene expression levels. Our data, therefore, reinforce the findings from MPRA that promoter regions enriched with CpG dinucleotides are functionally associated with increased gene expression levels (van Arensbergen et al., 2017; Hartl et al., 2019).

Looking forward, we envision the delineation of a set of mathematical functions for each cell type, which can accurately predict its mRNA expression level from genome sequence alone, to be a grand challenge for the field. As shown here, this framework will allow us to quantify and characterize the mechanisms of gene regulation of which we are aware and may draw our attention to ones that have yet to be discovered.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **LEAD CONTACT AND MATERIALS AVAILABILITY**
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- **METHOD DETAILS**
  - Gene Expression Data Collection and Pre-processing
  - Collecting Promoter Sequences and mRNA Half-life Features
  - Hyperparameter Optimization and Model Training
  - Whole-Transcriptome Predictions
  - SuRE MPRA Data
  - Baseline Models
  - Prediction on a Genomic Window
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.celrep.2020.107663>.

## ACKNOWLEDGMENTS

We thank Jacob Tome, Seungsoo Kim, and other members of the Shendure lab for critical commentary and helpful discussions, as well as Anandi Black for help with Xpresso logo artwork. This material is based upon work supported under NRSA NIH fellowship 5T32HL007093 (to V.A.) and NIH grants

5DP1HG007811, 5R01HG009136, 1UM1HG009408, 5R01CA197139, and 1R01HG006768 (to J.S.). J.S. is an investigator of the Howard Hughes Medical Institute.

## AUTHOR CONTRIBUTIONS

V.A. conceived of the study, conducted the research, and wrote the paper with feedback and supervision from J.S.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 25, 2018

Revised: June 11, 2019

Accepted: April 28, 2020

Published: May 19, 2020

## REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., and Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*, 1603.04467, Printed online March 14, 2016.
- Abdalla, M., Abdalla, M., McCarthy, M.I., and Holmes, C.C. (2018). A general framework for predicting the transcriptomic consequences of non-coding variation. *bioRxiv*. <https://doi.org/10.1101/27923>.
- Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4, 4.
- Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdrorf, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., et al. (2017). Ensembl 2017. *Nucleic Acids Res.* 45 (D1), D635–D642.
- Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838.
- Ancona, M., Ceolini, E., Oztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In 6th International Conference on Learning Representations (ICLR 2018). <https://research-collection.ethz.ch>.
- Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878.
- Antequera, F. (2003). Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci.* 60, 1647–1658.
- Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27, 1653–1659.
- Bergstra, J.S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, and F. Pereira, eds. (Curran Associates Inc.), pp. 2546–2554.
- Bergstra, J., Yamins, D., and Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In ICML'13: Proceedings of the 30th International Conference on Machine Learning, S. Dasgupta and D. McAllester, eds. (JMLR).
- Bessière, C., Taha, M., Petitprez, F., Vandel, J., Marin, J.-M., Bréhélin, L., Lèbre, S., and Lecellier, C.-H. (2018). Probing instructions for expression regulation in gene nucleotide compositions. *PLoS Comput. Biol.* 14, e1005921.
- Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441, 349–353.
- Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* 298, 1039–1043.



- Cheng, C., Yan, K.K., Yip, K.Y., Rozowsky, J., Alexander, R., Shou, C., and Gerstein, M. (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.* 12, R15.
- Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., Yan, K.K., Dong, X., Djebali, S., Ruan, Y., et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* 22, 1658–1667.
- Chollet, F. (2015). Keras. GitHub. <https://github.com/fchollet/keras>.
- Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* 16, 1–10.
- Denzler, R., McGeary, S.E., Title, A.C., Agarwal, V., Bartel, D.P., and Stoffel, M. (2016). Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Mol. Cell.* 64, 565–579.
- Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigó, R., Birney, E., and Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 13, R53.
- Duren, Z., Chen, X., Jiang, R., Wang, Y., and Wong, W.H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. USA* 114, E4914–E4923.
- Edfors, F., Danielsson, F., Hallström, B.M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., and Uhlén, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* 12, 883.
- Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216.
- FANTOM Consortium and the RIKEN PMI and CLST (DGT); Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberer, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470.
- Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* 354, 769–773.
- Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282.
- Gasparini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 176, 1516.
- Ghorbani, A., Abid, A., and Zou, J. (2017). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI Press)*. <https://doi.org/10.1609/aaai.v33i01.33013681>.
- Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018.
- Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774.
- Hartl, D., Krebs, A.R., Grand, R.S., Baubec, T., Isbel, L., Wirbelauer, C., Burger, L., and Schübeler, D. (2019). CG dinucleotides enhance promoter activity independent of DNA methylation. *Genome Res.* 29, 554–563.
- Herzog, V.A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T.R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S.L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* 14, 1198–1204.
- Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics* 106, 159–164.
- Jain, D., Baldi, S., Zabel, A., Straub, T., and Becker, P.B. (2015). Active promoters give rise to false positive ‘Phantom Peaks’ in ChIP-seq experiments. *Nucleic Acids Res.* 43, 6959–6968.
- Karlič, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* 107, 2926–2931.
- Kelley, D.R. (2019). Cross-species regulatory sequence activity prediction. *bioRxiv*. <https://doi.org/10.1101/660563>.
- Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26, 990–999.
- Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* 28, 739–750.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006.
- Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2017). CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* 35, 561–568.
- Klein, J.C., Chen, W., Gasparini, M., and Shendure, J. (2018). Identifying Novel Enhancer Elements with CRISPR-Based Screens. *ACS Chem. Biol.* 13, 326–332.
- Krebs, W., Schmidt, S.V., Goren, A., De Nardo, D., Labzin, L., Bovier, A., Ulas, T., Theis, H., Kraut, M., Latz, E., et al. (2014). Optimization of transcription factor binding map accuracy utilizing knockout-mouse models. *Nucleic Acids Res.* 42, 13051–13060.
- Landolin, J.M., Johnson, D.S., Trinklein, N.D., Aldred, S.F., Medina, C., Shulha, H., Weng, Z., and Myers, R.M. (2010). Sequence features that drive human promoter function and tissue specificity. *Genome Res.* 20, 890–898.
- Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2, e270.
- Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al.; FANTOM consortium (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 16, 22.
- Marco, E., Meuleman, W., Huang, J., Glass, K., Pinello, L., Wang, J., Kellis, M., and Yuan, G.C. (2017). Multi-scale chromatin state annotation using a hierarchical hidden Markov model. *Nat. Commun.* 8, 15011.
- Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 44 (D1), D110–D115.
- McLeay, R.C., Lesluyes, T., Cuellar Partida, G., and Bailey, T.L. (2012). Genome-wide in silico prediction of gene expression. *Bioinformatics* 28, 2789–2796.
- Moorthy, S.D., Davidson, S., Shchuka, V.M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N.N., and Mitchell, J.A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* 27, 246–258.
- Mullokandov, G., Baccarini, A., Ruza, A., Jayaprakash, A.D., Tung, N., Israelow, B., Evans, M.J., Sachidanandam, R., and Brown, B.D. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods* 9, 840–846.
- Nguyen, T.A., Jones, R.D., Snavey, A.R., Pfenning, A.R., Kirchner, R., Hemberg, M., and Gray, J.M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* 26, 1023–1033.
- Ouyang, Z., Zhou, Q., and Wong, W.H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* 106, 21521–21526.
- Parker, S.C., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al.; NISC Comparative Sequencing Program; National Institutes of Health Intramural

- Sequencing Center Comparative Sequencing Program Authors; NISC Comparative Sequencing Program Authors (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. USA* **110**, 17921–17926.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.
- Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilienky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330.
- Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J.K., Ebert, P., Nordström, K., Barann, M., Sinha, A., et al. (2017). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* **45**, 54–66.
- Schofield, J.A., Duffy, E.E., Kiefer, L., Sullivan, M.C., and Simon, M.D. (2018). TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods* **15**, 221–225.
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* **473**, 337–342.
- Segal, E., and Widom, J. (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.* **10**, 443–456.
- Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M.S. (2009). Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res.* **16**, 45–58.
- Spies, N., Burge, C.B., and Bartel, D.P. (2013). 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23**, 2078–2090.
- Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. USA* **110**, 18602–18607.
- van Arensbergen, J., FitzPatrick, V.D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H.J., and van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153.
- Vogel, C., Abreu, Rde.S., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6**, 400.
- Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319.
- Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G.C. (2017). Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285–299.e5.
- Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934.
- Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179.

## STAR★METHODS

### KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Ensembl v90 BioMart	Aken et al., 2017	<a href="http://uswest.ensembl.org/biomart/martview/9f79d60b2f33f0f2df1099520516c8ce">http://uswest.ensembl.org/biomart/martview/9f79d60b2f33f0f2df1099520516c8ce</a>
BEDTools	Quinlan and Hall, 2010	<a href="https://bedtools.readthedocs.io/en/latest/">https://bedtools.readthedocs.io/en/latest/</a>
Keras 2.0.8	Chollet, 2015	<a href="https://keras.io/">https://keras.io/</a>
TensorFlow 1.3.0	Abadi et al., 2016	<a href="https://www.tensorflow.org/">https://www.tensorflow.org/</a>
Other		
Resource website	This paper	<a href="https://xpresso.gs.washington.edu">https://xpresso.gs.washington.edu</a>
Reproducible codebase	This paper	<a href="https://github.com/vagarwal87/Xpresso">https://github.com/vagarwal87/Xpresso</a>
Google Colab	This paper	<a href="https://colab.research.google.com/gist/vagarwal87/bdd33e66fa2c59c41409ca47e7132e61/xpresso.ipynb">https://colab.research.google.com/gist/vagarwal87/bdd33e66fa2c59c41409ca47e7132e61/xpresso.ipynb</a>

### LEAD CONTACT AND MATERIALS AVAILABILITY

#### Lead Contact

Material and resource requests may be addressed to: Jay Shendure ([shendure@uw.edu](mailto:shendure@uw.edu)).

#### Materials Availability

This study did not generate new unique reagents.

#### Data and Code Availability

Original/source data for Figures 3 and 4 in the paper is available in Table S1. Associated code necessary to perform a hyperparameter search, train, and test the deep learning models is available at <https://github.com/vagarwal87/Xpresso>, our website (<https://xpresso.gs.washington.edu>), and associated Google Colab iPython notebook.

### METHOD DETAILS

#### Gene Expression Data Collection and Pre-processing

We retrieved a matrix of normalized expression values for protein-coding mRNAs across 56 tissues and cell lines from RNA-seq data gathered and quantified by the Epigenomics Roadmap Consortium (<https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz>) (Roadmap Epigenomics Consortium et al., 2015).

For mouse gene expression data, we gathered all ENCODE RNA-seq datasets satisfying the following constraints: i) datasets corresponded to “polyA-selected mRNA RNA-seq” or “total RNA-seq,” ii) reads were mapped to the *Mus musculus* mm10 genome assembly, iii) files were “tsv” files corresponding to gene-level quantifications, iv) biosamples were not treated with “DMS,” “LPS,” or “ $\beta$ -estradiol,” v) files were not derived from samples with “low replicate concordance,” “low read depth,” “insufficient read depth,” or “insufficient read length.” Only samples corresponding to the first replicate of each tissue or cell line were utilized. In total, 254 mouse RNA-seq datasets passed these criteria. For cell-type-specific questions in the mouse, we used gene expression data from mESCs that were computed previously (Ouyang et al., 2009).

For each species, we computed the median expression level across all cell types for each gene, and transformed all gene expression values  $y$  such that:  $\hat{y} \leftarrow \log_{10}(y + 0.1)$  to reduce the right skew of the data. Quantile normalizing the samples of the mouse gene expression matrix prior to computing the median values resulted in nearly an identical set of median expression levels (i.e., 0.9996 correlation before and after quantile normalization), making this step optional.

One-to-one human-to-mouse orthologs were acquired from the Ensembl v90 BioMart (Aken et al., 2017) by extracting the “Mouse gene stable ID” and “Mouse homology type” with respect to each human gene.

### Collecting Promoter Sequences and mRNA Half-life Features

Human and mouse promoter CAGE peak annotations were downloaded from the FANTOM5 consortium's UCSC data hub ([https://fantom.gsc.riken.jp/5/datahub/hg38/peaks/hg38.cage\\_peak.bb](https://fantom.gsc.riken.jp/5/datahub/hg38/peaks/hg38.cage_peak.bb), hg38 genome build; [https://fantom.gsc.riken.jp/5/datahub/mm10/peaks/mm10.cage\\_peak.bb](https://fantom.gsc.riken.jp/5/datahub/mm10/peaks/mm10.cage_peak.bb), mm10 genome build) (Fantom Consortium et al., 2014; Lizio et al., 2015). The best peak corresponding to each promoter, labeled with the keyword "p1@", was extracted. HUGO gene names or gene name synonyms were converted into Ensembl IDs using the Ensembl v90 BioMart, HGNC ID tables (<https://www.genenames.org/download/custom/>), and Mouse Genome Informatics gene model tables ([http://www.informatics.jax.org/downloads/reports/MGI\\_Gene\\_Model\\_Coord.rpt](http://www.informatics.jax.org/downloads/reports/MGI_Gene_Model_Coord.rpt)).

Gene annotations for protein coding genes were derived from Ensembl v90 (hg38 genome build) (Aken et al., 2017). Only protein-coding genes were carried forward for analysis, with the following genes filtered out as sources of bias: i) genes located on chrY, whose gene expression depended upon whether their cells of origin were male or female, ii) histone genes, whose expression was mis-quantified due to their mRNAs lacking poly(A) tails, therefore being undersampled in poly(A)-selected RNA-seq libraries. Out of all transcripts corresponding to each gene, the one with the longest ORF, followed by the longest 5' UTR, followed by the longest 3' UTR was chosen as the representative transcript for that gene. The G/C content and lengths of each of these functional regions (i.e., 5' UTRs, ORFs, and 3' UTRs), intron length, and ORF exon junction density (computed as the number of exon junctions per kilobase of ORF sequence) were gathered as additional features associated with mRNA half-life (Sharova et al., 2009; Spies et al., 2013). All length-related features were transformed such that:  $\hat{x} \leftarrow \log_{10}(x + 0.1)$  to reduce the right skew, and along with gene expression levels were then z-score normalized by subtracting their respective mean values and dividing by their standard deviations. The starting coordinate of the first exon of the representative transcript was defined as that gene's transcriptional start site (TSS). The vast majority of mRNAs possessed a dominant CAGE peak; if this was so, the TSS was re-centered to the coordinates of the CAGE peak. While considering CAGE data helped to improve our model, its use was optional as the performance of our models was only modestly worse when considering only Ensembl TSS annotations ( $r^2$  of 0.54 instead of 0.59 in the human). The  $\pm 10$  kilobase sequence centered at the TSS was extracted as the putative promoter region to consider. Intermediate steps such as extracting sequences from the genome or converting between bed formats were executed with BEDTools (Quinlan and Hall, 2010).

### Hyperparameter Optimization and Model Training

Matching gene expression levels to promoter sequences resulted in a total of 18,377 and 21,856 genes in human and mouse, respectively. All continuous variables were mean-centered and scaled to have unit variance, and promoter sequences were one-hot encoded into a boolean matrix. For each species, genes were then randomly partitioned into training, validation, and test sets such that the validation and test sets were allotted 1,000 genes each. We defined the objective function as the minimum mean squared error achieved on the validation set across 10 epochs of training. For the best set of hyperparameters specifying the neural network structure, we trained ten independent trials, and selected the parameters derived from the specific trial and epoch that minimized the validation MSE as our final model. For each trial, parameters were first randomly initialized by sampling from a Glorot normal distribution (Chollet, 2015). Next, the Adam optimizer was used to search for a local minima (Chollet, 2015). The search was performed for 100 epochs but cancelled if a lower validation MSE was not discovered within 7 epochs of the best model discovered so far for that trial. If a nested cross-validation strategy was implemented, we performed an identical strategy for each of the 10-folds of the data using the respective training and validation sets of the fold. The following software packages were required for model training and testing: Keras 2.0.8 (Chollet, 2015), TensorFlow 1.3.0 (Abadi et al., 2016), CUDA 8.0.61, cuDNN 5.1.10, and the Anaconda2 distribution of Python. We initialized a hyperparameter search space to specify the model architecture (Table 1), and used Hyperopt (Bergstra et al., 2013) to search for an optimal set of hyperparameters. All models were trained on an NVIDIA Quadro P6000 GPU equipped with 24Gb of video RAM.

### Whole-Transcriptome Predictions

To predict expression levels for all annotated genes, we implemented a nested 10-fold cross-validation procedure. Specifically, we partitioned the dataset into 10 equally sized bins. For each fold, we i) reserved 1/10 of the data as a test set, 1000 genes as a validation set, and the remaining genes as a training set; ii) trained 10 independent models until convergence, iii) selected the model with the minimum validation mean squared error, and iv) generated a prediction on the test set. We then concatenated all of the predictions together that were derived from the best model from each of the ten folds of the data (Table S1).

### SuRE MPRA Data

Pre-processed genome-wide, stranded SuRE MPRA data mapped to hg19 was acquired as bigwig files from GEO record GSE78709 (van Arensbergen et al., 2017). To compute SuRE activity at a specified promoter, we extracted the mean SuRE signal over covered bases on the correct strand as the gene, centered at 1000bp around the TSS, using utilities provided in the UCSC genome browser ("bigWigAverageOverBed -sampleAroundCenter=1000") (Kent et al., 2002). For the TSS annotations, we utilized our set of CAGE-corrected TSSs lifted over from hg38 to hg19 using liftOver (Kent et al., 2002), supplemented with TSSs for genes annotated in Gen-code release 27/Ensembl v90 for any missing IDs (<https://www.encodegenes.org/releases/27lift37.html>) (Harrow et al., 2012).

### Baseline Models

To train baseline models for comparison (Figure S4), we merged the training and validation sets used initially for hyperparameter optimization and model training. For each gene, we first extracted the  $\pm 1500$ bp window centered at each TSS and defined this as the promoter. For  $k$ -mer-based models, we counted the frequency of all  $k$ -mers occurring in the promoter region, varying  $k$  from 1 to 5 and 6 for human and mouse, respectively. To train transcription-factor-based models, we scanned promoters using FIMO (Grant et al., 2011) using positional weight matrices derived from the JASPAR 2016 Core Vertebrate set (Mathelier et al., 2016). Default parameters were used for the search, except that the set of promoter sequences to compute a first order Markov background model for the search. For the transcription factors matched to the promoters, we populated a binary matrix with a 1 if a significant motif was detected for the promoter, and 0 otherwise. We then trained multiple linear regression models explaining median mRNA expression levels as a function of i) the collection of  $k$ -mer counts, ii) the binary matrix of JASPAR matches, or iii) both of the former. These models were trained both with and without half-life features, with the  $r^2$  (i.e., the coefficient of determination, computed as the square of the Pearson correlation) evaluated on the test set.

### Prediction on a Genomic Window

We extracted 10.5Kb sequences tiling across the 600 Kb and 700 Kb regions of the human and mouse genomes, respectively, in 100bp increments (Genome coordinates: chr1:109500000-110300000, hg19 genome build; chr3:107800000-108500000, mm10 genome build). The “bedtools makewindows” (parameters “-w 10500 -s 100”) was used to generate these windows, and “bedtools getfasta” to extract the sequences (Quinlan and Hall, 2010). Predictions were then made using our cell-type-agnostic Xpresso model trained upon median gene expression data, using zero values for all half-life features (with zero corresponding to the mean values as these features had been z-score normalized).

### QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical details, including the statistical tests used, what  $n$  represents, and dispersion and precision measures (e.g., mean, median, SD, confidence intervals), can be found in the legends of the corresponding figures.