Evaluating the use of sequence-to-expression predictors for personalized expression prediction

Parth Baokar
Nilah Ioannidis, Ed.

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-120
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-120.html

May 13, 2022

Evaluating the use of sequence-to-expression predictors for personalized expression prediction

by

Parth Baokar

A thesis submitted in partial satisfaction of the

requirements for the degree of

Masters of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nilah Ioannidis, Chair
Professor Liana Lareau

Spring 2022

Evaluating the use of sequence-to-expression predictors for personalized expression prediction

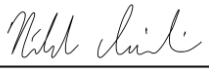
by Parth Baokar

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:



Professor Nilah Monnier Ioannidis
Research Advisor

5/13/22

(Date)

* * * * * * *

Professor Liana Lareau
Second Reader

May 10, 2022

(Date)

Abstract

Evaluating the use of sequence-to-expression predictors for personalized expression prediction

by

Parth Baokar

Masters of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Nilah Ioannidis, Chair

With rapid advances in deep neural network architectures, there has been recent interest in using these complex models to understand the regulatory factors that govern gene expression. Recent state-of-the-art models are trained to predict expression levels in different cell types from the reference genome sequence around the start site of each gene. These models explain a large fraction of the variation in expression across different genes in the genome, and have demonstrated an ability to recognize biologically relevant regulatory motifs. However, here we show that model performance is limited when applied to sequences from personal genomes to explain variation in expression across individuals. Our results suggest a relative insensitivity of these models to small but biologically meaningful perturbations in the input sequence. We also demonstrate that while the models identify some key sites of regulatory variation corresponding to those found in eQTL (expression quantitative trait loci) studies, they often fail to capture the correct direction of effect on expression. This work highlights potential shortcomings of these deep learning models when applied to personal genome interpretation in a clinical setting, and suggests further avenues of exploration for improving model performance on personalized genomes.

To my friends and family

Contents

List of Figures

Acknowledgments

I want to extend a huge thank you to Professor Nilah Ioannidis for mentoring me throughout the past few years. I came in as a student very fresh to research, and especially so to the field of machine learning, genomics, and the intersection between them. Although there were several hurdles in settling on the project, I'm incredibly grateful for the patience, advice, and insight that you've provided throughout the process. I also want to thank the rest of ni-lab for their helpful feedback and ideas. I especially want to thank Connie Huang and Richard Shuai for being such amazing students to work with on this project, and helping generate some of the results and figures used in this thesis.

Last but not least, I want to express my appreciation for my friends and family that provided me unwavering support these past few years, and gave me some much needed moments of respite from the Berkeley grind.

Chapter 1

Introduction

1.1 Background

Understanding the effects of genomic variation on transcriptional expression levels can provide key insights in functional genomics and personalized medicine. Traditional studies such as genome-wide association studies (GWAS) have affirmed the relationship between genetic variation and clinical phenotypes, identifying loci in the genome associated with phenotypic effects and providing an effect size for each [1]. However, the results of GWAS suffer from a few challenges in interpretation. First, the identified associated variants tend to be within non-coding regions of the genome, and the mechanisms governing the effects of non-coding variants are not well understood or easily studied [2, 3]. Additionally, the multiple testing burden imposes a strict significance threshold for determining associations, so GWAS can fail to capture intermediate frequency variants with small effects or rare variants with moderate effects [1]. Since these variants likely play an important role in disease, it limits the clinical impact of GWAS-based disease risk prediction [1]. Finally, the phenomenon of linkage disequilibrium obfuscates the true causal variants [4]. Variants at nearby positions tend to be inherited together and have correlated dosages, making it difficult to identify the true causal variant from statistical association studies (a process known as fine-mapping). An approach similar to GWAS is used to identify statistically-significant associations between genetic variants and translational output, or expression quantitative trait loci (eQTLs), but suffers from similar challenges.

As a result, several computational models for predicting expression directly from genomic sequences have been developed as a complementary tool for compiling and characterizing the effects of variants on gene expression. The regulation of transcription rate is controlled by a variety of mechanisms, involving the binding of transcription factors (TFs) to short DNA motifs [5], epigenomic modifications such as histone marks and methylation [6], and genomic regulatory elements such as promoters and enhancers [7]. The steady-state mRNA expression level of a given gene is also a function of the decay rate, which involves degradation mechanisms such as mi-RNA silencing or nonsense-mediated mRNA decay [8]. Advances in

deep learning techniques have unlocked the capability to model these complex interactions between different regulatory mechanisms, with the hope of uncovering new insights into biological pathways to motivate and guide future study. Here we focus on four specific sequence-to-expression prediction models.

1.2 Expression prediction models

Many state-of-the-art methods rely on deep convolutional neural networks (CNNs) to uncover important motifs within the genomic sequence. The input sequences are located around the transcription start site (TSS) of genes, as it has been shown that promoter sequences play a primary role in determining gene expression.

Xpresso. Xpresso [9] is one such method that consists of two convolutional blocks and two fully connected layers to predict gene expression levels. The model is trained on normalized RNA-seq data across 56 tissues and cell lines from the Epigenomics Roadmap Consortium [10]. There are two types of Xpresso models, one where the model is trained on median expression data across cell types, and another where models are trained on cell-type specific expression. As found through cross-validation, the optimal input sequence for Xpresso is a small 10.5kb region asymmetrically centered around the transcription start site (TSS). The median cell type model achieved an R^2 of 0.54 when predicting median expression, while cell-type specific models were initially only able to achieve a best R^2 of 0.51 when predicting cell-type specific expression. The residuals were shown to be consistent with our biological understanding of cell-type specific regulatory elements, where overrepresentation of certain enhancers and silencers skewed the distribution of residuals. Additionally, including mRNA half-life estimates as features improved performance, by capturing the impact of post-transcriptional regulation as well. When visualizing saliency scores computed for the input DNA sequence, it was found that the model assigns the greatest importance to the 1kb sequence (the core promoter sequence) centered on the TSS, in accordance with experimental results.

Basenji2. While promoters are close in proximity to the TSS, enhancer regions are frequently found within a 100-150kb region upstream and downstream [11]. In order to capture these more distal interactions, models would need to increase their receptive field to consider a significantly larger input sequence. Basenji2 [12, 13] makes progress towards this goal, increasing the input sequence to 131kb and the receptive field to 44kb, adding more vanilla convolutional layers, and introducing dilated convolutional layers that allow the model to consider exponentially larger areas in the input sequence and capture a greater diversity of interactions between different regulatory elements. The model is trained in a multitask fashion, predicting other epigenetic features and markers of regulatory activity in addition to transcript abundance as measured by CAGE. Basenji2 is able to significantly outperform Xpresso, achieving a Pearson correlation of 0.85 between log prediction and log experiment across all tested genes. Basenji2 performance does suffer for more variable genes, however. Although Basenji2 does learn that these genes display a higher diversity

in expression levels, it does not predict the full range of variability. However, the benefit of the wider receptive field of Basenji2 is reflected in the saliency scores. Basenji2 picks up on the effects of the promoter sequence, similar to Xpresso, but also identifies enhancer and silencing motifs. This implies that Basenji2 could be useful for fine-mapping eQTLs, and it was found that Basenji2 achieves statistically significant correlation in predicting the direction of effect of regulatory variants in the Gene Tissue Expression (GTEx) data [14]. However, the convolutional architecture of Basenji2 still limits its ability to learn relationships between distal elements.

Enformer. In order to overcome this burden of locality while still taking advantage of the propensity of CNNs to recognize sequence motifs, Enformer [15] replaces the dilated convolutions of Basenji2 with a self-attention mechanism inspired by work in natural language processing (NLP). Self-attention [16] has been utilized in Transformer architectures to facilitate the learning of long-range dependencies within sequence in a computationally efficient and simple manner. These architectural changes lead to improvements in cell-type specific expression predictions over Basenji2 [15]. In addition, the Enformer model takes in an even larger input sequence, 196kb, and is able to more reliably predict expression for genes with high and low variance in the CAGE datasets. Further study of the attention mechanisms revealed that the model developed a much better understanding of tissue-specific regulatory elements, which led to improved predictions of direction of effect for GTEx eQTLs.

ExPecto. The final method evaluated implements a hierarchical model, utilizing a convolutional neural network to predict various chromatin features based on DeepSEA [17], and using the results to build tissue-specific linear models to predict expression. This architecture, known as ExPecto [18], is trained on Roadmap, GTEx, and ENCODE [19] expression and regulatory feature data. Since epigenetic marks play a central role in regulating gene expression [20], predicting epigenetic features such as transcription factor (TF) binding and histone modifications as an intermediate task helps ExPecto learn differences in expression across tissue types [18]. This in turn helps ExPecto predict direction of effect when evaluated on significant GTEx eQTL variants. ExPecto was used to interpret GWAS data and prioritize putative causal variants; for example, the method identified variants that were pinpointed to be causal across several GWAS, and even prioritized variants that were later confirmed to be causal in follow-up analyses.

1.3 CAGE and RNA-Seq

Although all of the above methods utilize gene expression data in training, they use datasets with gene expression measurements from two different methods. In RNA-seq experiments, RNA is first isolated from a biological sample and converted into complementary DNA (cDNA). The sequences then go through an enrichment procedure where specific types of RNA (e.g. messenger RNA, microRNA, etc.) are selected by depleting non-desired forms of RNA [21]. The reads are then sequenced, aligned, and quantified to determine gene expression at the granularity of a gene. CAGE experiments, on the other hand, identify

and quantify the 5'-ends of capped RNA through cDNA cap trappers [22]. Expression is measured as a continuous output across the genome at single nucleotide resolution [23]. Due to the gene-level resolution of RNA-seq, it is difficult for models trained on RNA-seq data to be evaluated on CAGE data, but the opposite task is reasonable. Around the TSS, the results of both methods are highly correlated with each other, although CAGE is less stable and reproducible when considering expression at a specific single-nucleotide TSS [24]. In order to combat this issue, it is common practice to consider a window around the TSS when quantifying expression predicted by a computational model trained on CAGE.

1.4 Personalized expression prediction

Modern deep learning architectures such as those described above have helped make large strides in understanding patterns of gene expression in both cell-type specific and agnostic settings. However, the evaluation of these models has primarily been performed in the "reference" sequence setting; namely, by evaluating the models' ability to explain variation in expression levels across different genes in the transcriptome, using the reference sequence around each gene TSS as input. To our knowledge, the ability of these models to explain variation in expression of a given gene across individuals, incorporating personal genome variation around each promoter sequence, has not yet been evaluated. In this work, we assess the four above models when applied to personalized genomes for personalized expression prediction and aim to characterize the strengths and shortcomings of these approaches.

Chapter 2

Methods

2.1 Dataset

The data used for gene expression prediction was gathered from the Geuvadis consortium [25], which includes paired gene expression and whole genome sequencing data from individuals in the 1000 Genomes Project. The E-GEUV-1 release includes mRNA sequencing data from lymphoblastoid cell lines (LCLs) from a total of 465 samples. After excluding samples with unphased imputed genotypes, there were 421 Geuvadis individuals with phased whole genome sequencing data. These samples originated from five different populations with ancestry in Europe and Africa continents.

2.2 Creation of personalized input sequences

To prepare the samples for gene expression inference, we extracted the personal genome sequence around each gene for each individual in the dataset. In particular, we constructed the sequence of each haplotype for each individual centered around the TSS of each gene using the bcftools consensus command and the personal genome variation data in the Geuvadis VCF files. Information regarding the ENSEMBL Gene ID, TSS position, strand, and chromosome was also obtained from Geuvadis. The gene symbol was acquired using a converter from BioTools [26]. These tables were combined to create a csv file with all necessary metadata.

The different gene expression prediction methods each have different size receptive fields and thus required separate personalized input sequences. Xpresso uses an asymmetric input sequence of 7Kb upstream of the TSS and 3.5Kb downstream of the TSS; thus, the required input sequence for Xpresso depends on the orientation of the gene. For genes located on the positive strand, we directly computed the personal sequences. For genes located on the negative strand, we extracted the reference sequence with swapped boundaries of 3.5kb upstream and 7kb downstream, applied bcftools consensus, then took the reverse complement. The input sequences for Expecto, Basenji2, and Enformer are all symmetric about the TSS.

Expecto and Basenji2 both have a receptive field of approximately 40kb, while Enformer has a much larger receptive field of 196kb. All personalized sequences were computed with a combination of samtools and bcftools.

For our current analysis, we considered only single nucleotide variants (SNVs) and filtered out indels when creating the personalized input sequences, since SNVs do not change the length of the input sequence around each gene. Since the Geuvadis VCFs are based on hg19, we used hg19 as the reference genome for creating personal sequences. After creating these sequences, we performed a verification step by comparing the number of variants expected from the VCF file to the edit distance between the reference and personal sequences. Except where otherwise noted, results are shown across all genes that contained a significant eQTL in the Geuvadis EUR eQTL analysis.

2.3 Gene expression inference

We predicted gene expression levels for each individual using four state-of-the-art methods, and averaged the predictions from each individual's two haplotypes.

Xpresso. We used two pre-trained Xpresso models—human median expression and lymphoblastoid cell expression—to perform inference with a per-gene fasta as input. Input fasta sequences for all individuals were combined to create one fasta file for each gene, containing all of the individual data. The resulting mRNA expression predictions from Xpresso were stored in a two column plain text format containing the name of the sample according to the fasta and the predicted log RPKM expression value.

ExPecto. ExPecto predicts tissue-specific expression from sequence alone using deep-learning-based predictions of transcription factor binding, DNA accessibility, and histone marks in various cell types. The personalized sequences for each individual were fed into a pre-trained model (Beluga) that generates predictions of all of these epigenomic features for the 40-kb region surrounding the TSS, followed by a spatial transformation module as described in Zhou et al [18]. The resulting spatially transformed features are then used as input features for an L2-regularized linear regression model in order to predict expression for a given gene. To obtain expression predictions in a matching cell line to the Geuvadis experiments, we used the publicly available ExPecto model trained on EBV-transformed lymphocytes.

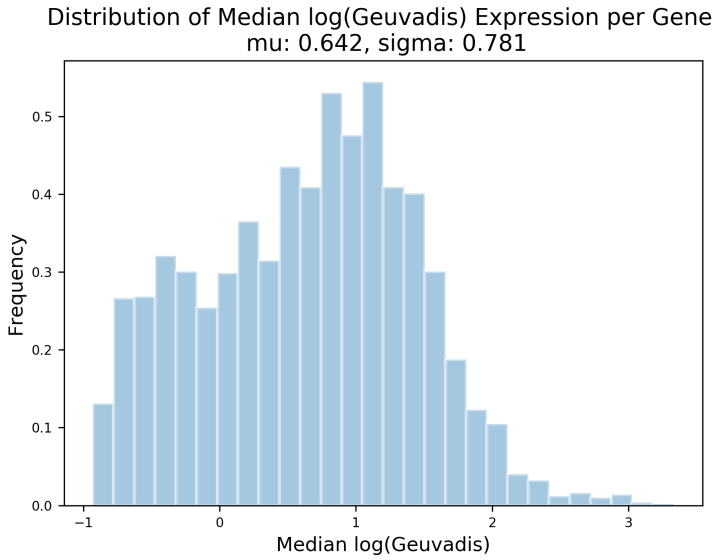
Basenji2. Basenji2 predicts expression for 5,313 epigenetic and transcriptional profiles taken from the ENCODE and FANTOM consortiums in 128-bp bins across the genome. To obtain expression predictions in a matching cell line to the Geuvadis RNA-seq experiments, we used Basenji2 predictions for CAGE measurements performed on the GM12878 lymphoblastoid cell line. The prediction of expression for a given gene was computed by averaging the predicted CAGE signals in the bin containing the TSS, the 5 bins upstream of the TSS, and the 5 bins downstream of the TSS. Since the region closest to the TSS has the greatest impact on gene expression, we use only these bins closest to the TSS of the gene to aggregate one value for the predicted expression of the gene of interest.

Enformer. Enformer utilizes 196kb of personalized input sequence and predicts expression for 5,313 epigenetic tracks in 128 bp bins across 114,688 bp, so the predictions are of length 896 for each haplotype. While the authors of Enformer averaged predictions within a 3-bin window around each gene TSS, we found that the 5-bin range (as also used above for Basenji2) had better performance on the Geuvadis data.

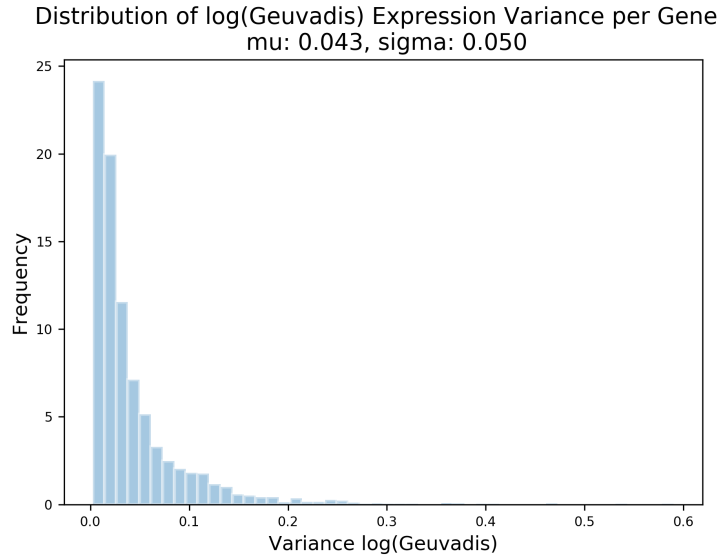
Chapter 3

Results

We started by visualizing the RNA-seq data provided in Geuvadis. We look at the distributions of median expression (Fig 3.1a) and variance of expression (Fig 3.1b), noticing the heavy right skew of the data. These patterns are to similar to those of the experimental expression data in GTEx [9] and serves as a baseline check for our expectations of these models on Geuvadis data. The measurements made in RNA-Seq are often noisy, so we could reasonably expect a small drop in performance.



(a)



(b)

Figure 3.1: Visualized distributions of median (a) gene expression and variance (b) of gene expression for all genes with a significant eQTL in Geuvadis. The expression values are all transformed by $\log_{10}(RPKM + 0.1)$.

3.1 Computational models tend to underperform on individuals

In order to characterize the general performance of these methods after inference, we calculated the correlation between the Geuvadis and predicted expression.

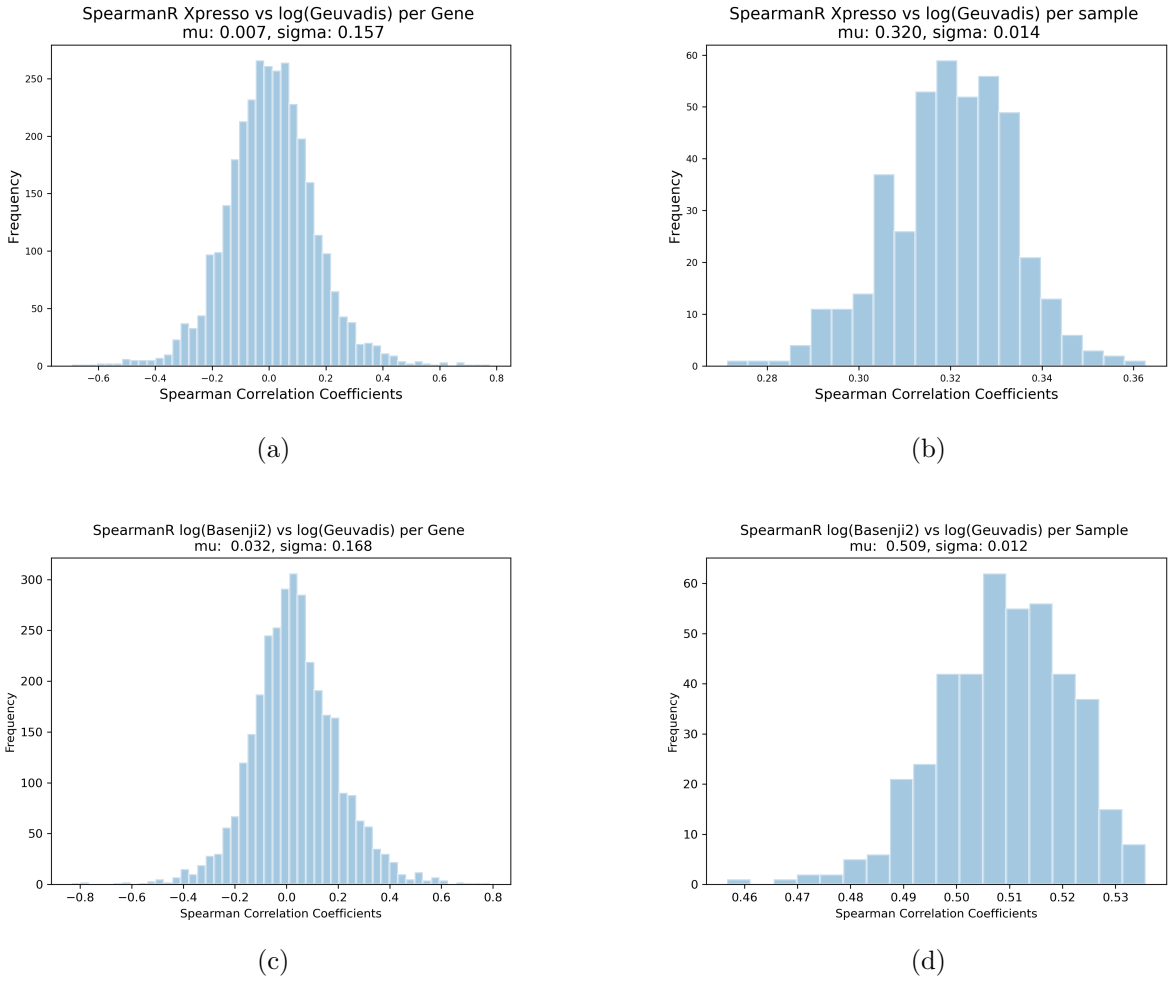
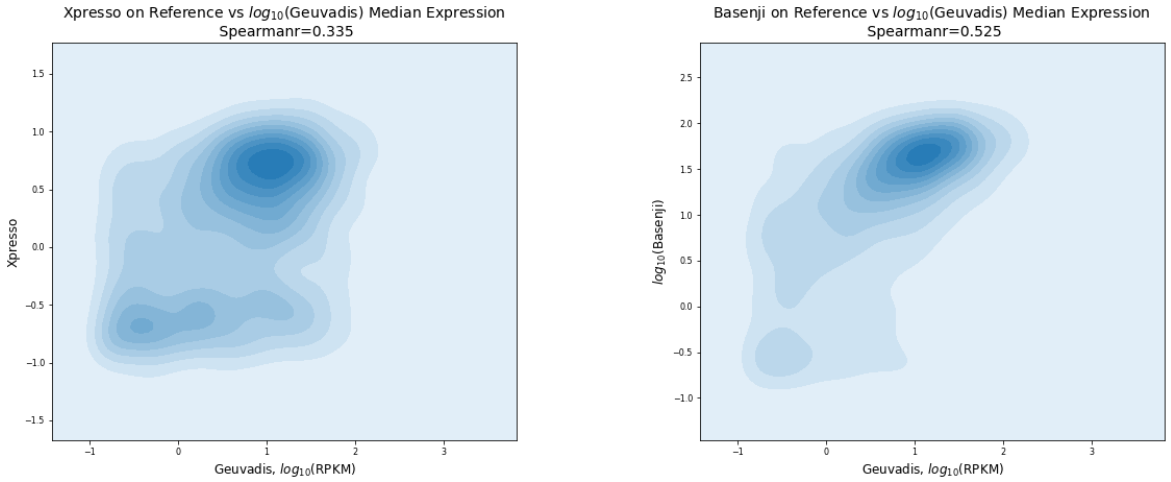


Figure 3.2: Distribution of Spearman correlations between predicted expression and Geuvadis expression. The correlations are calculated for Xpresso (a, b) and Basenji (c, d). The left distributions show the correlations computed for predictions across each gene (so there are 3000+ correlations), and the right distributions show the correlations computed for predictions across each individual (so there are 463 correlations). The data for all genes for Expecto and Enformer is not currently available.

The distribution of Spearman correlations for Xpresso and Basenji when computed across genes (Fig 3.2a and 3.2c) is approximately normal and centered at 0. This suggests that these methods tend not to be capturing the relationships between variants for many of these



(a) (b)

Figure 3.3: Xpresso (a) and Basenji (b) predictions on reference sequence for all genes versus the median $\log_{10}$ Geuvadis expression. The data is currently unavailable for Expecto and Enformer.

genes. Since, Xpresso and Basenji are trained on reference sequences for several different genes, this likely indicates that with this training procedure, the models are able to learn a representation for general important regulatory motifs, but fail to understand the nuances of single nucleotide variants in these motifs. We also see the models misconstrue associations since a large chunk of genes have negative correlations. We do, however, note that when the correlations are computed across each sample, the distribution for Xpresso (Fig 3.2b) is centered at a 0.320 Spearman r versus Basenji (Fig 3.2d) at 0.509 Spearman r. These results testify that Xpresso and Basenji have shown a propensity to learn the general expressivity of each gene. Although we do expect the correlations to be lower than what was reported in the respective articles, the drop in correlation from predictions (Fig 3.3) on reference sequence to predictions on individuals entails that these sequence-to-expression models underperform when tested on individuals.

To investigate this phenomenon a bit further, we specifically focused on genes that were measured to have high variance in expression across individuals (Fig 3.4). The models exhibit a relatively high variance in the predictions, although there is a systematic reduction in absolute range of the expression when compared with that of the experiment. Basenji2 (Fig 3.4c) more heavily compresses the variance even further, consistent with the findings when it was trained that the model consistently underestimates the degree of variance [12]. RNA-seq experiments often suffer from biological and technical noise generated from transcription, enrichment procedures, and other factors associated with sequencing [27, 28]. The deeper network with several convolutional and pooling layers likely acts as a denoising mechanism that can limit the dynamicity in prediction variance. We additionally analyzed the edit distance of each of the predictions to gather insight into how these models react to the

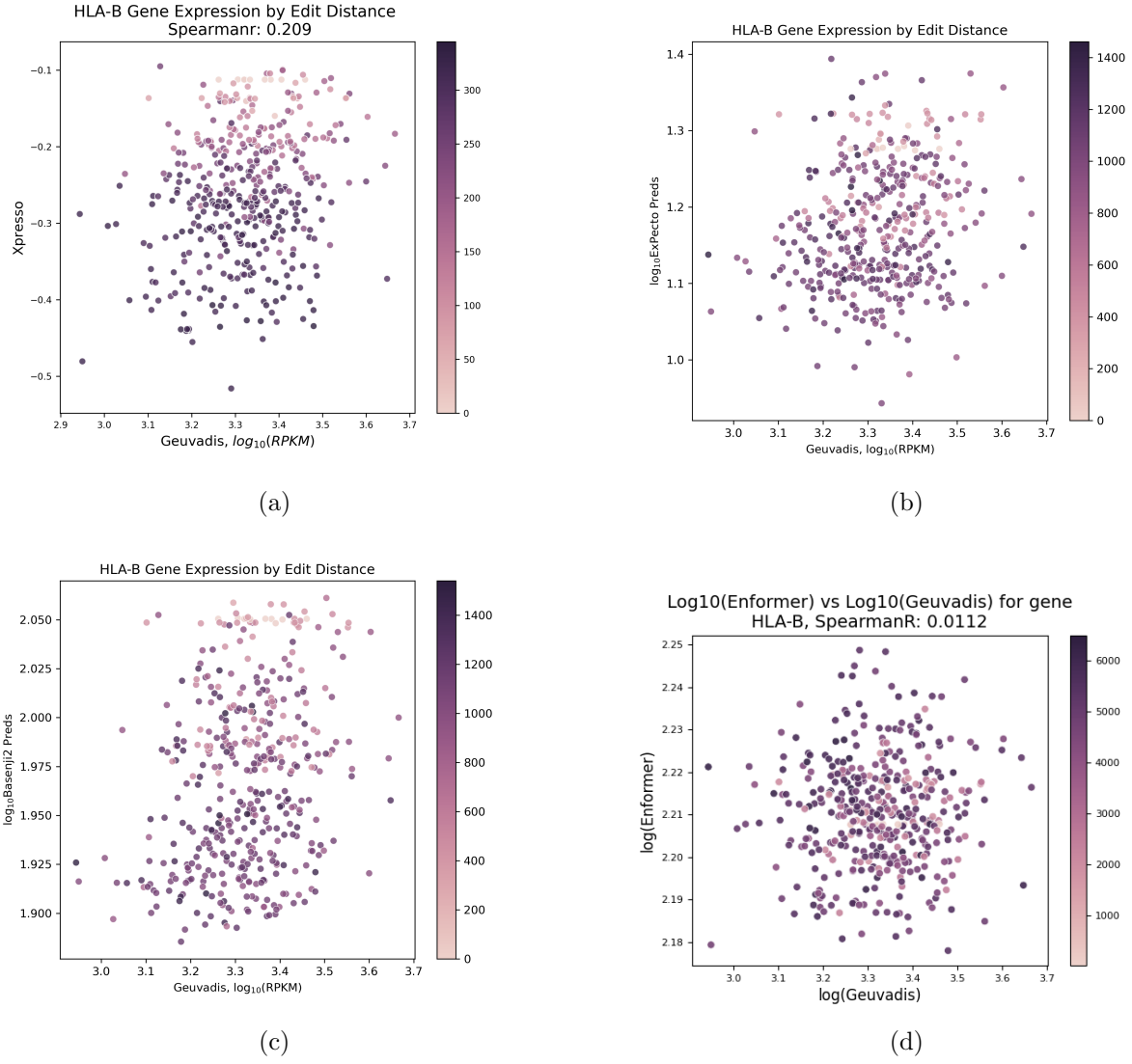


Figure 3.4: Shows predicted expression versus Geuvadis expression for the high variance gene *HLA-B* colored by the edit distance for an individual from the reference sequence. The resulting expressions are shown for Xpresso (a), Expecto (b), Basenji (c), and Enformer (d). Spearman r is only shown for Xpresso and Enformer.

consensus sequences. These models, especially Xpresso (Fig 3.4a), tend to understand that a higher volume of variants in the sequence lead to more varied expression.

We further inspected lower variance genes to see if the trend still holds. In accordance with our expectations, the variance in predictions has been significantly restricted (Fig 3.5). Low abundance-low variance scenarios such as *CPAMD8* seem to be much more difficult for the models to understand as the relationship between number of variants and expression starts to break down. Interestingly, in both the high variance and low variance scenarios,

Xpresso predictions (Fig 3.4a and 3.5a) have the wrong sign, which could be a result of the smaller receptive field.

3.2 Methods can learn signficant eQTLs but struggle with direction of effect

While the methods had poor correlation with the high and low variance genes, we were next interested in understanding how these methods performed on genes with the most significant eQTLs as identified by Geuvadis. We subsetted to find genes with significant eQTLs within the Xpresso's receptive field since it is the smallest, and focused on the top 10 most significant eQTLs. We cannot reasonably expect Xpresso to perform well in predictions when the significant determinants of expression lie outside the input sequence. Two of the chosen genes, *SNHG5* (Fig 3.6) and *PEX6* (Fig 3.7), are shown here. Within both genes, we see a clear three clusters of predictions forming with a high magnitude in correlation. While the clustering could initially seem to be related to the number of variants, analyzing the predictions as a function of dosage of the putatively causal SNP (Fig 3.8 and 3.9) reveals that the methods are capable of identifying very significant loci contributing to expression.

Crucially, while the models may learn to recognize significant variants, the models may not always understand the correct direction of effect. In *SNHG5*, Expecto (Fig 3.8b), Basenji (Fig 3.8c), and Enformer (Fig 3.8d) all predict a negative direction of effect while *rs1059307* variant actually increases expression. Xpresso (Fig 3.8a) is the only method that manages to capture this relationship, which could be due to a combination of the close proximity of the SNP to the TSS and the smaller receptive field. Up-regulation of *SNHG5* has been shown to be associated with elevated tumor growth [29] so the methods with longer receptive fields could be considering silencing regulatory features when predicting expression. The models are all shown to agree for *PEX6*, but further experiments would need to be conducted as to the difference in importance of features between the two genes.

Additionally, we consider the difference in expression between the reference sequence and the reference sequence with the alternate allele at the most significant SNP (alternate sequence), and are shown in Figures 3.8 and 3.9. For all methods, the difference between reference and alternate do not span the entire range of variance of the predictions. Given that the predictions have more than one variant and there are explicit clusters of predictions, this likely suggests that these significant variants are in LD. Groups of variants within these receptive fields are likely inherited together or are more likely to co-occur, leading to the additional scope of predicted abundance.

The discordant opinions of the models on direction of effect prompted a more extensive study of the top eQTLs. We computed SNP expression difference (SED) scores, which is defined as the difference in expression between the earlier computed reference and alternate sequences. When plotted against the experimentally determined directions of effect, we see gradual improvements with respect to the complexity of the models (Fig **??**). As these

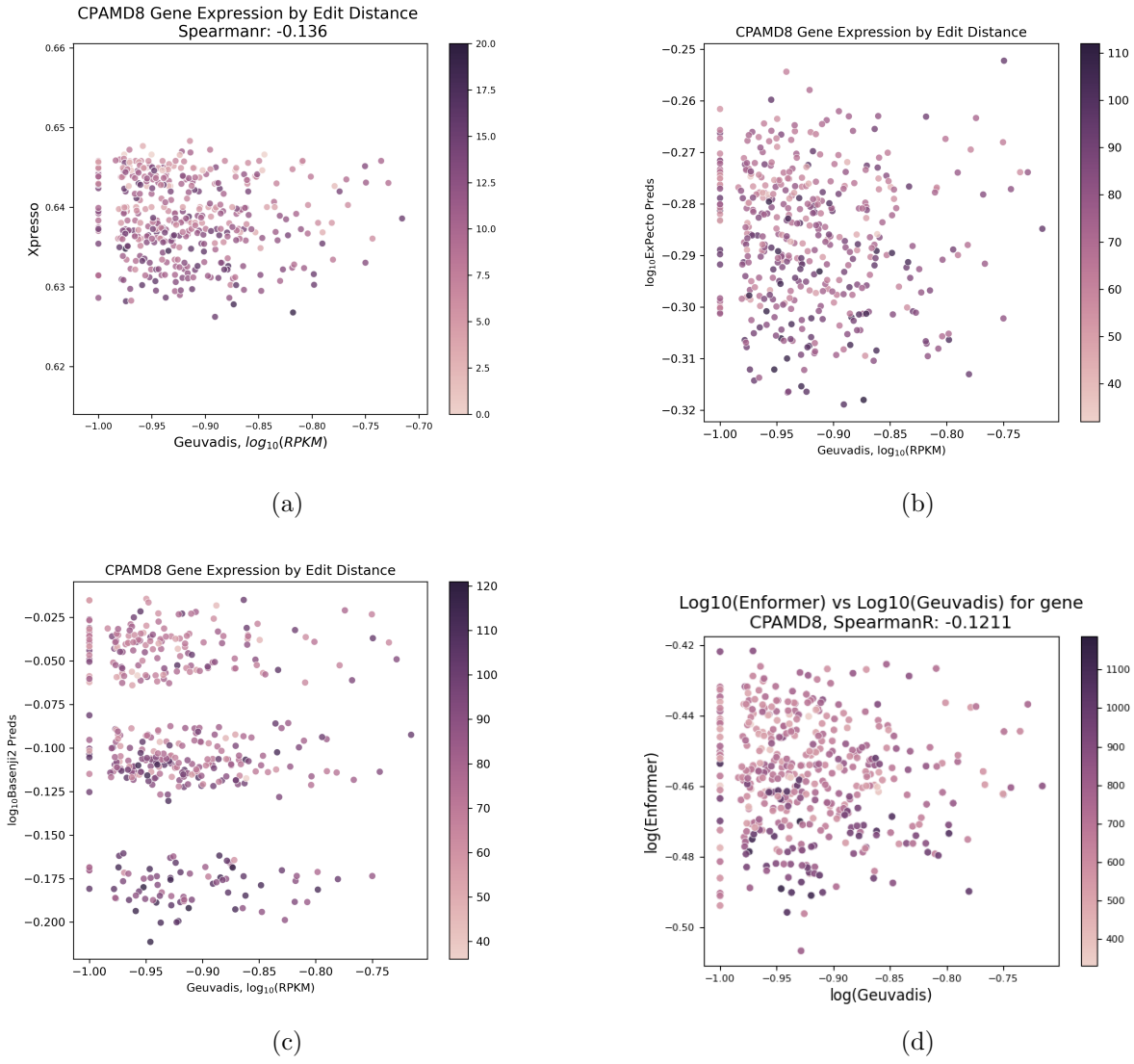


Figure 3.5: Shows predicted expression versus Geuvadis expression for the low variance gene *CPAMD8* colored by the edit distance for an individual from the reference sequence. The resulting expressions are shown for Xpresso (a), Expecto (b), Basenji (c), and Enformer (d). The Geuvadis expression is log transformed for all methods, and Spearman r is only shown for Xpresso and Enformer.

models are able to capture and incorporate more distal information, they are more capable in ascertaining the appropriate direction of effect. The multi-task nature of Basenji also entails that developing a common representation for a more diverse set of regulatory features proves more helpful when considering direction of effect. Aggregating and integrating information from all the changes in regulatory features offers a much more complete picture of the mechanisms involved in affecting expression.

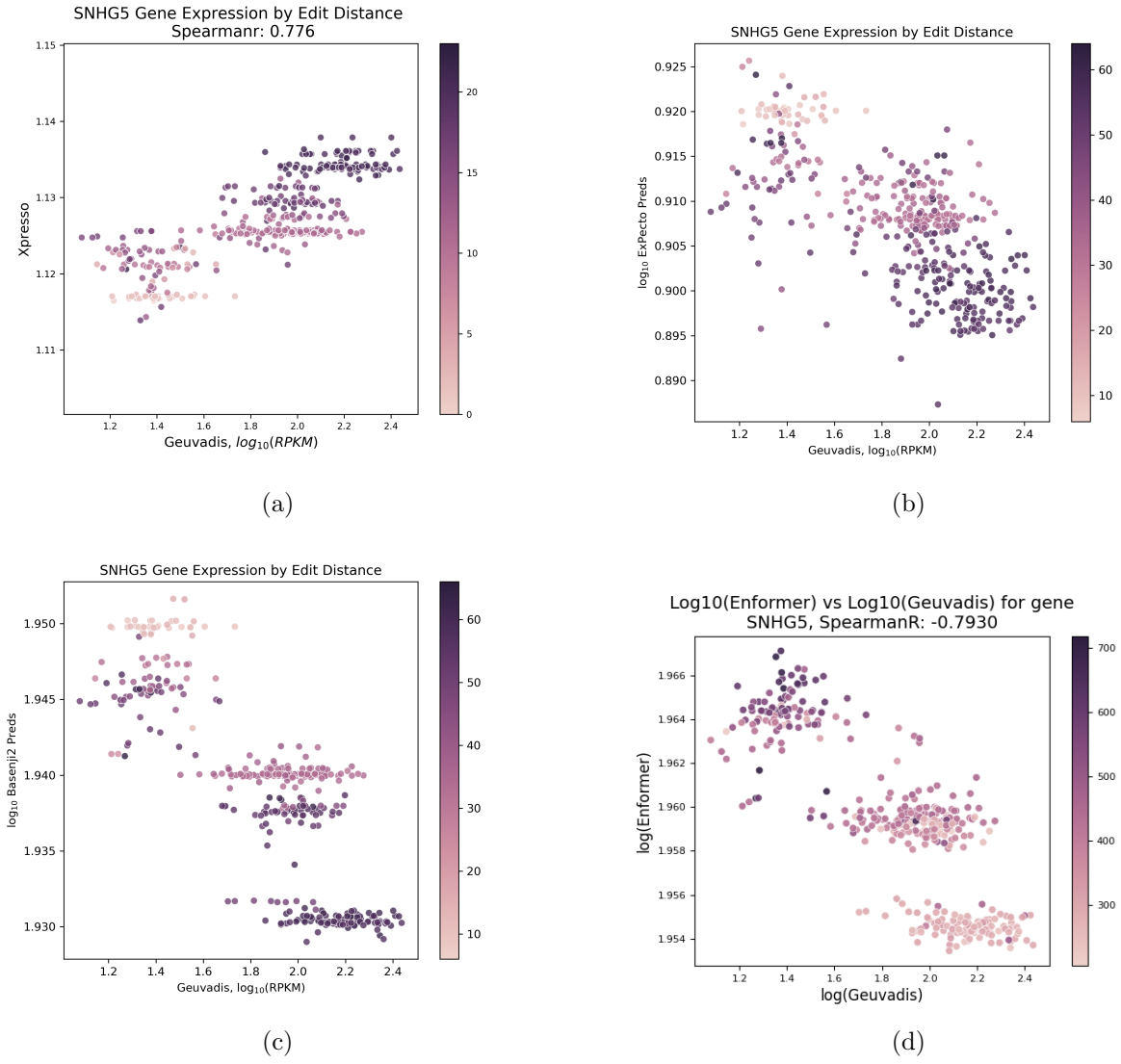


Figure 3.6: Shows predicted expression versus Geuvadis expression for the gene *SNHG5* by the edit distance for an individual from the reference sequence. The resulting expressions are shown for Xpresso (a), Expecto (b), Basenji (c), Enformer (d). Spearman r is only shown for Xpresso and Enformer.

Finally, we measured the agreement of these models on predictions within the set of 582 eQTLs we studied. The average correlation between these four methods (Fig 3.11a) seems suggest that Xpresso predictions are relatively random in comparison. However, when observing the correlations on individual genes, we see that Xpresso tends to strongly agree or disagree with the other methods. It is reasonable to expect Basenji and Enformer to have relatively similar prediction patterns given that Enformer inherits from Basenji [15], however we see that in the case of *KLHL7-DT* 3.11d, this is not always true. The transformer layers

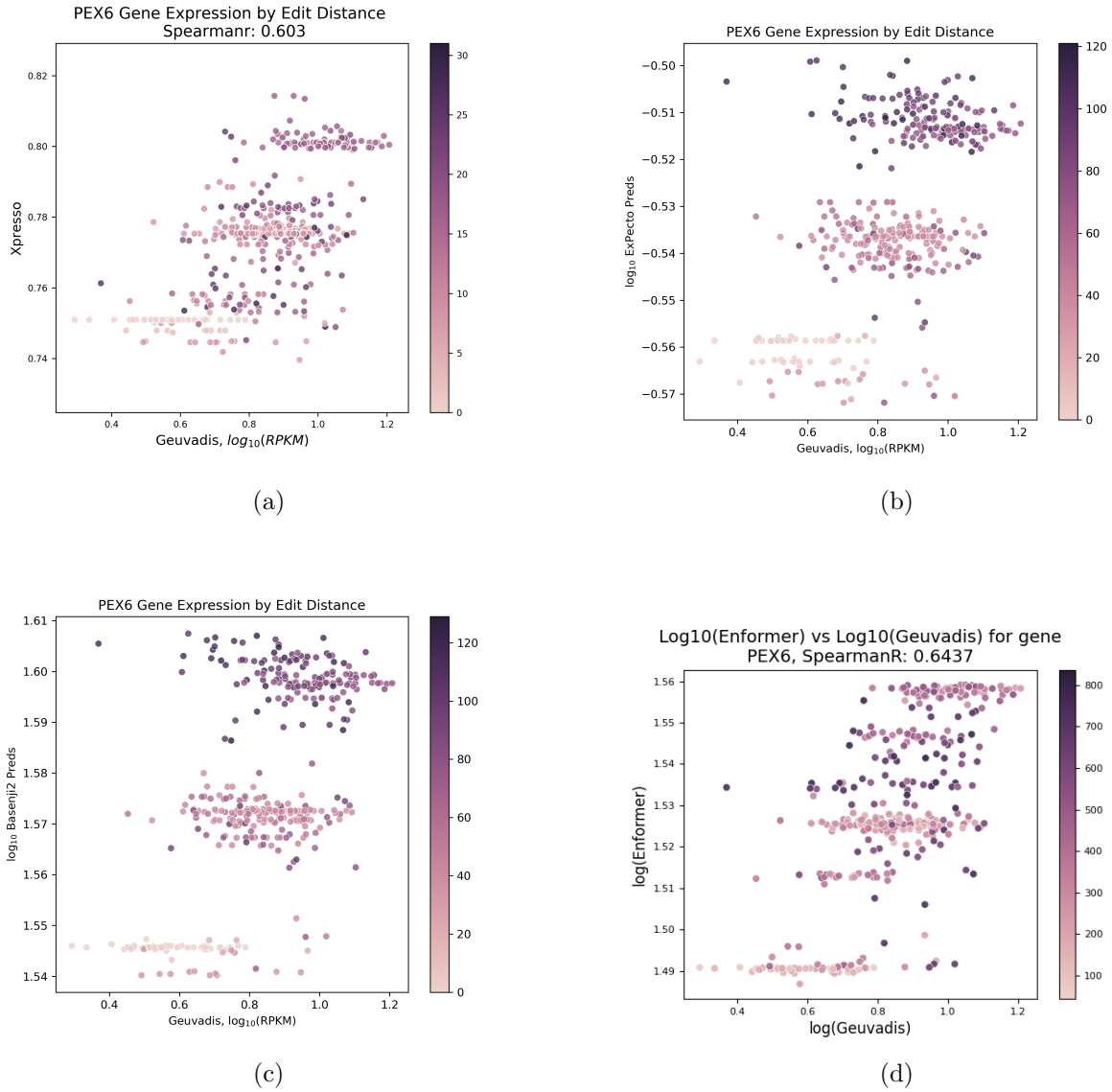


Figure 3.7: Shows predicted expression versus Geuvadis expression for the gene *PEX6* colored by the edit distance for an individual from the reference sequence. The resulting expressions are shown for Xpresso (a), Expecto (b), Basenji (c), Enformer (d). Spearman r is only shown for Xpresso and Enformer.

may allow Enformer to more effectively attend to signals directly around the TSS.

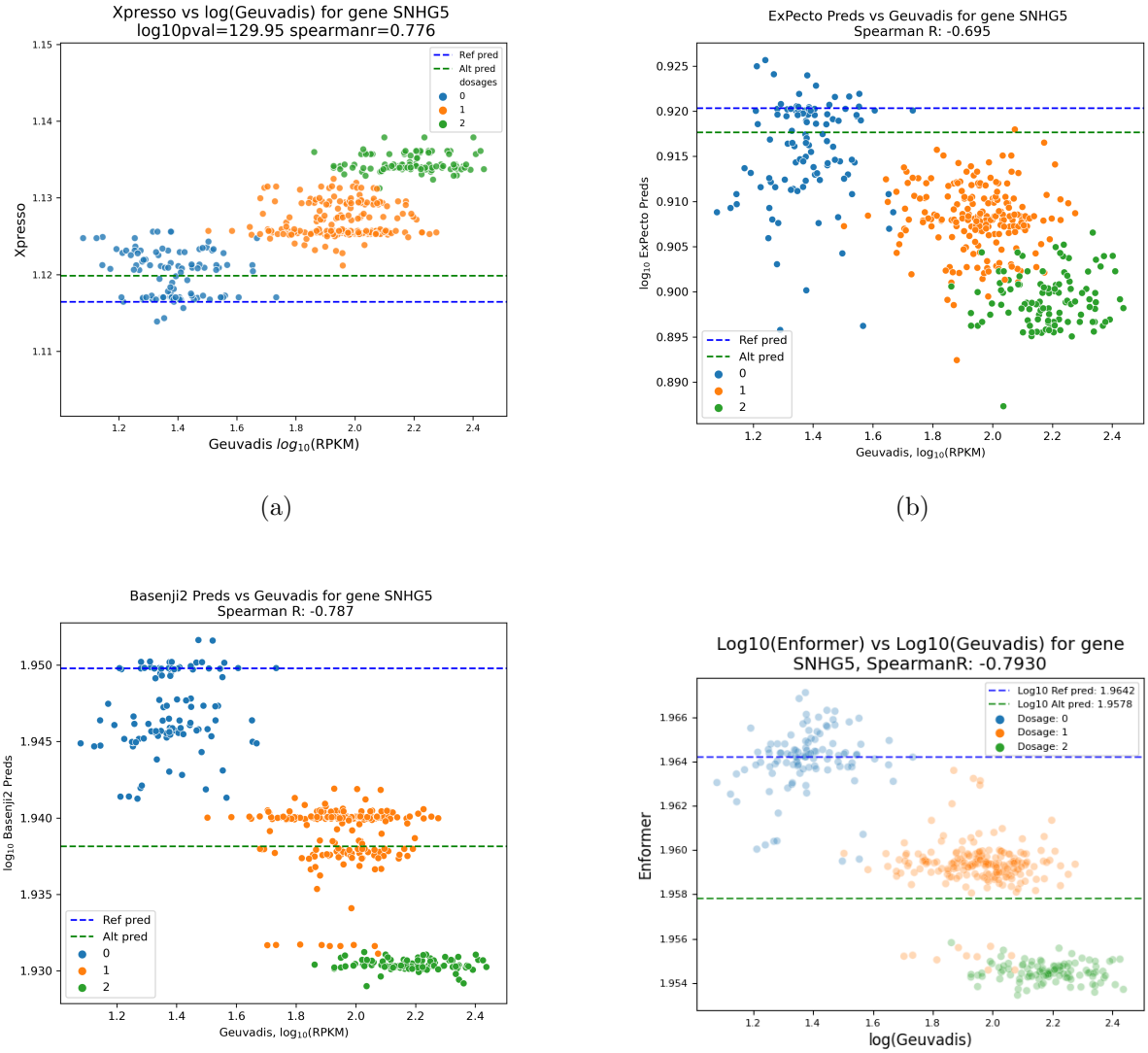


Figure 3.8: Shows predicted expression versus Geuvadis expression for the gene *SNHG5* colored by the dosage for the most significant SNP. The resulting expressions are shown for Xpresso (a), Expecto (b), Basenji (c), Enformer (d). Spearman r is only shown for Xpresso and Enformer.

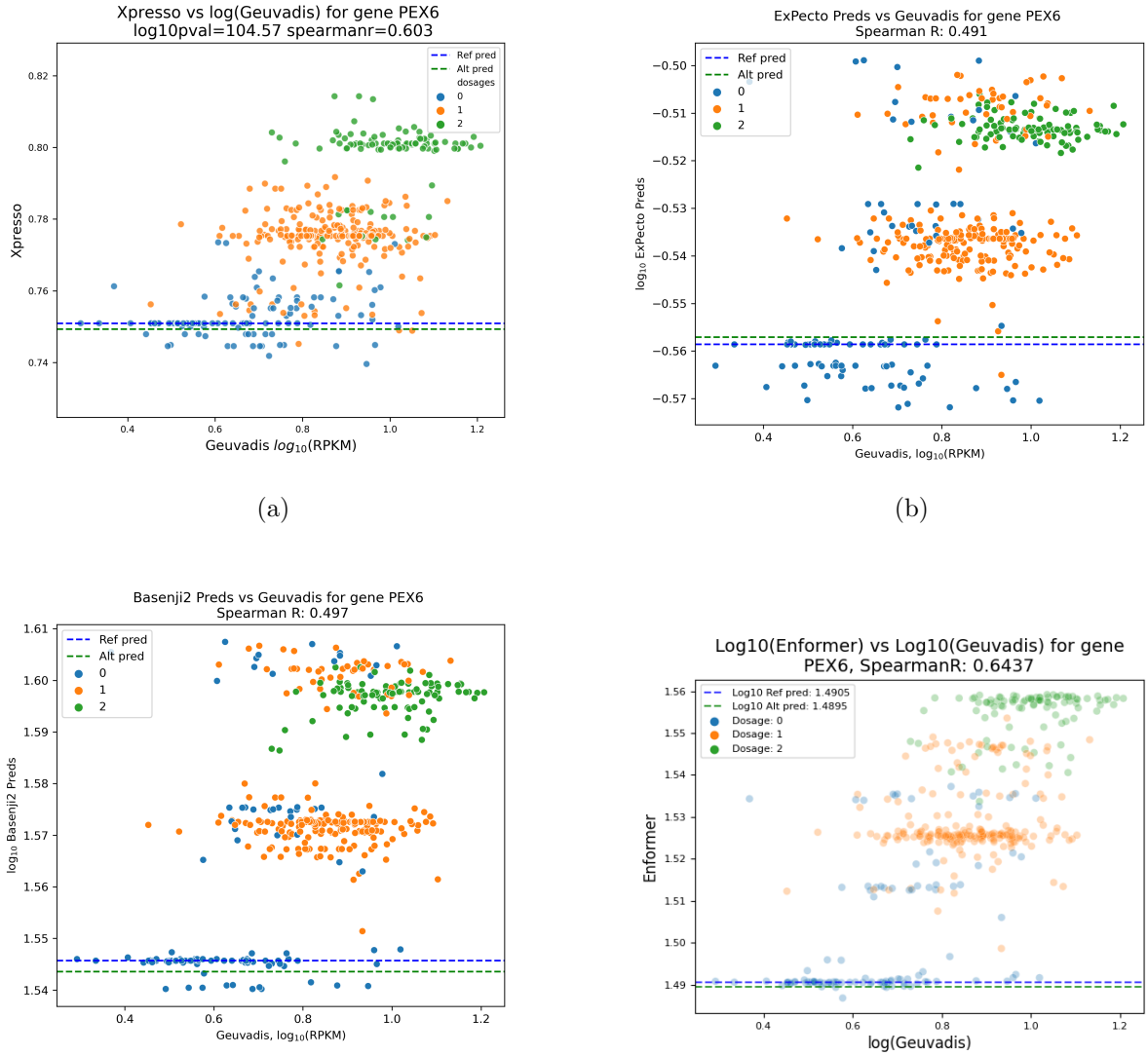
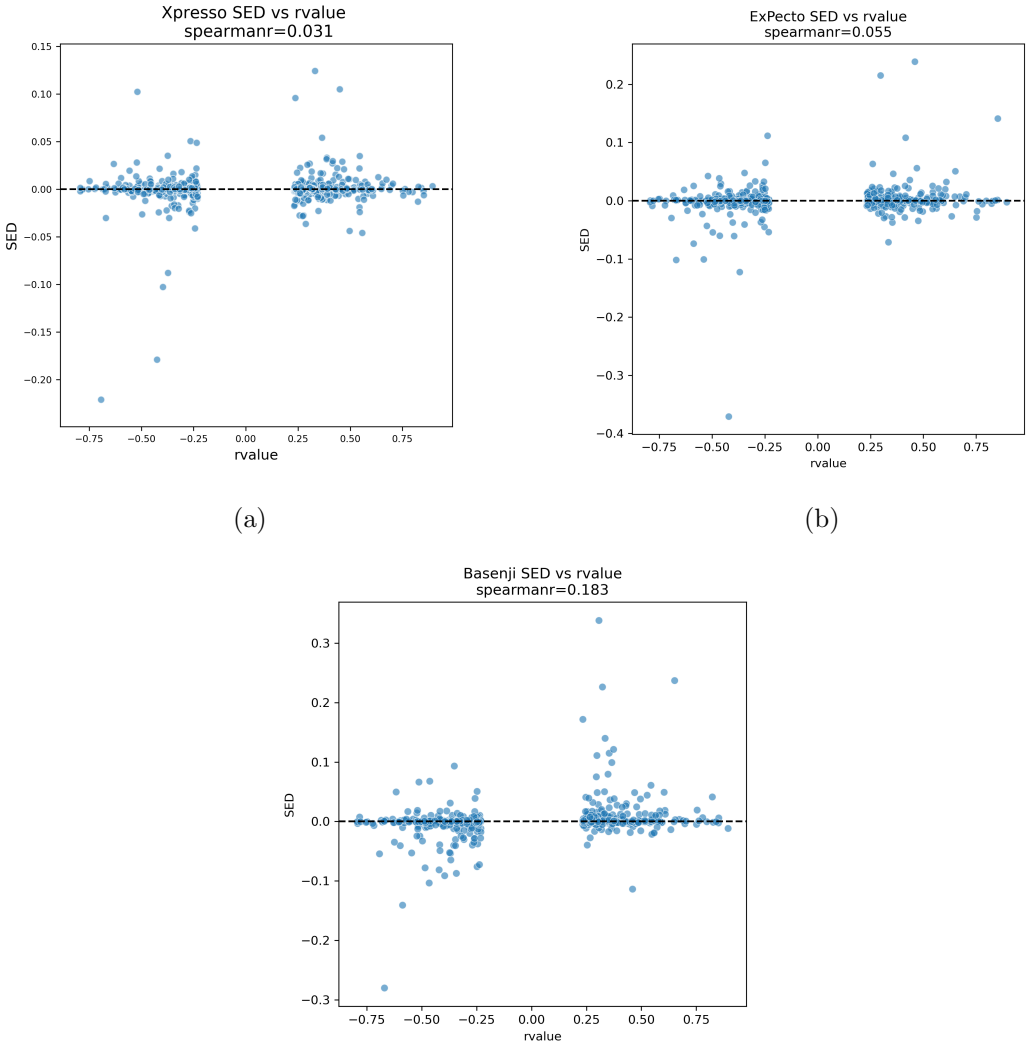


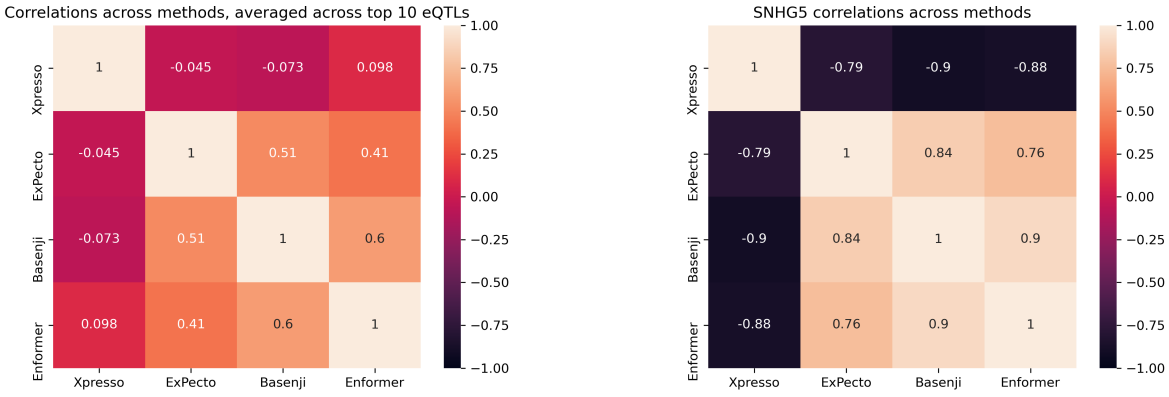
Figure 3.9: Shows predicted expression versus Geuvadis expression for the gene *PEX6* colored by the dosage for the most significant SNP. The resulting expressions are shown for Xpresso (a), Expecto (b), Basenji (c), Enformer (d). Spearman r is only shown for Xpresso and Enformer.



(a) (b)

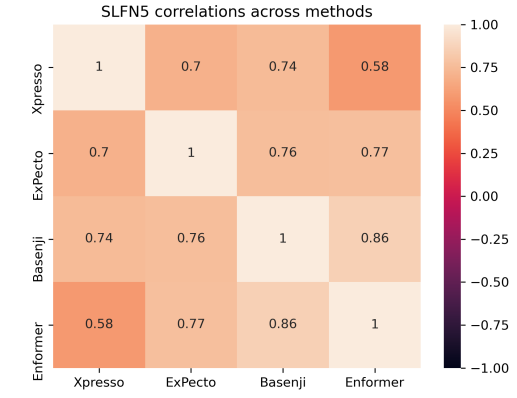
(c)

Figure 3.10: SED scores for Xpresso (a), Expecto (b), and Basenji (c) plotted against direction of effect of all eQTLs within the receptive field of Xpresso. The black dashed line represents a SED score of 0, which means a variant has no effect on the expression from reference. The data all of these genes is not yet available for Enformer.

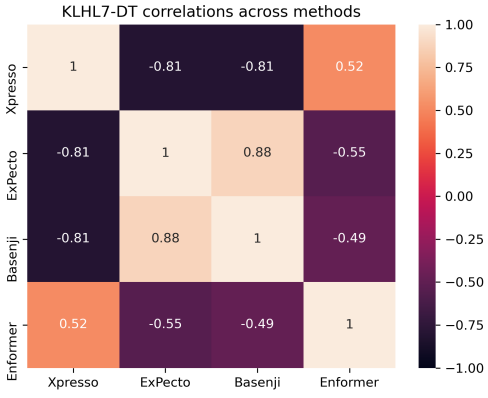


(a)

(b)



(c)



(d)

Figure 3.11: Heatmap of Spearman correlations across the predicted expressions of the four methods. The correlations are averaged (a) across all 10 genes containing the most significant eQTLs. The correlations for 3 of those individual genes are also shown: SNHG5 (b), SLFN5 (c), and KLHL7-DT (d).

Chapter 4

Discussion

4.1 Conclusions

Recent advances in deep learning have led to improved models for predicting gene expression from genetic sequences. These large networks are able to learn increasingly complex functions; in the biological setting, this has translated to an improved capacity to understand, characterize, and uncover complex regulatory interactions within long input sequences. However, the clinical applications of these sequence-to-expression models are largely unexplored.

We establish a framework and analyze the performance of four state-of-the-art sequence-to-expression architectures—Xpresso, ExPecto, Basenji, and Enformer—on personalized expression prediction. These models, while successful in learning sequence features that explain variation in expression across different genes in the genome, consistently underperform when predicting differences in expression across individuals based on a smaller set of inter-individual differences in the input DNA sequence. Although these models possess the ability to recognize some regulatory variants, they do not reliably predict the effects of those variants on expression. While current state-of-the-art models may still be unsuited to personal genome interpretation, we propose some strategies for improving these models below.

4.2 Future Work

We suggest several further experiments to confirm our findings and offer a few avenues of exploration to consider for improving performance on personal genomes. We first plan to repeat these experiments on additional datasets, such as GTEx [14], to confirm the results of this study. In addition, a current shortcoming of the analyses above is the lack of consideration of insertions and deletions within the personal genome sequences. Indels are an important contributor to variance in human gene expression [30] that could partially explain differences between predicted and experimentally measured expression levels. The addition of mRNA half-life features could also improve the accuracy of these predictions; for example,

new methods that aim to predict half-life from sequence could be incorporated into these models [31].

Finally, we plan to implement new training and fine-tuning procedures for these models to determine whether it is possible to improve their sensitivity to small changes in input sequences. In particular, instead of training solely on reference sequence, we aim to include paired personal genome and expression data in the training of these models and analyze changes in the important features learned by the models.

Bibliography

[1] Vivian Tam et al. “Benefits and limitations of genome-wide association studies”. In: *Nature Reviews Genetics* 20.8 (2019), pp. 467–484.

[2] Feng Zhang and James R Lupski. “Non-coding genetic variants in human disease”. In: *Human molecular genetics* 24.R1 (2015), R102–R110.

[3] Khader Shameer et al. “Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment”. In: *Briefings in bioinformatics* 17.5 (2016), pp. 841–862.

[4] Montgomery Slatkin. “Linkage disequilibrium—understanding the evolutionary past and mapping the medical future”. In: *Nature Reviews Genetics* 9.6 (2008), pp. 477–485.

[5] Thanasis Mitsis et al. “Transcription factors and evolution: an integral part of gene expression”. In: *World Academy of Sciences Journal* 2.1 (2020), pp. 3–8.

[6] Warren A Cheung et al. “Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome”. In: *Genome biology* 18.1 (2017), pp. 1–21.

[7] Alvaro Sanchez et al. “Effect of promoter architecture on the cell-to-cell variability in gene expression”. In: *PLoS computational biology* 7.3 (2011), e1001100.

[8] Ann-Bin Shyu, Miles F Wilkinson, and Ambro Van Hoof. “Messenger RNA regulation: to translate or to degrade”. In: *The EMBO journal* 27.3 (2008), pp. 471–481.

[9] Vikram Agarwal and Jay Shendure. “Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks”. In: *Cell reports* 31.7 (2020), p. 107663.

[10] Bradley E Bernstein et al. “The NIH roadmap epigenomics mapping consortium”. In: *Nature biotechnology* 28.10 (2010), pp. 1045–1048.

[11] Iris Zhu et al. “A model of active transcription hubs that unifies the roles of active promoters and enhancers”. In: *Nucleic acids research* 49.8 (2021), pp. 4493–4505.

[12] David R Kelley et al. “Sequential regulatory activity prediction across chromosomes with convolutional neural networks”. In: *Genome research* 28.5 (2018), pp. 739–750.

[13] David R Kelley. "Cross-species regulatory sequence activity prediction". In: *PLoS computational biology* 16.7 (2020), e1008050.

[14] John Lonsdale et al. "The genotype-tissue expression (GTEx) project". In: *Nature genetics* 45.6 (2013), pp. 580–585.

[15] Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature methods* 18.10 (2021), pp. 1196–1203.

[16] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[17] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nature methods* 12.10 (2015), pp. 931–934.

[18] Jian Zhou et al. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". In: *Nature genetics* 50.8 (2018), pp. 1171–1179.

[19] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), p. 57.

[20] ER Gibney and CM Nolan. "Epigenetics and gene expression". In: *Heredity* 105.1 (2010), pp. 4–13.

[21] Kimberly R Kukurba and Stephen B Montgomery. "RNA sequencing and analysis". In: *Cold Spring Harbor Protocols* 2015.11 (2015), pdb–top084970.

[22] Piero Carninci et al. "High-efficiency full-length cDNA cloning by biotinylated CAP trapper". In: *Genomics* 37.3 (1996), pp. 327–336.

[23] Sebastian Boegel. *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*. Springer, 2020.

[24] Hideya Kawaji et al. "Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing". In: *Genome research* 24.4 (2014), pp. 708–717.

[25] Tuuli Lappalainen et al. "Transcriptome and genome sequencing uncovers functional variation in humans". In: *Nature* 501.7468 (2013), pp. 506–511.

[26] Anderson Rodrigo da Silva. *biotools: Tools for Biometry and Applied Statistics in Agricultural Science*. R package version 4.2. 2021. URL: https://cran.r-project.org/package=biotools.

[27] Gökcen Eraslan et al. "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature communications* 10.1 (2019), pp. 1–14.

[28] Ales Varabyou, Steven L Salzberg, and Mihaela Pertea. "Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments". In: *Genome research* 31.2 (2021), pp. 301–308.

[29] Yarui Li et al. "Long non-coding RNA SNHG5 promotes human hepatocellular carcinoma progression by regulating miR-26a-5p/GSK3β signal pathway". In: *Cell death & disease* 9.9 (2018), pp. 1–15.

[30] Julienne M Mullaney et al. "Small insertions and deletions (INDELs) in human genomes". In: *Human molecular genetics* 19.R2 (2010), R131–R136.

[31] Vikram Agarwal and David R Kelley. "The genetic and biochemical determinants of mRNA degradation rates in mammals". In: *bioRxiv* (2022).