Performance and comparison of current sequence-based models on individual gene expression prediction

Connie Huang
Nilah Ioannidis, Ed.

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-83
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-83.html

May 10, 2023

Acknowledgement

I would like to express my deepest gratitude to Professor Nilah Ioannidis, who introduced me to the field of machine learning in genomics, guided me through research for the last 3 years, and made sense of the computations and figures I came to her with. Additionally, I would like to extend a huge thank you to Richard Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, and Pooja Kathail for essential contributions with the models analysed in the paper.

Performance and comparison of current sequence-based models on individual gene expression prediction

by

Connie Huang

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 Aaron Streets

Spring 2023

Performance and comparison of current sequence-based models on individual gene expression prediction

by Connie Huang

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**.

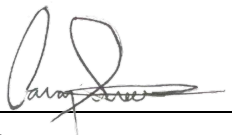
Approval for the Report and Comprehensive Examination:

Committee:

Professor Nilah Ioannidis
Research Advisor

(Date)

* * * * * * *



Professor Aaron Streets
Second Reader

05/09/2023

(Date)

Abstract

Performance and comparison of current sequence-based models on individual gene expression prediction

by

Connie Huang

Masters of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Nilah Ioannidis, Chair

State-of-the-art deep learning models have been recently utilized to understand factors that affect gene expression. Current models are able to predict gene expression well across different genes in the reference genome directly from the DNA sequence alone, but their performance on individual level variation has not yet been studied deeply. To investigate model capabilities on individual variation, we evaluate four models - Enformer, Basenji2, Expecto, and Xpresso - on personal genome data and find limited performance when explaining variation in expression across individuals. Some genes have strong negative correlations between predicted and observed expression levels, suggesting that the models have identified causal regulatory variant(s) but incorrectly predicted their direction of effect. Our comparison of all four models reveals that the models often disagree with one another on the predicted direction of genetic effects on expression, and that models agree more often on the magnitude than on the direction of genetic effects.

Contents

List of Figures

Acknowledgments

I would like to express my deepest gratitude to Professor Nilah Ioannidis, who introduced me to the field of machine learning in genomics, guided me through research for the last 3 years, and made sense of the computations and figures I came to her with. Additionally, I would like to extend a huge thank you to Richard Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, and Pooja Kathail for essential contributions with the models analysed in the paper.

Chapter 1

Introduction

1.1 Background

The human genome carries instructions for the expression of every gene in every cell type in the body. Perturbations in the genomic sequence from person to person make us unique; however, adverse mutations in the genome may cause disregulation of important genes and may lead to diseases. Here we are interested in understanding effects of variation in the genome and evaluating recent deep learning methods that have been built to predict gene expression from just the DNA sequence alone.

Previously, thousands of genome-wide associations studies (GWAS) have been performed to scan the entire genome for variants that have associations with disease [1]. Similar association studies help find expression quantitative trait loci (eQTLs), locations in the genome that explain the genetic variance in expression for many genes. One problem with GWAS is that it requires genome data from tens of thousands of individuals to obtain statistically significant results; which means these studies are only feasible for common phenotypes that exist in many individuals, and they require lots of genetic data, which is expensive to obtain. In addition, GWAS may find many variants in association with a particular disease, but true causal variants are difficult to pinpoint due to linkage disequilibrium. Finally, many other biological factors such as transcription factors, chromatin accessibility, histone markers, etc. also have effects on the expression of a gene, and these effects are again difficult and costly to measure in a lab [2].

Most recently, deep learning models have utilized convolutional neural nets to predict epigenetic features and gene expression from the DNA sequence alone [3]. Using the DNA sequence alone to predict gene expression (as opposed to using additional epigenetic features) allows us to predict the role of sequence variation in regulating expression. Convolutional neural networks treat the sequence as a 4xn pixel image (where n is the input sequence length) and find filters based on important motifs in the sequence. However, due to the limited range of the receptive field, the model captures genetic effects on expression for a certain gene only within a limited window around the gene transcription start site (TSS).

This is one critical weakness of existing deep learning methods, because the model could miss important causal distal variants that lie outside of its receptive field. The latest and highest performing of these methods, Enformer [4], uses a transformer architecture in addition to the convolutional neural network architecture from its predecessor, Basenji2 [5], to model the sequence, which allows the model to capture much longer range interactions than the previous convolutional neural net architectures, as described below.

1.2 Model Architectures

Each of the four models tested here has a unique architecture which affects how it predicts expression. Enformer [4] uses 7 standard convolutional layers and 11 additional transformer layers trained on 5,313 tracks (including the lymphoblastoid cell line CAGE data which we focus our study on) and allows for a receptive field of up to 196kb. Predictions are outputted in 128 base pair bins along a 114kb sequence. Enformer's transformer architecture allows for this long reaching receptive field and therefore its predictions have a large sequence context. By contrast, Basenji2 [5], trains on the same 5,313 tracks, utilizes the same standard convolutional layers, but instead of transformer layers, they use additional dilated convolutional layers to widen the sequence context to 44kb. Despite the difference in receptive field, we will show that Basenji2 and Enformer perform similarly on personal transcriptome data, likely due to their shared underlying convolutional layers and the fact that they use the same training data.

The third model, Expecto [6] is built upon an older convolutional model, DeepSea [7], which uses a convolutional neural net to predict epigenetic features from sequence. Expecto uses the epigenetic features from DeepSea and applies a spatial transformation with exponential decay functions to reduce the dimensionality of the data and compute predictions. They use Roadmap, GTEx, ENCODE expression and regulatory feature data, and CAGE peaks to train the model. Expecto's architecture is able to attain a receptive field of 40kb.

The last model we will evaluate, Xpresso [8], uses the smallest receptive field of 10.5kb asymmetrically around the TSS of the gene of interest, focusing on the promoter region of the gene, as this is the region which contributes the most to the gene expression. However, with this sequence context size, Xpresso's prediction would not be able to incorporate effects from variants farther away from the TSS that may be regulating the expression of the gene. Xpresso's architecture contains two convolutional blocks followed by two fully connected layers and is trained on RNA-sequencing data from Epigenomics Roadmap Consortium across 56 tissues and cell lines.

1.3 Limitations of current models

In this paper, we highlight an important problem seen in four state of the art sequence-to-expression models (namely, Enformer, Basenji2, Expecto, and Xpresso): while all of the

models perform well predicting gene expression on the reference sequence, they all perform poorly on personalized sequences. Each method is trained with the reference genome, where training examples are made up of windows along the reference genome. As a result, the methods perform well at explaining variation across genes at different locations in the genome, but we show that they perform much worse when tasked with identifying the much more granular differences between individuals for a given gene. Previously Karollus et al [9] evaluate several deep learning models including Enformer, Basenji, and Xpresso, and find that models can predict effects within promoters near the TSS of a gene, yet perform poorly when predicting effects of more distal enhancers, highlighting yet another issue of the models.

Here we examine and compare the performance of Enformer, Basenji2, Expecto, and Xpresso on paired whole genome sequencing and RNA-sequencing data from the Geuvadis consortium [10] and show that model predictions poorly explain the variation between individuals. In addition, models sometimes predict the wrong direction of effect of some single nucleotide genetic variants (SNVs). A new study by Sasse et al. [11] similarly evaluates Enformer on whole genome sequencing and paired transcriptome data, and agrees with our finding that the model often predicts the incorrect direction of variant effects on gene expression.

Chapter 2

Methods

2.1 Dataset

We use paired gene expression and whole genome sequencing data from individuals in the 1000 Genomes Project [12] from the Geuvadis consortium [10]. 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. In comparisons to the most significant eQTL per gene, we select the most significant eQTL identified from the Geuvadis European samples within 20kb of the gene TSS. For all analysis, we evaluate all models on the set of genes that contained at least one significant eQTL in the Geuvadis European eQTL analysis.

2.2 Constructing Personlized Input Sequences

For each gene, the ENSEMBL gene ID, TSS position, strand, and chromosome were obtained from Geuvadis. The gene symbol was converted from the gene ID using BioTools [13]. Each method has a different receptive field and thus required separate personalized input sequences of the appropriate length. Xpresso uses an asymmetric input sequence, and therefore, the input sequence depends on the orientation of the gene. For genes located on the positive strand, we directly computed the personal sequences using bcftools consensus [14]. For genes located on the negative strand, we extracted the reference sequence from 3.5kb upstream to 7kb downstream of the TSS using samtools, applied bcftools consensus, and then took the reverse complement. The input sequences for Expecto, Basenji2, and Enformer are all symmetric about the TSS and were computed identically to genes on the positive strand for Xpresso. We considered only single nucleotide variants (SNVs) and did not include indels when creating the personalized input sequences. We used hg19 as the reference genome for creating personal sequences, since the Geuvadis dataset is available in hg19. After creating

these sequences, we verified correctness by comparing the number of variants expected from the VCF file to the edit distance between the reference and personal sequences. We predicted gene expression levels for each individual by averaging the predictions from both haplotypes.

2.3 Obtaining predictions

Basenji2

Basenji2 runs on input sequences of 131kb with an effective receptive field of 44kb for each prediction. The model outputs predictions in 128-bp bins for 5,313 epigenetic and transcriptional tracks from the ENCODE, Roadmap, and FANTOM consortiums. We used Basenji2 predictions for the GM12878 lymphoblastoid cell line CAGE track, as it is the most relevant cell type for the Geuvadis expression data. For a given input sequence centered at a gene TSS, we averaged predictions over the forward and reverse complement sequence and minor sequence shifts to the left and right (1 nucleotide in each direction). To compute the final expression prediction for the gene, we averaged the predicted CAGE signal over the 128-bp bin containing the TSS, the 5 bins upstream of the TSS, and the 5 bins downstream of the TSS.

Enformer

Enformer replaces the dilated convolutions of Basenji2 with a self-attention mechanism, which facilitates the learning of long-range dependencies. Enformer has a receptive field of 196kb and outputs predictions in 128 bp bins for the same 5,313 tracks that Basenji2 is trained on. We used Enformer predictions for CAGE measurements performed on the GM12878 lymphoblastoid cell line. While the Enformer authors averaged predictions within a 3-bin window around each gene TSS, we found that averaging over a 10-bin window led to better performance on the Geuvadis dataset.

Expecto

ExPecto predicts gene expression by first using a convolutional neural network (Beluga, an updated version of DeepSEA [7]) to predict chromatin features within a 40kb region around each gene TSS. Specifically, Beluga outputs predictions in 200-bp bins for 2,002 epigenetic tracks from the ENCODE and Roadmap consortiums. To predict expression for a given gene, Beluga is first used to predict chromatin features for 200 bins centered around the TSS, averaging predictions over the input sequence and its reverse complement. The resulting predictions are spatially transformed with a set of basis functions and used as input features for an L2-regularized linear regression model to predict expression for the given input sequence. For our expression predictions, we used a publicly available ExPecto

model trained on EBV-transformed lymphocytes from GTEx, which we chose as the most relevant cell type to compare with the Geuvadis expression data.

Chapter 3

Results

We use RNA-sequencing data from the Geuvadis consortium, measured on lymphoblastoid cell lines, and whole genome sequences from 421 individual genomes to evaluate model performance on individual level variation. We focus on the 3,259 genes for which Geuvadis found at least one significant eQTL to ensure that all of our genes have some genetic association to gene expression. Sequences for each individual are constructed by inserting each single nucleotide variant (SNV) from that individual into the input sequence around the transcription start site of each gene; then for all four models and all 3,259 genes, gene expression predictions for each individual and the reference sequence are computed. We use the lymphoblastoid CAGE track outputs from Enformer and Basenji2, the EBV-transformed lymphocyte expression output from Expecto, and the LCL-specific Xpresso model for our final predictions since these outputs are closest to cell type used for the measured expression from Geuvadis.

3.1 Models explain variation across genes, but not across individuals

First, we validate the performance of the models on the reference sequence to confirm what was found in the original papers. Comparing reference sequence predictions with the median Geuvadis expression levels over all individuals for each gene, we find Spearman R correlations of 0.57 for Enformer (Fig 3.1a), 0.52 for Basenji2 (Fig 3.1b), 0.54 for Expecto (Fig 3.1c), and 0.33 for Xpresso (Fig 3.1d), which are similar to the findings in the original publications and show that these models can explain variation across genes.

To further confirm this finding, we compute for each individual Spearman R correlations between predicted expression and measured expression across genes, and find that the average correlation coefficient closely resembles the reference genome correlation (Fig 3.2). Each individual sequence is made up of the reference sequence with a set of single nucleotide variants; therefore, the models treat these individual sequences for each gene almost the same as they treat the reference sequence. The correlations within each individual across

genes in general show how similar the model treats the reference sequence compared to individual sequences when explaining variation in expression across all genes, likely due to similar sets of genes being highly or lowly expressed in all individuals.

We then compute the Spearman R correlation between predicted and measured expression for each of the 3,259 genes across all 421 individuals, and find that the correlations are centered around 0, demonstrating the models' poor performance in explaining variation across individuals (Fig 3.3). We see from the histograms in Fig 3.3, that for the majority of genes, model predictions across individuals have almost no correlation with measured expression levels. For some genes, however, the story is different. The tail ends of the histograms represent some genes with very strong positive and negative correlations with the measured expression levels across individuals, showing that, for some genes, models perform very well, and for other genes, they get the magnitude of effect right but the direction of effect wrong. We next investigate this result in more detail.

3.2 Models sometimes predict the wrong direction of effect even for genes with strong eQTLs

Interestingly, for some genes, model expression predictions have very strong positive correlation with measured expression, and others have very strong negative correlation. To investigate why this occurs or if certain genes are better predicted than others, we consider the strength of eQTLs for these genes, as measured in the Geuvadis cis-eQTL analysis, because eQTL variants have significant associations with gene expression.

First, we look closely into some of the genes with a top (most significant) eQTL within 20kb of the gene's TSS, which is inside the receptive field for three of the models, Enformer, Basenji2, and Expecto. We compare the individual predictions for each model to the prediction of the reference sequence with or without the alternate allele inserted. For some genes, the model predicts that the alternate allele of the top eQTL has expression significantly different from the predicted reference expression (e.g. SNHG5) (Fig 3.4). For other genes, the model predicts that the alternate allele has no effect on the gene expression (e.g. PEX6) (Fig 3.5). These results suggest that the models find some variants that are important to gene expression; however, the direction of effect of the alternate variant is sometimes wrong, and this causes the correlation of expression prediction to be strongly negative, as seen with the SNGH5 predictions. Xpresso was the only model which correctly predicted the direction of effect for SNHG5, and also the only model which predicted the alternate variant to increase the expression level from the reference prediction (Fig 3.4d). For PEX6, the models all make predictions that have a strong positive correlation with measured expression, despite predicting no effect of the top eQTL variant, suggesting that the models likely pick up on a different variant as being causal.

We then look at the trend of the significance of the strongest eQTL compared with the cross-individual correlation of the gene to see if the p-value significance is related to the

correlation (Fig 3.6). We find that while the genes with the most significant eQTLs often have the largest magnitude correlations, there is no strong trend towards these genes being positively correlated rather than negatively correlated, indicating that the direction of effect is often predicted incorrectly even for these genes.

Similarly, we investigate the effect sizes (R-values) of the top eQTLs compared with cross-individual correlation (Fig 3.7), and find that genes with eQTLs with the strongest effect sizes have generally high magnitude correlations, but no trend towards positive correlations. Our results indicate that when genes have variants with strong effects on expression, the models can often pick up those variants and incorporate their predicted effects into the overall expression prediction, while the models struggle more with predicting cross-individual expression of genes with no strong eQTL within the receptive field of the model. In general, though, the models struggle to predict aggregate effects of variation in personal genome sequences, and are just as likely to predict incorrect rather than correct directions of effect of causal regulatory variants.

Lastly, we also hypothesized that the range of predicted expression levels (maximum expression minus minimum expression) might be related to the predicted direction of effect, with smaller ranges (indicating less predicted effect of genetic variation on expression overall) being more likely to have an incorrect predicted direction of effect (negative cross-individual correlation). We find a slight trend towards strongly positively correlated genes having a larger range of predicted expression, and strongly negatively correlated genes having a smaller range (Fig 3.8). This may imply that models struggle more to predict the direction of effect for genes with a smaller range of predicted expression.

3.3 Model comparisons

We compare the Spearman R correlation between models using all predictions over all individuals and genes (Fig 3.9), and find that the models generally have higher correlations with each other than with the Geuvadis measured expression levels. For Basenji2 and Enformer, high correlation can be explained by similar architectures and training strategies, with Enformer performing a little better on the observed expression than Basenji2 due to Enformer's addition of the transformer architecture in place of Basenji2's dilated convolution layers. For both Expecto and Xpresso, their predictions are also more correlated with the other models than they are with the measured Geuvadis levels, despite very different training data and modeling approaches.

Finally, we compare the cross-individual Spearman R correlations between Enformer and all other models pairwise, and we see enrichment of correlations along the $y = x$ and $y = -x$ axes (Fig 3.10). This implies that the models tend to agree more on the magnitude of correlation than on the direction of correlation, further supporting the conclusion that these deep learning models recognize the presence of important regulatory variation but struggle with understanding the direction of effect of such variation.

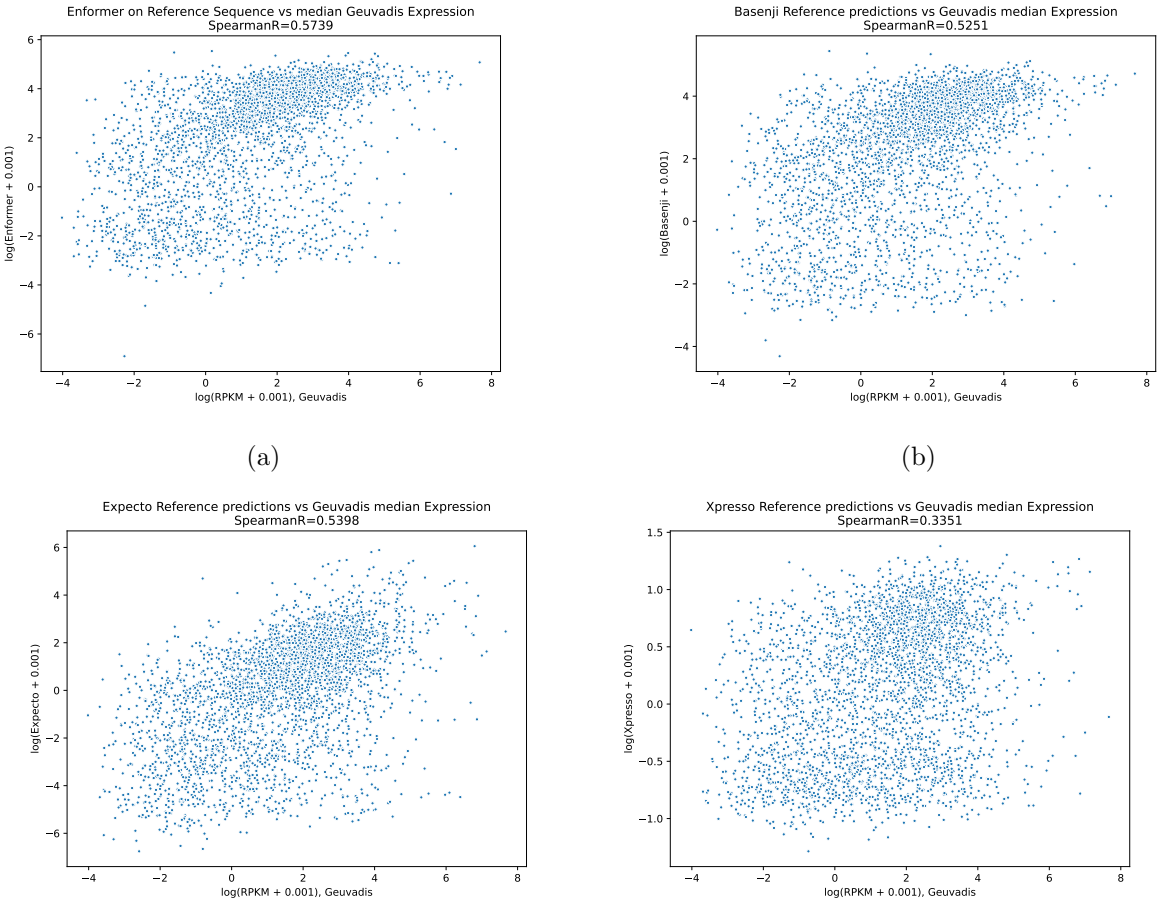


Figure 3.1: Reference predictions on all 4 models compared to observed expression.

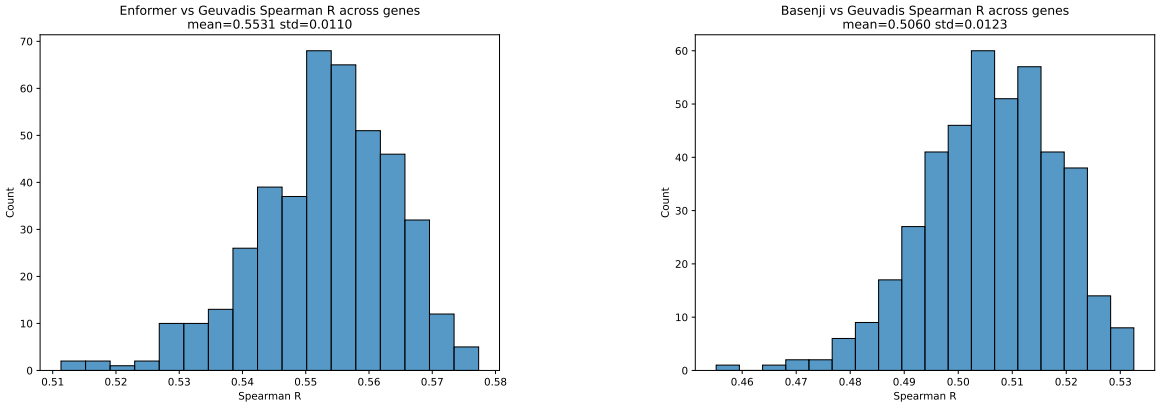


Figure 3.2: Histogram of Spearman R correlations across genes. We see high average correlations across genes in each individual, which are similar to reference prediction correlations.

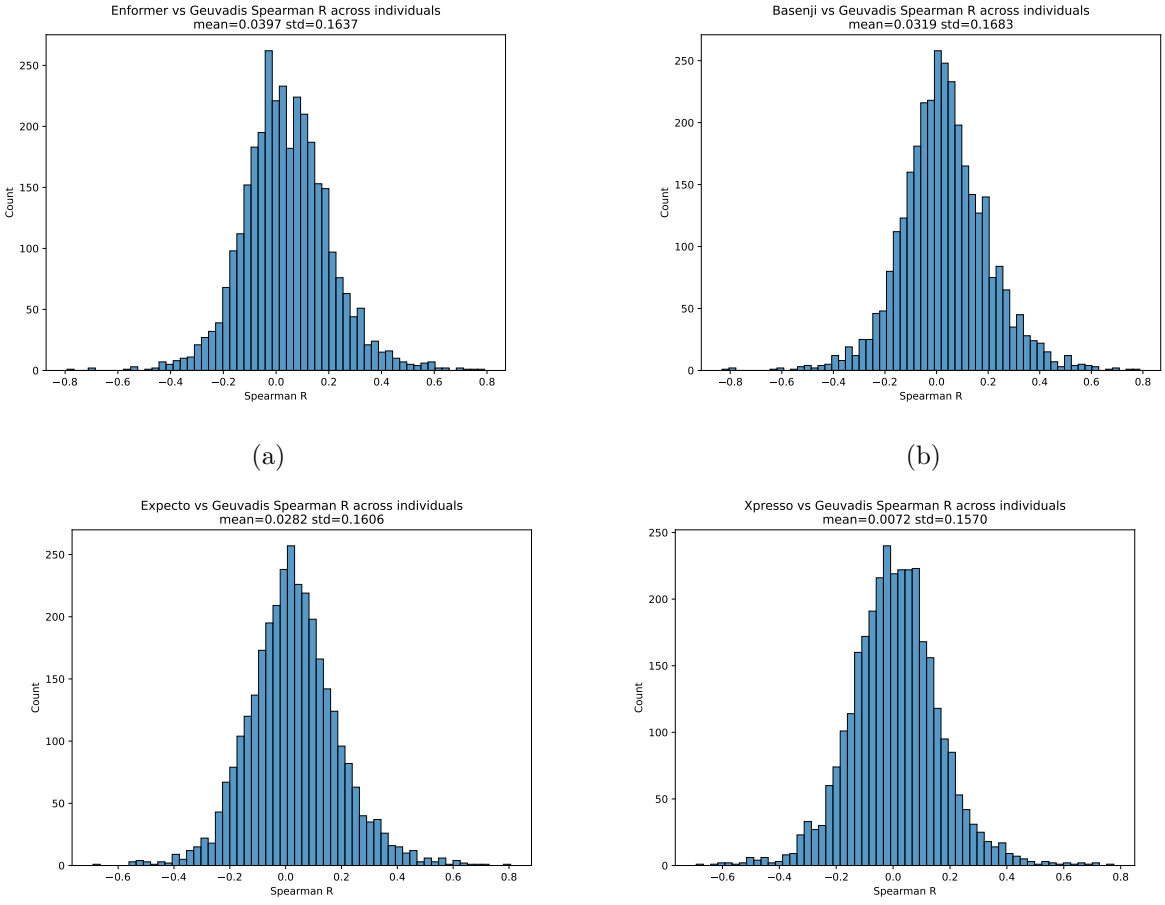


Figure 3.3: Histogram of Spearman R correlations across individuals. The correlations across individuals are centered around 0.

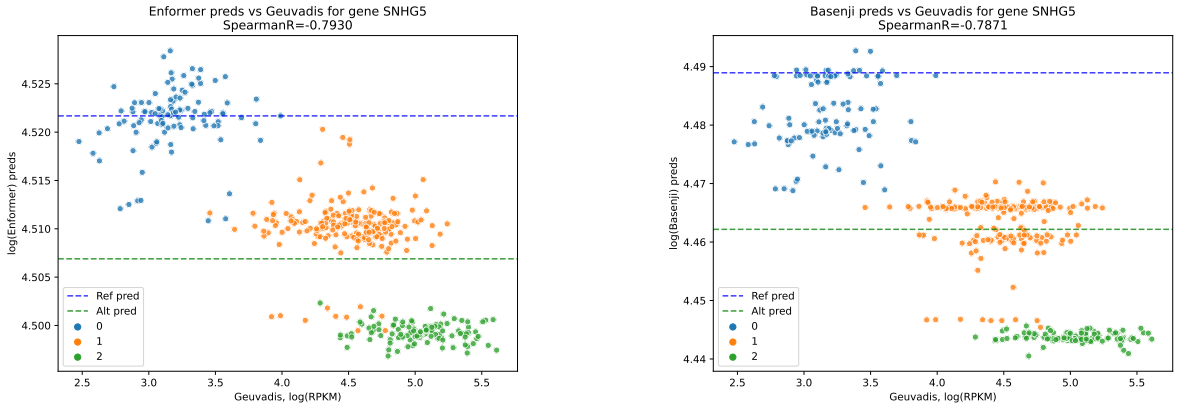


Figure 3.4: SNHG5 predictions. The 3 striations seen in the plots align with the number of copies each individual has of the most significant eQTL within 20kb of the TSS (dosage 0, 1, or 2). The alternate prediction is the model's prediction of the reference sequence with the most significant eQTL inserted. For SNHG5, we see Enformer, Basenji, and Expecto predict a large effect on the expression value from the eQTL variant, although in the wrong direction.

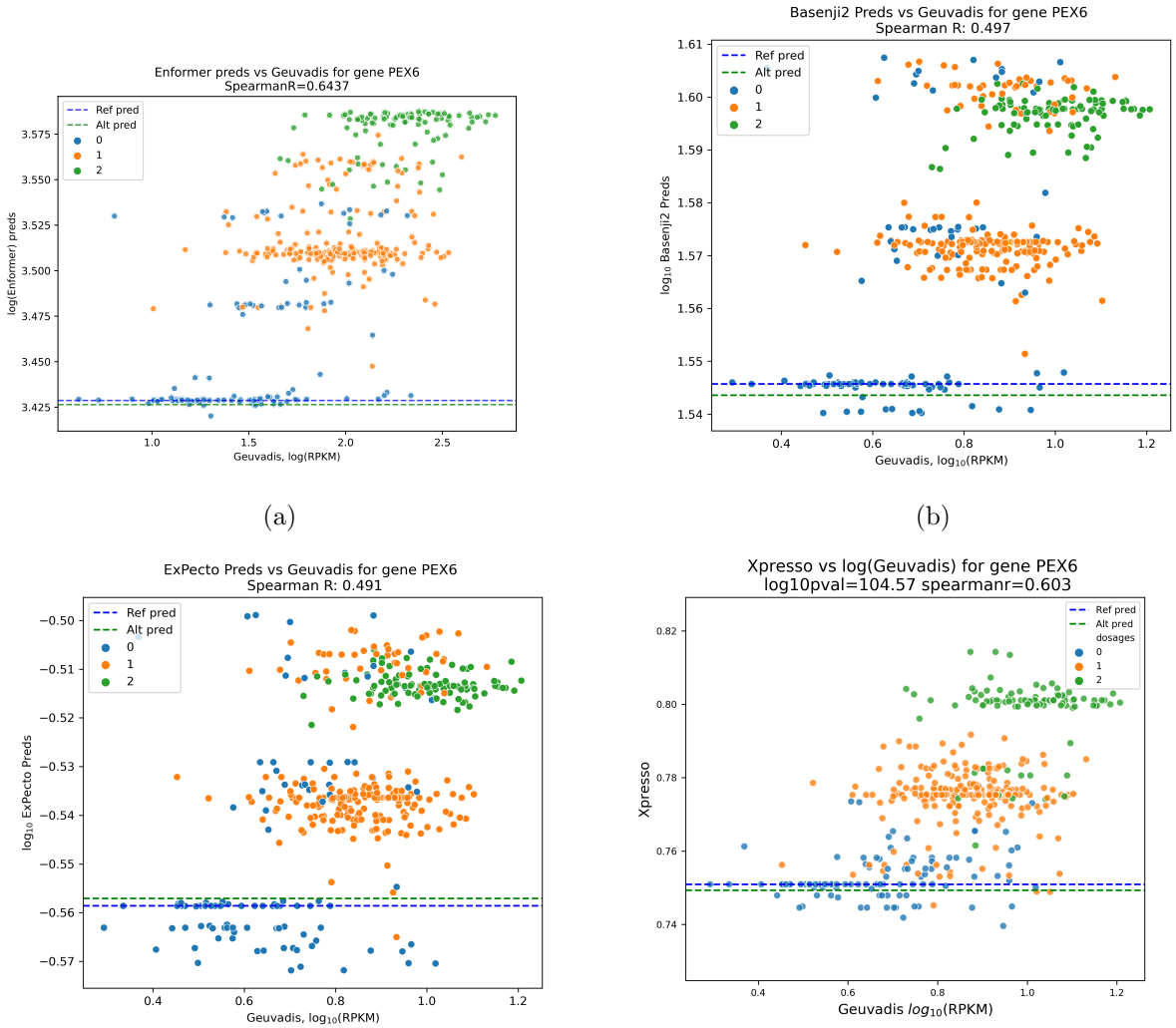
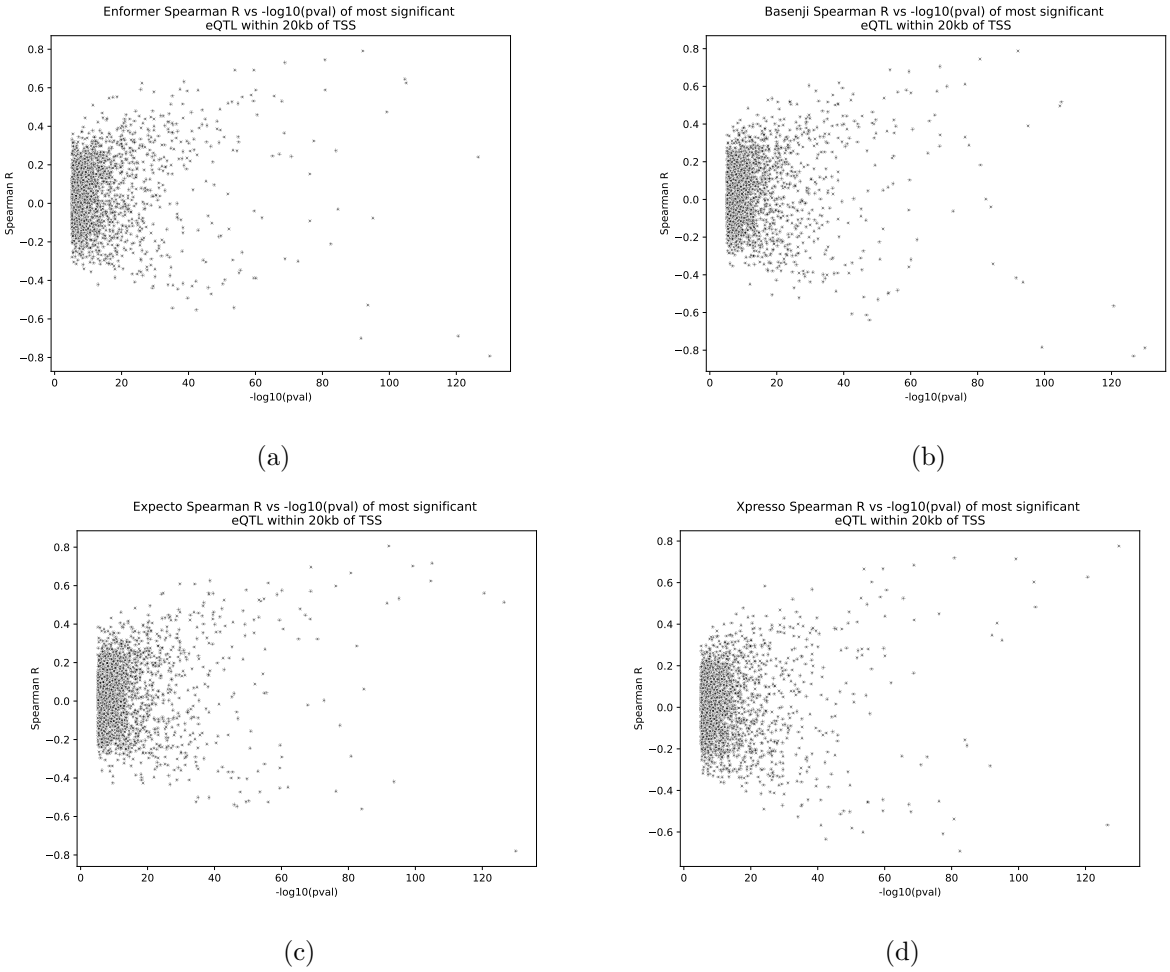


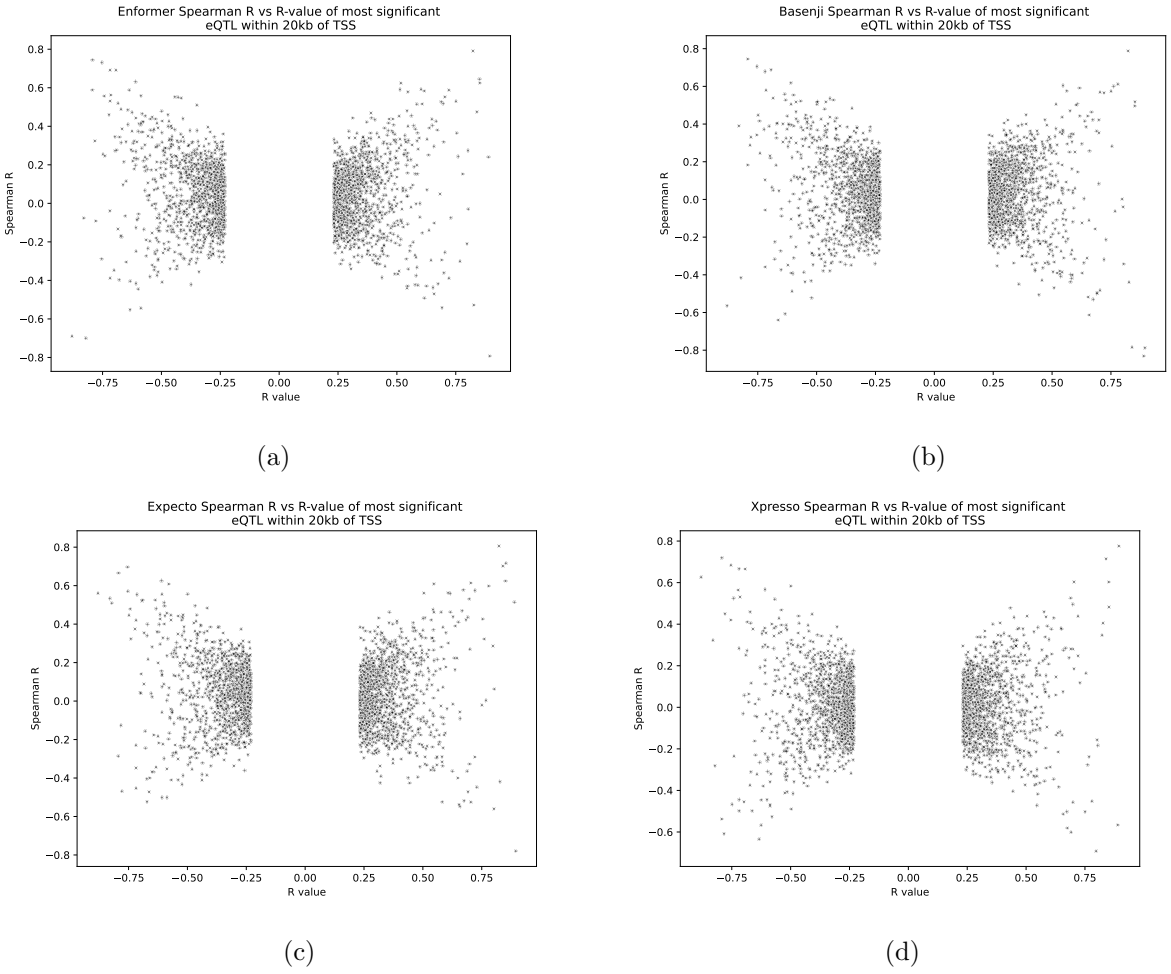
Figure 3.5: While the models all have strong positive correlations for predicting PEX6, none of the models attribute the expression variance to the most significant eQTL, as seen from the alternate prediction being almost the same as the reference prediction for all models.



(a) (b)

(c) (d)

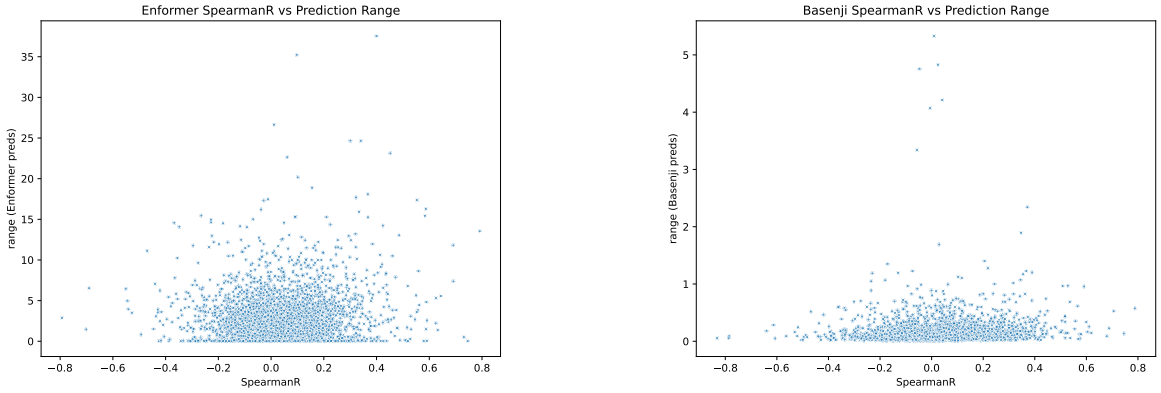
Figure 3.6: Spearman R Coefficient of predictions and Geuvadis expression levels across individuals for all genes plotted against -log10(pvalue) of the gene's most significant eQTL as measured by Geuvadis.



(a) (b)

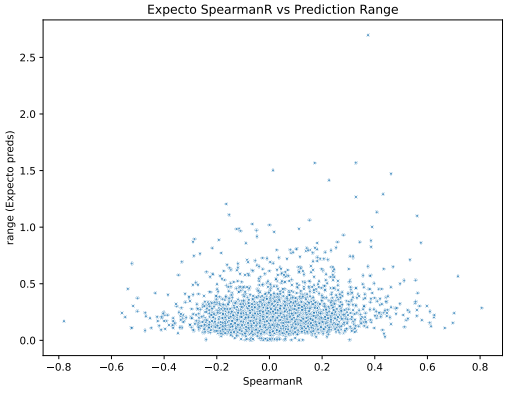
(c) (d)

Figure 3.7: Spearman R Coefficient of predictions and Geuvadis expression levels across individuals for all genes plotted against the effect size (R value) of the gene's most significant eQTL as measured by Geuvadis.

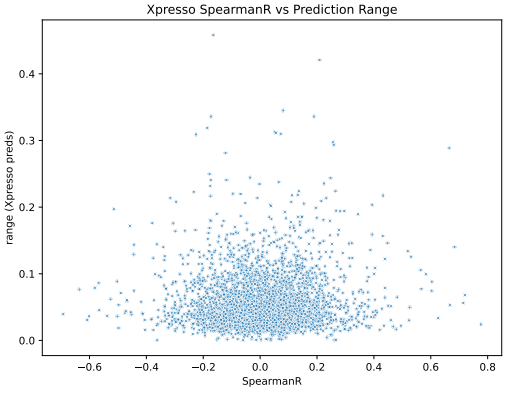


(a)

(b)

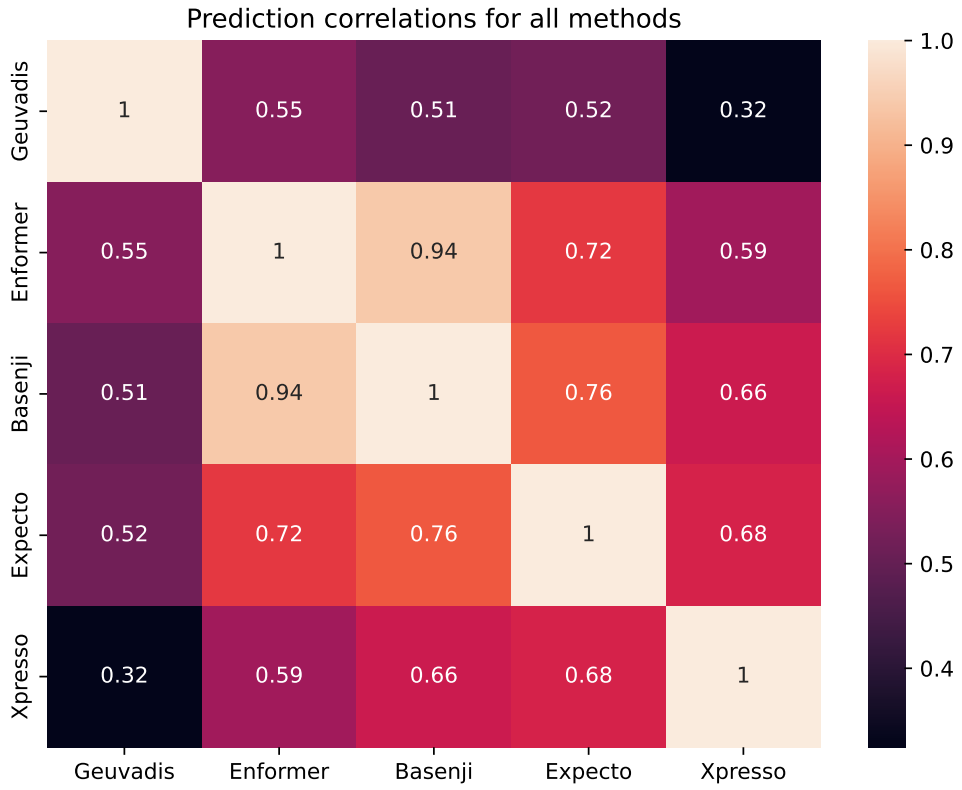


(c)



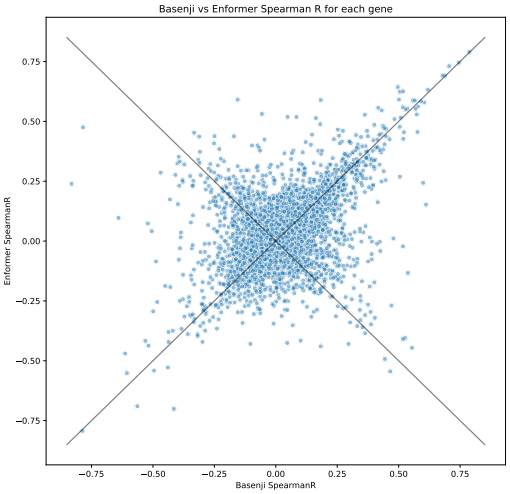
(d)

Figure 3.8: Range of predicted expression values vs. Spearman R across individuals for each gene.

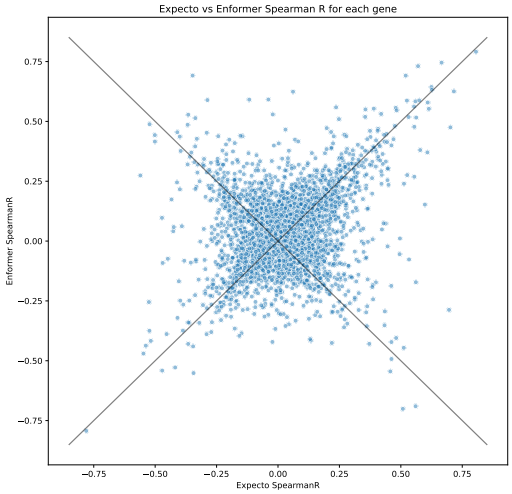


(a)

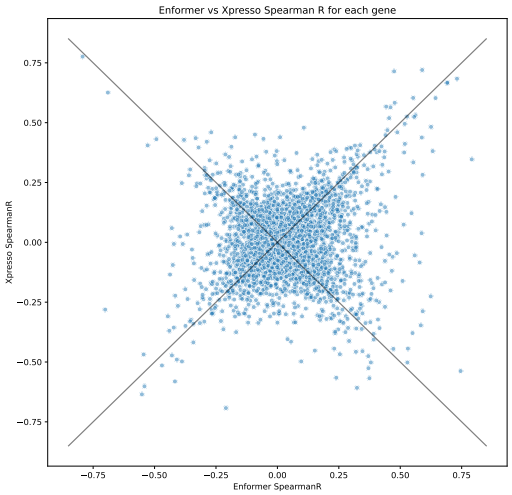
Figure 3.9: Spearman R correlations of all predictions from all models, as well as observed expression. Models are more correlated with each other than with the observed expression.



(a)



(b)



(c)

Figure 3.10: Cross comparison of cross-individual correlations between Enformer and other models. The increased number of genes along the $y = -x$ axis shows that models often disagree on the direction of effect of genetic variation on expression.

Chapter 4

Conclusion

In conclusion, we implement an analysis framework for evaluating sequence-to-expression deep learning methods, Enformer, Basenji2, Expecto, and Xpresso, on personal genomes, and we find that while models perform well predicting reference genome expression across different genes in the genome, they fail at predicting expression across individual genomes. We find that for genes that have strongly associated variants (eQTLs), the models are more likely to identify causal regulatory variants; however, they sometimes predict the wrong direction of effect. These errors in direction of effect are not consistent across models - in fact, we find that models often disagree on the direction of effect of strong eQTLs but often agree on the magnitude of correlation for these genes.

Preceding this work, Baokar [15] did significant work uncovering the same problem of poor expression predictions across personal genomes, primarily for Xpresso and Basenji2. The current work expands on [15] by extending the analysis of all 3,259 genes to Enformer and Expecto, by comparing predictions across all four models, and by further investigating the properties of genes with strong negative correlations. In addition, the recent work of Sasse et al. [11] finds similar results for Enformer using a different gene expression dataset, where Enformer also predicts the wrong direction of effect for some genes. Other recent studies have been done on Enformer showing that Enformer captures gene expression factors near the promoter region of the gene more accurately than it captures factors that affect gene expression far from the TSS [9].

Following this work, further investigation of the underlying reason that models fail at predicting individual level expression would help provide insight into how to improve the design and training of such models for the task of personal genome interpretation. Ideally, with additional data we could compare the performance of the models not only on expression, but on other epigenetic tracks as well, to see if the models have more trouble predicting gene expression than they do for other regulatory factors, such as histone modifications and chromatin accessibility. With more investigation and an ever more complete picture of the weaknesses of these models, better models can be constructed to pinpoint the molecular and cellular consequences of genetic variation across individuals.

In the future, if genomic deep learning models are successful in predicting personal gene

expression levels from personal genome sequence in the promoter region near the gene TSS, we propose to conduct eQTL analyses after subtracting out the predicted effect of the promoter region on gene expression. Since the promoter region of a gene has the biggest impact on expression, we hypothesize that removing the effect of promoter region variation on the expression level of a gene and considering only its residual expression level across many individuals may improve the power of association tests for distal eQTLs and help discover other causal variants that do not affect the promoter region but that contribute to the expression of the gene. Such analyses and the insights that deep learning methods offer us in understanding the sequence determinants of gene expression can help to shed light on the function and organization of the non-coding regions of the genome.

Bibliography

[1] Melinda C. Mills and Charles Rahal. "A scientometric review of genome-wide association studies". In: *Communications Biology* 2.1 (2019), p. 9. DOI: 10.1038/s42003-018-0261-x. URL: https://doi.org/10.1038/s42003-018-0261-x.

[2] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. *From genome-wide associations to candidate causal variants by statistical fine-mapping*. Aug. 2018. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050137/.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. "Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers". In: *bioRxiv* (2022).

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

[11] Alexander Sasse et al. "How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks?" In: *bioRxiv* (2023), pp. 2023–03.

[12] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68.

[13] 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`.

[14] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2 (Feb. 2021). ISSN: 2047-217X.

[15] Parth Baokar. "Evaluating the use of sequence-to-expression predictors for personalized expression prediction". In: ().