Context-aware sequence-to-activity model of human gene regulation

Ekin Deniz Aksu[1], Martin Vingron[1]
[1] Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract

Sequence-to-activity models have been very successful in predicting gene expression, chromatin accessibility, and epigenetic marks from DNA sequences alone. However, current state-of-the-art models have a fundamental limitation: they cannot extrapolate beyond the cell types and conditions included in their training dataset. Here, we introduce a new approach that is designed to overcome this limitation: Corgi, a new context-aware sequence-to-activity model that accurately predicts genome-wide gene expression and epigenetic signals, even in previously unseen cell types. We designed an architecture that strives to emulate the cell: Corgi integrates DNA sequence and trans-regulator expression to predict the coverage of multiple assays including chromatin accessibility, histone modifications, and gene expression. We define *trans*-regulators as transcription factors, histone modifiers, transcriptional coactivators, and RNA binding proteins, which directly modulate chromatin states, gene expression, and mRNA decay. Trained on a diverse set of bulk and single cell human datasets, Corgi has robust predictive performance, approaching experimental-level accuracy in gene expression predictions in previously unseen cell types, while also setting a new state-of-the-art level for joint cross-sequence and cross-cell type epigenetic track prediction.

Introduction

Sequence-to-activity models of gene regulation have been highly successful in predicting gene expression, chromatin accessibility, and epigenetic marks from DNA sequences alone[1]. Trained on the genome and large functional genomics datasets, these models learn gene regulatory grammar and can accurately predict the activity of previously unseen sequences[2,3]. They are useful in predicting effects of non-coding variants and generating synthetic sequences[4–6]. However, current state-of-the-art models have a fundamental limitation: they cannot extrapolate beyond the cell types and conditions included in their training dataset. Since they operate on DNA sequence alone, these models are blind to the biological context in which the sequence is interpreted.

Context-aware sequence-to-activity models overcome this fundamental limitation by enabling generalization to previously unseen cellular contexts. These models operate on not only the DNA sequence but also a context vector. The context vector is computed from knowledge about the cell state: usually gene expression or chromatin accessibility. Previous models used transcription factor (TF) expression[7–10], ATAC-seq[11] and single-cell multiome (joint RNA and ATAC)[12] data. However, these models are limited either in their predictive performance, low number of predicted assay types, or reliance upon known TF binding motifs.

Here we introduce **Corgi**, a new context-aware sequence-to-activity model that predicts genome-wide coverage of 16 different assays including RNA-seq, chromatin accessibility, DNA methylation and histone modifications in human cells. Corgi stands for "**Co**ntext-aware **R**egulatory

Genomics **I**nference" and it continues the canine-themed naming tradition in sequence-to-activity models. Importantly, we designed Corgi's architecture to imitate cellular gene regulation. In the cell, *trans*-regulatory proteins bind DNA and RNA to drive transcription, modulate chromatin state and control mRNA decay. So for its context vector, Corgi uses the expression of *trans*-regulators, which are defined as transcription factors, transcriptional co-activators, chromatin modifiers, and RNA binding proteins.

We address a central challenge in context-aware sequence-to-activity models: it is not straightforward to integrate DNA sequence with the context vector due to the difference in their dimensions. The context vector is one-dimensional, while the DNA sequence has two dimensions (length and features). We leverage the feature-wise linear modulation[13] (FiLM) technique to combine the two. FiLM was previously employed in various tasks such as image generation[14] and speech recognition[15], and has proven technically and conceptually appropriate for this task.

After training on a diverse set of human bulk and single-cell contexts, we show that Corgi can accurately generalize to previously unseen contexts. Corgi approaches experimental-level accuracy in predicting gene expression in new cell types, while accurately predicting DNase-seq, ATAC-seq, DNA methylation and various histone marks. This allows the imputation of epigenetic signals using only the genome and *trans*-regulator expression data, for example in cases where low cell numbers hinder the quality of ChIP-seq experiments. Remarkably, Corgi's performance remains robust on the toughest benchmarks which test the model on held-out sequences and held-out cell types.

Methods

Data sources

The bulk of our data comes from the ENCODE consortium. We scanned the ENCODE database for human samples which have any RNA-seq experiment, plus either one of i) CAGE or RAMPAGE data, ii) DNase-seq and at least 1 histone modification ChIP-seq data, or iii) any combination of at least 4 experiments. It is important that all the experiments are performed on replicates of the same sample. Data that did not comply with ENCODE data quality standards and were marked with an error tag were excluded. In total, we identified 373 samples with these constraints.

We enriched our dataset with CAGE-seq data from the FANTOM5 consortium. In order to extract maximum information from the available data, we manually curated a list of samples where 65 FANTOM5 CAGE-seq experiments were matched with previously selected ENCODE samples. In addition to these, we selected 119 new samples from FANTOM5 and added them to our set.

In order to capture variation in *trans*-regulator expression more extensively, we expanded our set using single-cell sequencing experiments. We included 71 pseudobulked single-cell RNA-seq samples from Tabula Sapiens that are matched with single-nucleus ATAC-seq experiments from Catlas. ATAC-seq data was downloaded from Catlas directly as bigwig files, and corresponding matched gene expression data was from Fu. et al[11]. Finally, we used 10x multiome single-nucleus joint RNA-seq and ATAC-seq data from brain and peripheral blood mononuclear cells. Data was

downloaded from 10x Genomics and after processing and generating pseudobulks, 17 samples were added to the set.

In total, we have 580 biological samples in our dataset, representing a wide range of human cells and tissues. The full list can be found in Supplementary Table 1.

Defining *trans*-regulatory factors

To determine the set of *trans*-regulatory factors, we need to consider all genes that may play a role in gene regulation. More precisely, we need to find the genes that directly or indirectly regulate DNA accessibility, histone modifications, chromosome folding, DNA methylation or RNA expression. We defined this set to be transcription factors, transcriptional co-activators, chromatin modifiers and RNA binding proteins. Transcription factors directly bind to DNA to regulate transcription and thus constitute the first step of interpretation of regulatory regions on the genome. We used the list of TFs as compiled by the Aerts Lab[16] which includes 1892 TF genes in the human reference genome hg38[17]. Secondly, transcriptional co-activators interact with transcription factors and regulate transcription and chromatin state, without directly binding to the DNA. The list of transcriptional co-activators was curated from the literature and contains 324 genes, including CREB binding protein, genes involved in the mediator complex and TRIM family proteins[18,19]. Third, chromatin modifiers regulate chromatin state via histone modifications and DNA methylation, thus directly influencing many of our output tracks. We used the dbEM database which contains 167 genes, such as histone acetyltransferases and DNA methyltransferases[20]. Finally, we decided to include all known RNA binding proteins into the set of *trans*-regulators. Since these proteins regulate splicing, transport and decay of RNA molecules, they directly influence RNA-seq coverage. We used a manually curated version from the RBPWorld database[20,21] which includes 741 genes. After filtering out duplicate entries and genes not contained in the ENCODE RNA-seq processing pipeline, we are left with a set of 2891 *trans*-regulatory factor genes. The full list of genes can be found in Supplementary Table 2.

We have chosen gene expression as a proxy for the in vivo activity of these proteins, due to the high availability of gene expression data. The unit of gene expression is defined as the logarithm of transcript per million (TPM) values, which is ordinarily calculated from RNA-seq experiments.

Training-test split

In order to prevent any data leakage across training and test sets, we carefully split our samples into subsets that are as dissimilar as possible. First, we organized the 580 samples into 40 clusters through hierarchical clustering according to their gene expression levels, and left out entire clusters out of the training set until a suitable number of validation and test samples were reached. Second, for all selected validation and test samples, similar cell types or tissues still left in the training set were determined by a keyword match and matching samples were taken out of the training set. Lastly, a random sample was selected from all training clusters into an “easy-test” set, for a testing scenario where the exact samples are unseen by the model, but similar cell types and tissues were seen during training.

In summary, we have 488 training samples, 28 validation samples, 27 easy-test samples and 37 test samples. The test set mainly consists of liver tissue, hepatocytes, testis, prostate, skin, keratinocytes and embryonic stem cell lines.

The genomic sequences were split according to Borzoi's folds[3], with fold 3 reserved for testing and fold 4 reserved for validation similar to Borzoi. Furthermore, regions overlapping with gene bodies of *trans*-regulatory factor genes were excluded from the validation and test folds, as their TPM values are used in the input. This resulted in 11496 sequences of length 524,288 bp for training, 1462 sequences for validation and 1437 for testing, in a tiled fashion across most of the human reference genome. In total, we train on 2.2 Gbp of DNA sequence and test on 292 Mbp.

Data preprocessing

Input data: The model uses expression levels of *trans*-regulators in units of log(TPM). Importantly, the diversity in gene expression experiments in our training data needs to be addressed because of systematic differences in TPM distributions across different assays. We have 4 types of gene expression assays in our training set: total bulk RNA-seq, polyA-capture bulk RNA-seq, pseudobulked single-cell RNA-seq and CAGE-seq. Since we want to represent biological context, or the cell state, as a universal *trans*-regulator expression vector, ideally this should not depend on the type of gene expression assay that was performed.

In order to control for batch effects across different assays, we harmonized gene expression TPMs of coding genes via taking the logarithm, then quantile normalization and finally batch correction using pyComBat[22]. Quantile normalization and batch correction was performed by using the total RNA-seq data as reference, as this is the most abundant type of gene expression assay in our data. This allows the user to directly use log(TPM) measurements from common RNA-seq assays without the need for further preprocessing.

Output signals training data: ENCODE, FANTOM5 and Catlas data was downloaded as processed in the bigwig format. For the 10x multiome experiments, data was downloaded in the h5ad and BAM formats. This data was preprocessed using scanpy[23], celltypist[24], multiVI[25], sinto[26] and deepTools[27] in order to generate separate bigwig coverage files for each pseudobulk.

In order to normalize coverage signals across experiments with different dynamic ranges, coverage values were scaled and soft clipped with assay-specific parameters similar to the processing pipeline of Borzoi. Afterwards, coverage values were aggregated at a 64 bp resolution by taking the sum of CAGE and RAMPAGE tracks, the square root of the sum for RNA-seq tracks and the mean for the rest of the tracks.

Model architecture

Corgi builds upon a hybrid convolutional-transformer architecture, similar to recent state-of-the-art DNA sequence models[28]. The input 524 kb one-hot encoded DNA sequence is first processed through several layers of convolution and max pooling operations, aiming to extract meaningful features from the sequence (e.g. TF binding motifs, TATA boxes) and to reduce the length of the sequence by decreasing resolution. This step computes a matrix of sequence features that

summarizes the local *cis*-regulatory landscape. As a completely new feature, Corgi processes the *trans*-regulatory context vector using a multi-layer perceptron which computes shift and scale parameters for FiLM layers, which represent the *trans*-features. Subsequently, *cis*- and *trans*-features are integrated by the FiLM layers, which apply an affine transformation to the *cis*-features matrix along its feature axis. These are followed by transformer layers that aim to capture long-range genomic interactions such as enhancer-promoter loops. The final layer outputs coverage predictions for 16 different genomic assays at a 64 bp resolution.

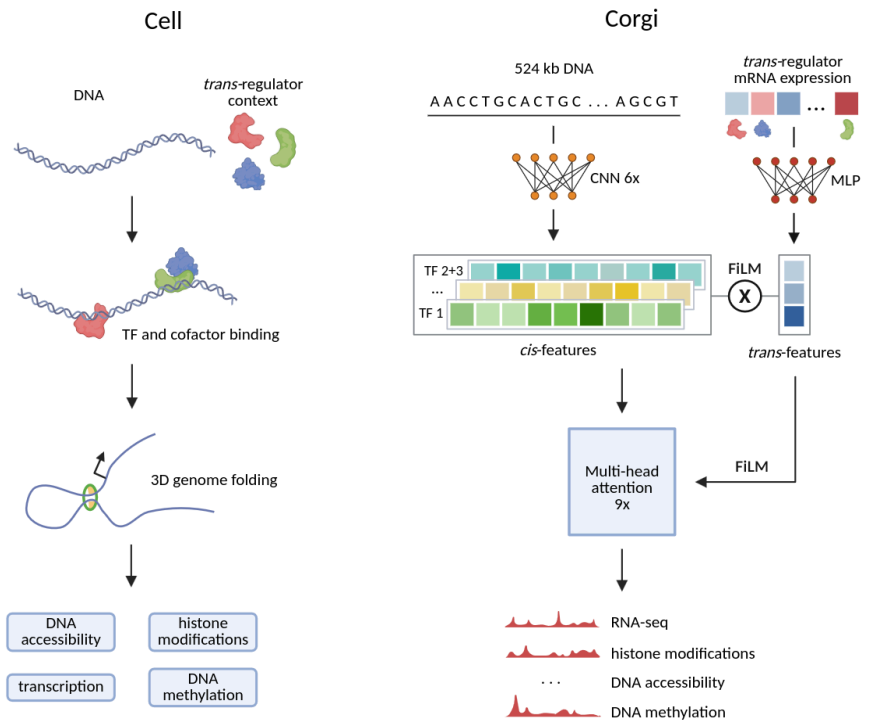

Figure 1. Corgi's architecture imitates cellular gene regulation
We designed an architecture that strives to emulate how cells read the regulatory genome using *trans* factors. The model first extracts *cis*-regulatory features via convolutional layers and *trans*-regulatory features via an MLP operating on *trans*-regulator expression. These are integrated using FiLM layers, which apply feature-wise affine transformations akin to biophysical "affinity × concentration" models. Instead of single TFs, these features are combinations of *trans*-regulators learned by the model. Afterwards, transformer blocks with FiLM layers capture long-range genomic interactions in a cell-type specific manner.

This design mirrors cellular gene regulation: for a regulatory element to be active, both a *cis*-element and its corresponding *trans*-regulator must be present. Mathematically it is a multiplication of *cis*- and *trans*- features, and if either one of them is zero, the product is also zero. This is conceptually similar to models like activity-by-contact (ABC)[29], which multiply activity with contact. In Corgi, these features can be combinations of different factors that are learned during the training process. There are two FiLM layers within the convolutional blocks and two within the transformer blocks, ensuring that *trans*-regulatory context can be integrated into the sequence at multiple points. Each FiLM layer has an independent set of parameters.

The transformer block consists of nine layers that all use multi-head attention at a 64 bp resolution. We believe that this resolution represents a good balance between memory cost and prediction quality, as it allows Corgi to precisely identify splice sites while avoiding the need for upsampling (such as U-net in Borzoi). We use FlashAttention v2 which has speed and memory benefits compared to older implementations of multi-head attention.

Corgi predicts coverage tracks for the central 393 kb of the input window. The predicted assays are DNase-seq, ATAC-seq, ChIP-seq for multiple histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, and H3K79me2), CTCF ChIP-seq, DNA methylation (whole genome bisulfite sequencing) and transcriptomic assays (CAGE-seq, RAMPAGE-seq, total RNA-seq, polyA-enriched RNA-seq, 10x scRNA-seq). Transcriptomics tracks are strand-specific (i.e. reads aligning to the positive and negative strands are split), with the exception of scRNA-seq. In total, Corgi has 22 channels in its final layer.

Model training

We trained Corgi with an adapted version of the Poisson multinomial loss function similar to Borzoi, which decomposes the loss to magnitude and shape terms and allows for weighting the shape loss. In order to standardize the loss values across different output tracks, we balanced the loss by weighting each channel's loss with a learnable weight parameter, inspired by Kendall et al[30]. Thus we minimize a new objective (poisson multinomial * weights + log[weights]) which empirically yields improved performance in some tracks, especially CAGE-seq (Supplementary Fig. 2).

The model was trained for 5 epochs on NVIDIA A100 GPUs with a batch size of 2, using the AdamW optimizer with learning rate 0.0001 and weight decay 0.01 for the multi-layer perceptron and 0.001 otherwise. Mixed precision was used as it is required by FlashAttention. Furthermore, we used grouped query attention with 4 groups and 8 heads with a total of 192 dimensions. Rotary positional encodings were applied to the first 128 dimensions, similar to Flashzoi[31]. A cosine annealing learning rate scheduler was used which ramps up over 3000 warmup steps and then decays the learning rate. Due to the lack of computational resources, we did not perform any hyperparameter tuning, which means that the model can possibly be tuned for increased performance.

The number of epochs is considerably lower than most other models, since we consider all pairs of training sequence and training cell type as one training sample and thus our training set size becomes around 5.6 million. In one epoch, each DNA sequence is seen approximately 500 times, which can quickly lead to overfitting before the model has time to learn the rules of cell type specific gene regulation. To combat this redundancy, we applied several dynamic augmentation techniques. First, we use the custom hg38 version from Borzoi in which Gnomad SNPs with a higher allele frequency than the hg38 reference allele were edited in. For all such alleles overlapping an input DNA sequence, each one is randomly edited in with a probability of 0.5. The reasoning is that such alleles could have been the reference allele, and they would mostly have small effects on chromatin and gene regulation, if any. Afterwards, as in previous work, input DNA sequences were randomly shifted up to 3 bp to the left or right, and then randomly reverse complemented.

Results

First, we assessed the impact of TPM harmonization by focusing on samples with at least two different RNA expression assay available. Harmonization reduced batch effects between different assay types, and improved correlation across genes in a given sample between two different assay types, with the most striking examples being CAGE-seq vs RNA-seq (Supplementary Figure 1).

Then, we evaluated Corgi's performance under three different benchmarking regimes: cross-cell type, cross-sequence, and cross-both. In the cross-cell type setting, the model predicts on held-out cell types using sequences seen during training, thus testing generalization to new biological contexts. In the cross-sequence benchmark, predictions are made on unseen genomic regions within training cell types. This is the only benchmark that traditional non-context-aware models can be evaluated in. Finally, the stringent cross-both setting evaluates predictions on both unseen cell types and unseen sequences. This benchmark represents a very hard challenge for sequence-to-activity models, as accurate prediction requires generalization capabilities across new cell types and new sequences. For all benchmarks we applied test-time data augmentation by averaging predictions for slightly shifted and reverse complemented sequences.

Corgi has strong overall performance across a multitude of experiments (Fig. 2). In the cross-cell type setting, Corgi achieves an average Pearson's *r* of 0.84 for DNase-seq predictions across genomic bins. The performance changes considerably across assays: histone modification ChIP-seq tracks show the highest variability ranging from 0.39 for H3K9me3 up to 0.83 for H3K4me3. The model also predicts gene expression very accurately, achieving an average of 0.59 for CAGE and 0.79 for bulk RNA-seq tracks. Single-cell RNA-seq cannot be predicted as well as the bulk experiments, probably due to low abundance of data in our training set and possibly due to its inherent sparsity. Finally, DNA methylation can be predicted with very high accuracy (mean *r* 0.92), and in some samples approaching near-perfect predictions.

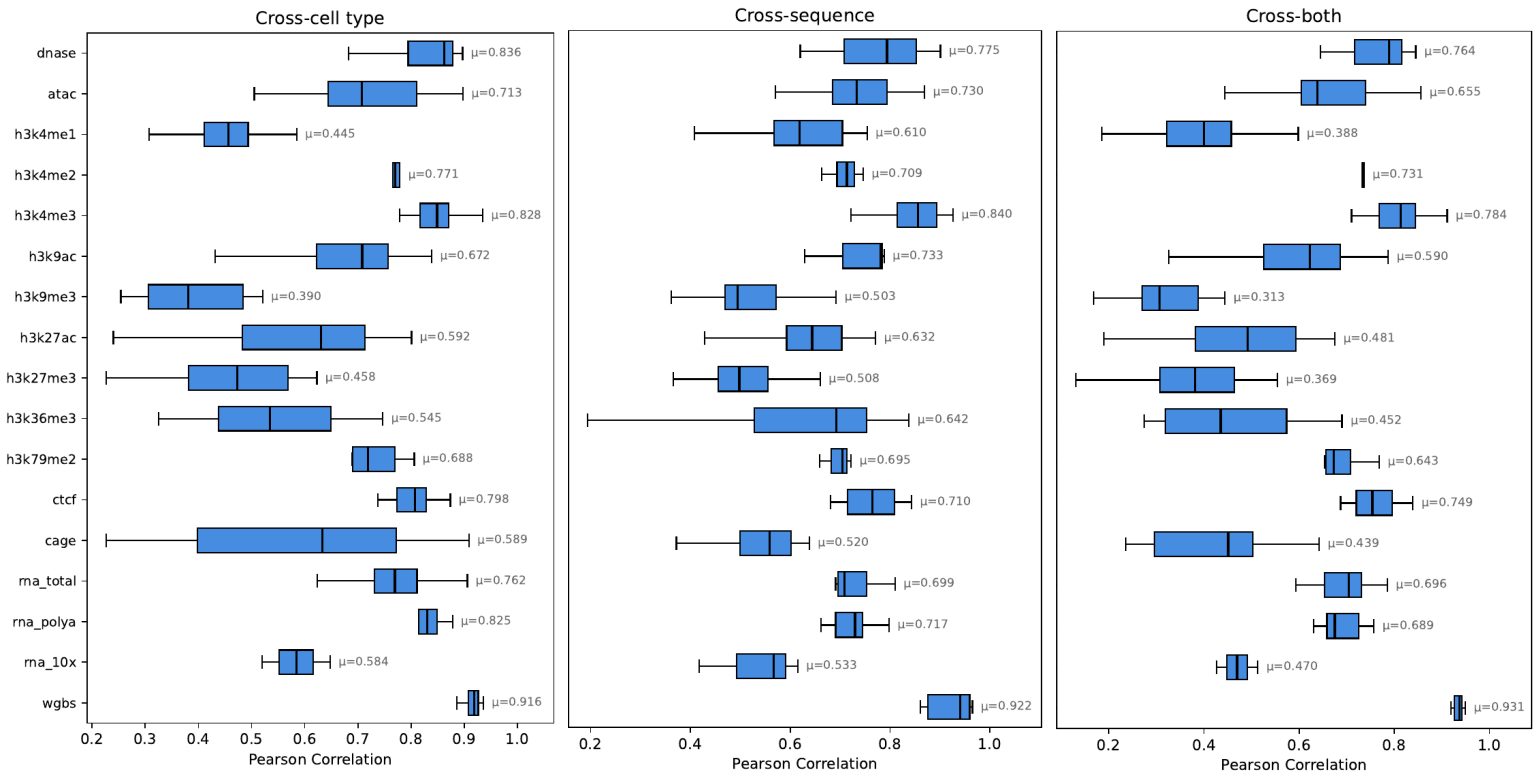

Figure 2. Genomic coverage prediction performance across different tracks.
Boxplots showing model performance across different assays in cross-cell type, cross-sequence and cross-both settings. Correlations between predictions and ground truth data across genomic bins are reported, with the variance in the boxplots coming from different biological samples. Pearson's *r* values are reported with the mean values shown next to the boxplots.

In the cross-sequence setting, Corgi maintains strong predictive performance, with increases in performance in 10 assays including all histone modifications, and modest drops in performance in seven assays. DNA accessibility and bulk RNA-seq predictions are not far from their cross-cell type counterparts, with 0.78 and 0.71 respectively.

Under the most challenging cross-both setting, which involves generalization to both unseen sequences and unseen cell types, Corgi still performs robustly, reaching 0.76 for DNase and 0.69 for RNA-seq. DNA methylation is predicted again with very high accuracy, reaching 0.93.

These results significantly outperform previously reported figures for models that use DNA sequence and transcription factor expression. For example, EpiGePT, the leading model in this category, reports a median Pearson's *r* of 0.515 for DNase-seq in a cross-both benchmark, while Corgi has a performance of 0.76. However, we note that this comparison is not directly one-to-one: the authors' reported figure comes from a test which was performed under different conditions, such as different input sequence length (128 kb vs 524 kb), different sampling (EpiGePT used a random partitioning for both sequence and cell types, while we use a stringent partitioning that minimizes data leakage), and a different subset of ENCODE data. However, we were unable to reliably run EpiGePT using either its web server or the provided code.

For gene expression prediction, it is often more informative to use log-transformed TPM values instead of the exact RNA-seq coverage across the gene. To simulate this, we computed gene-level expression predictions by aggregating coverage predictions over annotated exons and applied a log-transformation. Corgi achieves gene-level Pearson's *r* values of 0.85 in the cross-cell type setting, and 0.91 in the cross-sequence and cross-both settings (Fig. 3 a,b,c). Importantly, Corgi's performance is robust in different benchmarks and does not drop off when generalizing to held-out sequences or cell types. Gene expression predictions show that while predictions and measurements overlap significantly, there is a set of genes which show zero measured expression and nonzero predictions (Fig. 3d), though Borzoi's predictions also show a similar pattern (Supplementary Fig. 3 c,d).

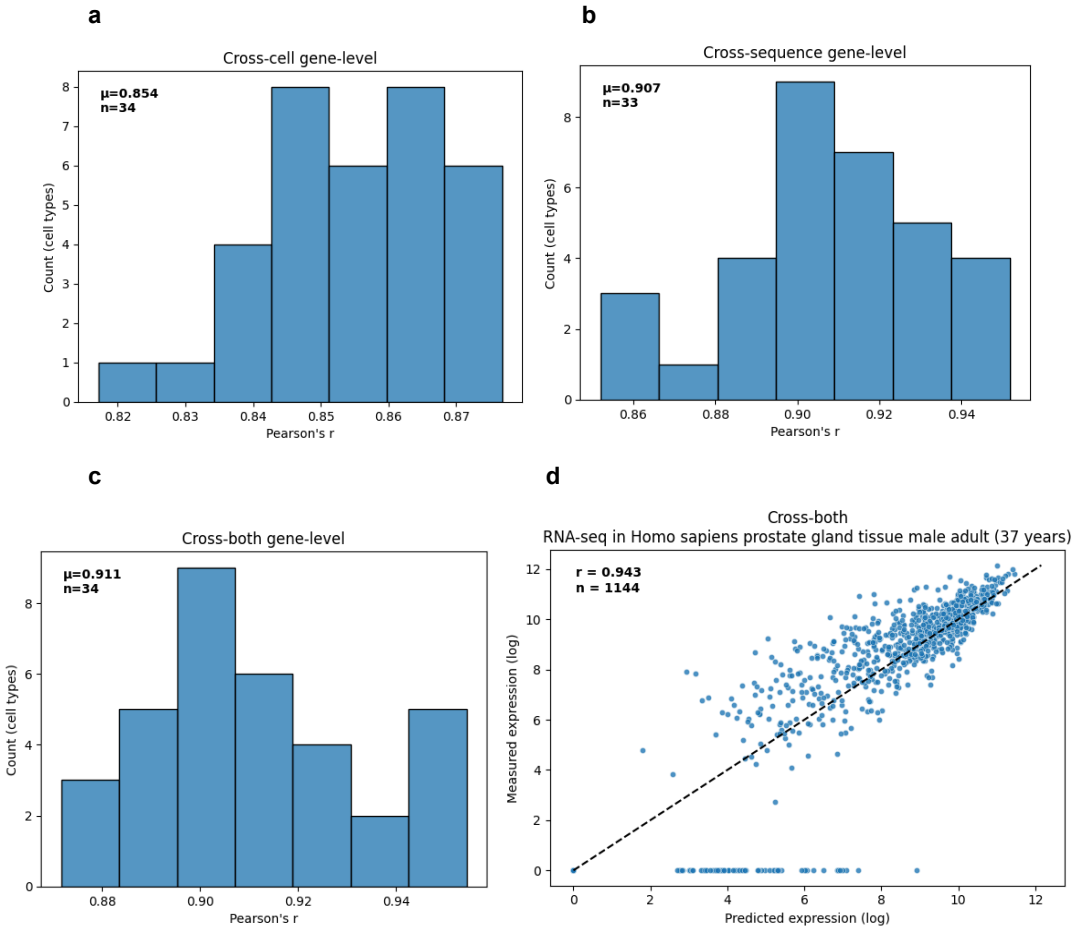

Figure 3. Gene-level expression prediction performance.
(a), (b) and (c): Histograms that show the distribution of gene-level expression prediction correlations across samples using Pearson's *r.* The prediction performance is robust across all three benchmarks. (d) Scatterplot showing the predicted vs measured expression gene expression values for RNA-seq in adult prostate tissue (cross-both benchmark). Each point represents the natural logarithm of total coverage overlapping exons. 1144 protein-coding genes are shown that are found within test sequences.

We next compared Corgi to the state-of-the-art DNA sequence-to-activity model Borzoi in a cross-sequence setting. To ensure a fair comparison, we selected 30 samples from our training set that matched specific tracks in the Borzoi training set. Both models were trained on the same

sequence folds, enabling a direct cross-sequence comparison. As Borzoi predicts at a 32 bp resolution, we applied average pooling to match the 64 bp resolution used by Corgi. Overall, performance across the set of tracks was similar (Fig. 4). Given that discrepancies in Pearson's *r* for some tracks may reflect minor differences in output signal scales between Corgi and Borzoi, we consider Spearman's *ρ* to be a more robust metric. Nevertheless, we report both metrics for reference. Gene-level expression correlations using Spearman's *ρ* were 0.86 for both Corgi and Borzoi, reaffirming that cross-sequence performance is comparable (Supplementary Fig. 3 a,b,e).

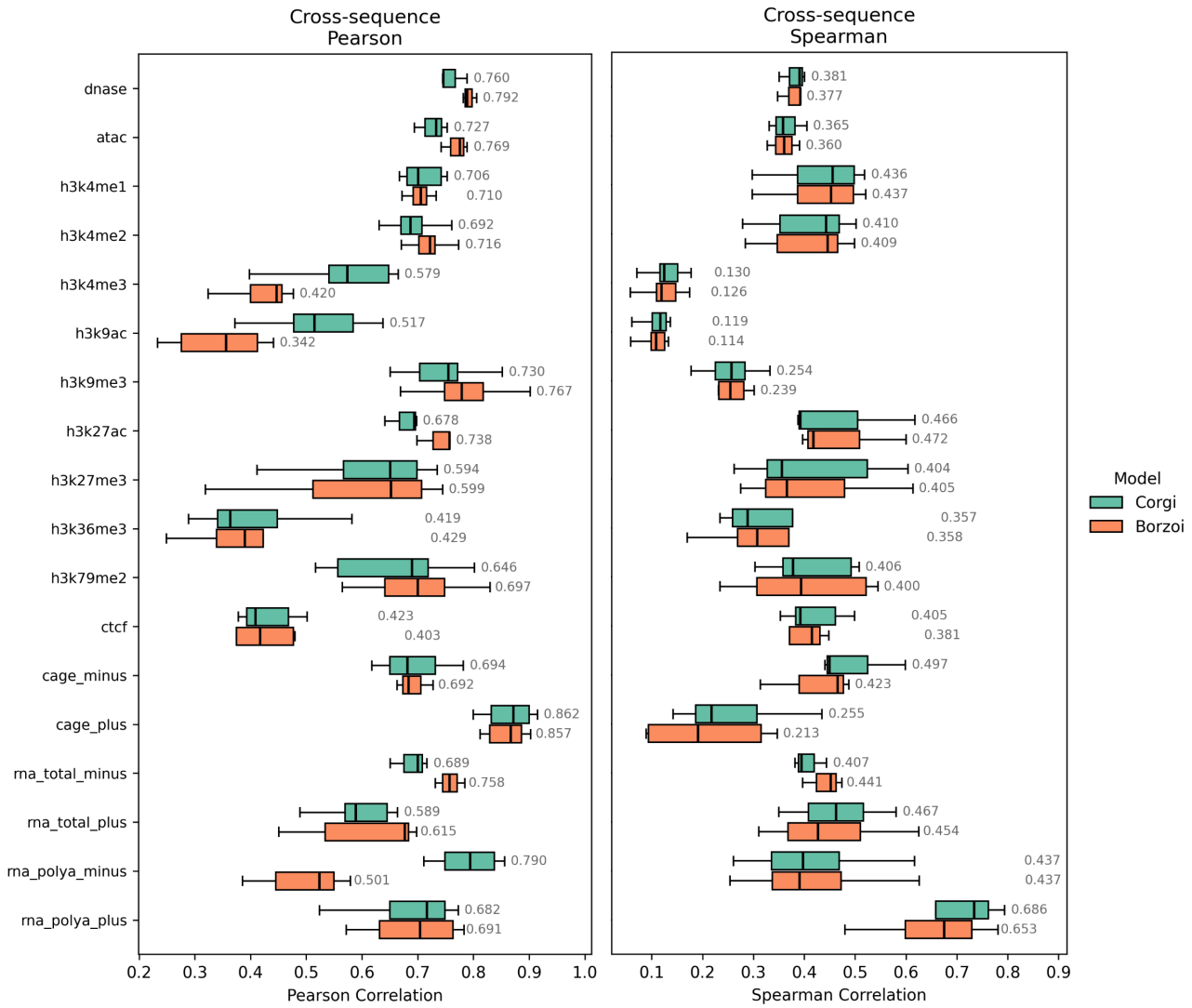

Figure 4. Cross-sequence performance comparison of Corgi and Borzoi
Boxplots showing model performance across 30 matched tracks in a cross-sequence setting. Pearson's *r* and Spearman's ρ values are reported, reflecting correlation between model predictions and ground truth data across genomic bins, with the mean values shown next to the boxplots. Corgi and Borzoi show comparable predictive performance across tracks.

A representative genome browser view illustrates Corgi's predictive capacity across a 35 kb region on chromosome 6, encompassing the *SESN1* locus (Fig. 5). This example is taken from the cross-both setting, which shows a held-out sequence in a held-out cell type. The model faithfully reconstructs a wide range of assays, including chromatin accessibility, histone modifications, transcription initiation, gene expression, and DNA methylation. Corgi captures both the sharper peaks characteristic of DNase hypersensitive sites as well as the broader domains marked by

active histone modifications such as H3K4me3 and H3K27ac. The predicted expression landscape across *SESN1* closely mirrors experimental RNA-seq and CAGE signals, with the exon-intron boundaries perfectly predicted in the RNA-seq track and the exact TSS predicted in the CAGE track. Finally, the model can also recapitulate DNA methylation patterns with high precision.

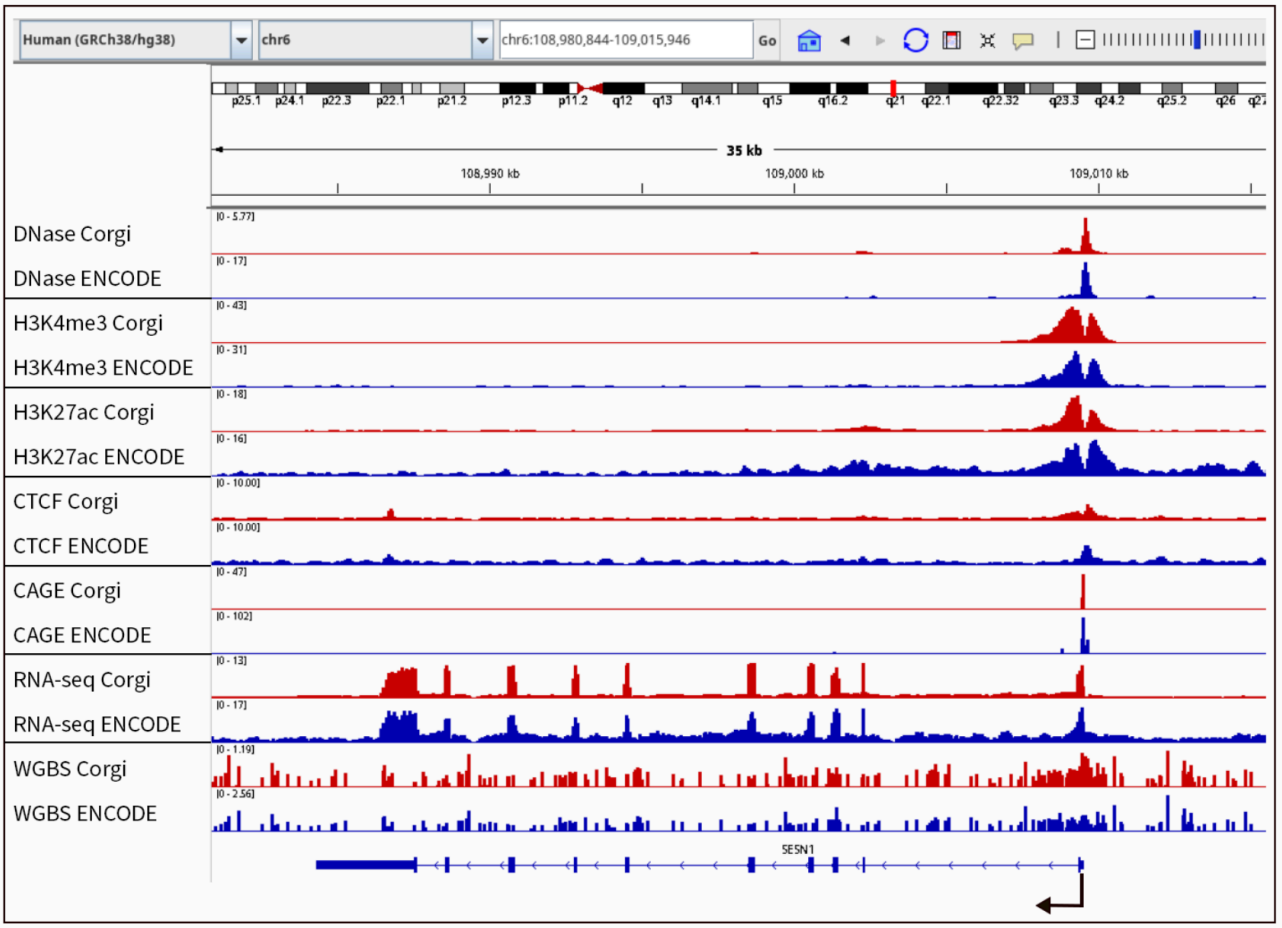

Figure 5. Genome browser view of predictions and ground truth data
Corgi predictions (red) and ENCODE ground truth data (blue) are shown in an example locus from chr6 containing the SESN1 gene in sample #323 (testis), from the cross-both benchmark. From top to bottom, DNase-seq, H3K4me3 ChIP-seq, H3k27ac ChIP-seq, CTCF ChIP-seq, CAGE-seq, RNA-seq and WGBS are shown. Remarkably, the model captures histone mark signal shapes, the site of the CAGE-signal, DNA methylation patterns, and exact exon-intron boundaries in RNA-seq.

Discussion

In this work, we present a novel approach that has the potential to shift the paradigm in sequence-to-activity models from using sequence alone to context-aware models that can generalize to new cell types as well as new DNA sequences. Compared to previous context-aware models, Corgi shows markedly increased accuracy, a larger context window, increased resolution and the most extensive output experiment types.

When compared to previous sequence-to-activity models such as Borzoi, the context-aware framework offers a possible advantage in using the same data in a more organized manner. Borzoi treats every genomics experiment as an independent observation, with its multi-task learning

framework. However, genomics data is highly correlated across samples: there are constitutively active regions that show a signal in many cell types and different assay types. Note that having separate output channels for each genomics experiment confers some technical advantages: batch effects across samples can be absorbed into the weights of the final convolutional layer. In addition, while Corgi's framework requires each output assay type to have a reasonable number of samples in the training data, there is no such constraint in Borzoi's framework. This allows Borzoi to be trained on a much larger dataset, including mouse sequences and data, and TF ChIP-seq experiments. Due to slight differences in *trans*-regulator genes, Corgi does not natively support predictions in the mouse genome. However, this problem can be circumvented by using orthologous genes. Another possibility is to fine-tune Corgi on the mouse genome.

Corgi needs cellular context in the form of *trans*-regulator expression, which includes transcription factors, transcriptional co-activators, chromatin modifiers and RNA binding proteins. This set of genes does not represent all *trans*-regulatory elements in the human genome. We purposefully excluded microRNAs and long non-coding RNAs, as including them would limit us to using samples with available total RNA-seq experiments and discard samples with polyA tail-enriched assays, which includes most single-cell protocols.

We believe that the expression of *trans*-regulators is the most "natural" representation of the cell state for sequence-to-activity models, as these regulators directly interpret the regulatory genome, modulate chromatin states and also regulate stability of RNA molecules. Other methods have used only transcription factors, ATAC-seq signal and abstract representations of the entire transcriptome, each having certain drawbacks. While constituting the largest subset of *trans*-regulators, transcription factors alone are not enough to regulate gene expression and chromatin on their own. Using ATAC-seq is a completely different approach, because it allows the context vector to be simply concatenated to DNA sequence, yielding five channels (four for the nucleotides, one for ATAC). This is a very rich source of information, for example such a model could use promoter accessibility as a baseline for gene expression predictions. However, ATAC (or DNA accessibility) cannot be easily manipulated experimentally, and its availability is lower than RNA-seq. Lastly, some models use an abstract representation of the cell state (scooby[12], DragoNNFruit[32]) which is not directly manipulatable and hard to interpret.

One of the key advantages of using *trans*-regulator expression is that *trans*-regulators can be experimentally manipulated with relative ease, using plasmid constructs for overexpression, knockdown or CRISPR knockout of key regulators. Corgi has the ability to simulate these experiments *in silico,* by running the model with a baseline cell type *trans*-regulator expression and manually changing the expression of selected genes to the desired value. This can also be done simultaneously with perturbations in any *cis*-regulatory sequence.

Additionally, Corgi can be run to impute missing experiments such as ATAC-seq or histone modification ChIP-seq, only using the reference genome and *trans*-regulator expression data from an RNA-seq experiment. This can be useful in situations where only RNA-seq data is available in a given context (rare cell type, patient sample etc.) and predictions of DNA accessibility, DNA methylation and histone modification ChIP-seq is desired.

References

1. Sasse, A., Chikina, M. & Mostafavi, S. Unlocking gene regulation with sequence-to-function models. *Nat. Methods* **21**, 1374–1377 (2024).
2. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
3. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.* 1–13 (2025).
4. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
5. de Almeida, B. P. *et al.* Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. *Nature* **626**, 207–211 (2024).
6. Taskiran, I. I. *et al.* Cell-type-directed design of synthetic enhancers. *Nature* **626**, 212–220 (2024).
7. Wnuk, K. *et al.* Deep learning implicitly handles tissue specific phenomena to predict tumor DNA accessibility and immune activity. *iScience* **20**, 119–136 (2019).
8. Nair, S., Kim, D. S., Perricone, J. & Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35**, i108–i116 (2019).
9. Liu, Q., Hua, K., Zhang, X., Wong, W. H. & Jiang, R. DeepCAGE: Incorporating transcription factors in genome-wide prediction of chromatin accessibility. *Genomics Proteomics Bioinformatics* **20**, 496–507 (2022).
10. Gao, Z., Liu, Q., Zeng, W., Jiang, R. & Wong, W. H. EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. *Genome Biol.* **25**, 310 (2024).
11. Fu, X. *et al.* A foundation model of transcription across human cell types. *Nature* **637**, 965–973 (2025).
12. Hingerl, J. C. *et al.* scooby: Modeling multi-modal genomic profiles from DNA sequence at single-cell resolution. *bioRxivorg* 2024.09.19.613754 (2024) doi:10.1101/2024.09.19.613754.

13. Perez, E., Strub, F., Vries, H. D., Dumoulin, V. & Courville, A. C. FiLM: Visual reasoning with a general conditioning layer. *National Conference on Artificial Intelligence* 3942–3951 (2017).
14. Brock, A., Donahue, J. & Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv [cs.LG]* (2018).
15. Yang, D.-H. & Chang, J.-H. FiLM conditioning with enhanced feature to the transformer-based end-to-end noisy speech recognition. in *Interspeech 2022* 4098–4102 (ISCA, ISCA, 2022).
16. TF List Aerts Lab. https://resources.aertslab.org/cistarget/tf_lists/allTFs_hg38.txt.
17. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
18. Talukdar, P. D. & Chatterji, U. Transcriptional co-activators: emerging roles in signaling pathways and potential therapeutic targets for diseases. *Signal Transduct. Target. Ther.* **8**, 427 (2023).
19. Alerasool, N., Leng, H., Lin, Z.-Y., Gingras, A.-C. & Taipale, M. Identification and functional characterization of transcriptional activators in human cells. *Mol. Cell* **82**, 677–695.e7 (2022).
20. Singh Nanda, J., Kumar, R. & Raghava, G. P. S. dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes. *Sci. Rep.* **6**, 19340 (2016).
21. Liao, J.-Y. *et al.* RBPWorld for exploring functions and disease associations of RNA-binding proteins across species. *Nucleic Acids Res.* **53**, D220–D232 (2025).
22. Behdenna, A. *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics* **24**, 459 (2023).
23. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
24. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
25. Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods* **20**, 1222–1231 (2023).
26. Stuart, T. Sinto. https://github.com/timoast/sinto.
27. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).

28. Barbadilla-Martínez, L., Klaassen, N., van Steensel, B. & de Ridder, J. Predicting gene expression from DNA sequence using deep learning models. *Nat. Rev. Genet.* 1–15 (2025).
29. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
30. Kendall, A., Gal, Y. & Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv [cs.CV]* (2017).
31. Hingerl, J. C., Karollus, A. & Gagneur, J. Flashzoi: An enhanced Borzoi model for accelerated genomic analysis. *bioRxiv* 2024.12.18.629121 (2024) doi:10.1101/2024.12.18.629121.
32. Schreiber, J. DragoNNfruit. https://github.com/jmschrei/dragonnfruit.

Acknowledgements

We thank Ibrahim Ilik for providing an extended list of RNA binding proteins. We thank the IT department of the Max Planck Institute for Molecular Genetics and the Max Planck Computing and Data Facility for the usage of computing resources. Figure 1 was made with BioRender with an institutional license.