Using Pair-wise Scores Derived From Epigenomic Features to Predict Phenotypic Similarity Between Genetic Variants

Elijah Jones[1], Saiyang Liu[1], Ha Vu[2], Jason Ernst[3]

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA [2] Bioinformatics Interdepartmental Ph.D. Program, UCLA
[3] Department of Biological Chemistry, David Geffen School of Medicine, UCLA

Abstract

Genome-wide association studies (GWAS) provide an unprecedented opportunity to uncover genotype-phenotype associations. Identifying genetic variants whose epigenetic properties influence their shared associations with certain phenotypes is an important challenge, especially in studies of the biological implications of GWAS variants. Here, we explore the possibility of developing a pair-wise score for genetic variants to measure their associations with similar phenotypes. We first analyzed the patterns of epigenetic signals that differentiate pairs of shared-phenotype GWAS variants based on data from Enformer–a deep learning framework that predicts >5,000 epigenetic features from DNA sequences. Then, we trained a fully-connected Siamese Network with a contrastive loss function to measure the epigenetic distances among variants and hence offer a proxy for their phenotypic similarity. This study offers evidence that epigenetic signals present at each variant in a pair can be predictive of their associations with similar phenotypes and diseases. The resulting score will be a resource for studies of human genetic variants and their associated phenotypes based on the closely-related variants at functional levels.

(1) Background

- GWAS studies have generated a myriad of associations between genetic variants and many traits and diseases. The CAUSALdb database identifies credible sets of potential causal GWAS variants of 3,683 unique phenotypes using fine-mapping (Wang et al., 2020).
- In this study, we explore the possibility of developing a measure of distance between a pair of genetic variants that is predictive of whether they share an association with certain phenotypes. We propose that such a measure should be developed through data on the variants' epigenomic features.
- Enformer, a deep learning framework, accurately predicts signals of 5,313 functional assays (DNase, Chip-seq, transcription factor bindings, etc.) in various cell types from DNA sequences by recognizing the effects of distal regulatory elements up to 100 kilobases (kb) away from the variant of interest (Avsec et al., 2021) (**Fig 1**). Variant effect can be evaluated *in silico* using the Enformer model by taking the difference between predicted signals of the 5,313 epigenetic features at the reference and alternative allele.

Enformer model architecture and outputs

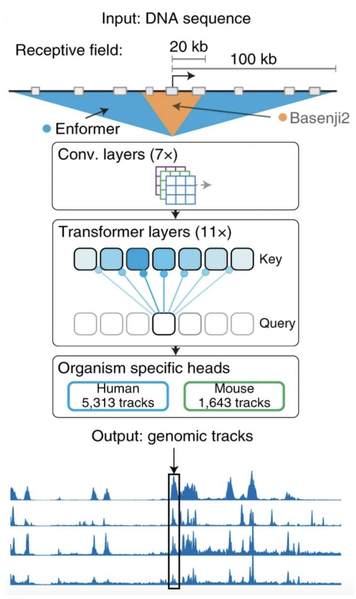

Figure 1: Enformer model architecture and outputs. The Enformer model increases the receptive field of a previous state-of-the-art model, Basenji2, through use of 7 convolutional layers and 11 transformer layers. The model was trained to predict 5,313 human and 1,643 mouse genomic tracks. (Figure adapted from Avasec et al., 2021.)

(2) Exploration of effect of shared phenotype associations on variants' pairwise correlations of functional assays

- We obtained data of **(1)** predicted variant effects on 5,313 functional assays for all common variants (MAF >= 0.05) in the 1000 Genomes project from Avsec et al., 2021, and **(2)** fine-mapped variants associated with 3,683 phenotypes from CAUSALdb (Wang et al., 2020).
- We generated pairs of variants with shared phenotypes, denoted positive pairs, and pairs of variants without a shared phenotype, denoted negative pairs.
- For each pair of variants, we calculated the Pearson correlation between the vectors of 5,313 functional assay SNP Activity Difference (SAD) scores for each variant of the pair. To focus on the magnitude of the variant effect, we used the absolute values of the SAD scores for 5,313 assays as input features.
- We generated histograms of correlations between pairs of variants that were generated from within the same chromosome (**Fig 2A**) and pairs sampled from different chromosomes (**Fig 2B**).
- We observed a significant difference in the average correlations between positive and negative pairs given both sampling methods. However, the signal of correlation difference between (+) and (-) variant pairs was much larger when the variant pairs were sampled from within similar chromosomes. Thus, the difference is driven largely by the proximity of variant pairs.

Pearson correlation of predicted variant effect vectors between variant pairs

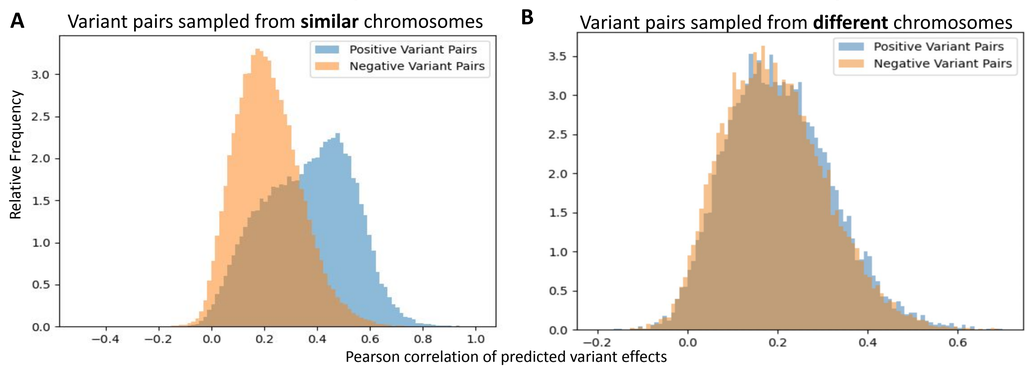

Figure 2: Pearson correlation of predicted variant effect vectors between variant pairs. (A) Histogram of Pearson correlations across all pairs of variants. Approximately 5000 positive and 5000 negative variant pairs were sampled from each of the 22 chromosomes. The mean correlation between positive pairs was greater than that between negative pairs ($P < 0.001$, one-tailed t-test, H_a: pos > neg). **(B) Similar to (A), but each variant pairs were sampled from different chromosomes.** The mean correlation between 18,000 positive pairs was greater than that between 18,000 negative pairs ($P = < 0.001$, one-tailed t-test, H_a: pos > neg).

(3) Development of a Neural Network For Scoring the Distance Between Pairs of Variants

- We developed a neural network model with a Siamese architecture consisting of three fully connected layers and a contrastive loss function to differentiate positive pairs of variants from negative pairs. The model was trained on 1 million positive and 1 million negative pairs, sampled from different chromosomes, based on the CAUSALdb variant database (**Fig 3**). Each variant's input features included the Enformer's predicted variant effects on the 5,313 function assays. We searched for optimal hyperparameters set based on the model's performance in recovering positive pairs of variants.
- To evaluate the model's performance, we first sampled 6,000 variants from 100 GWAS studies and generated the predicted pairwise distance for all possible pairs of variants. We also computed the Pearson correlation and Manhattan distance between the given pairs. For each variant v in the set, we define ρ_v as the number of variants $\hat{v}$ among the top 100 closest variants with v, based on each distance metric, that show shared phenotype associations with v. All three metrics performed better than expected performance ($E(\rho_v)=0.01$, i.e. we expected 1 variant with shared phenotype with v among the top 100 closest variants by chance). The neural network-trained distance metric, Pearson correlation, and Manhattan distance resulted in average ρ_v of 0.094, 0.099, and 0.089, respectively (**Fig 4A**).
- We next generated 5,000 variants taken from 100 GWAS studies and generated their predicted pairwise distance. We stratified the distance by the *non-cell-type-specific* chromatin state annotations overlapping each variant pair (Vu and Ernst, 2022). We observed that promoter-associated states show a higher average distance from other states (**Fig 4B**). In general, states of similar functional groups (transcription, heterochromatin, etc.) showed lower distances compared to those of distinct groups (**Fig 4B**).

Neural network procedures to predict variants' pairwise distance

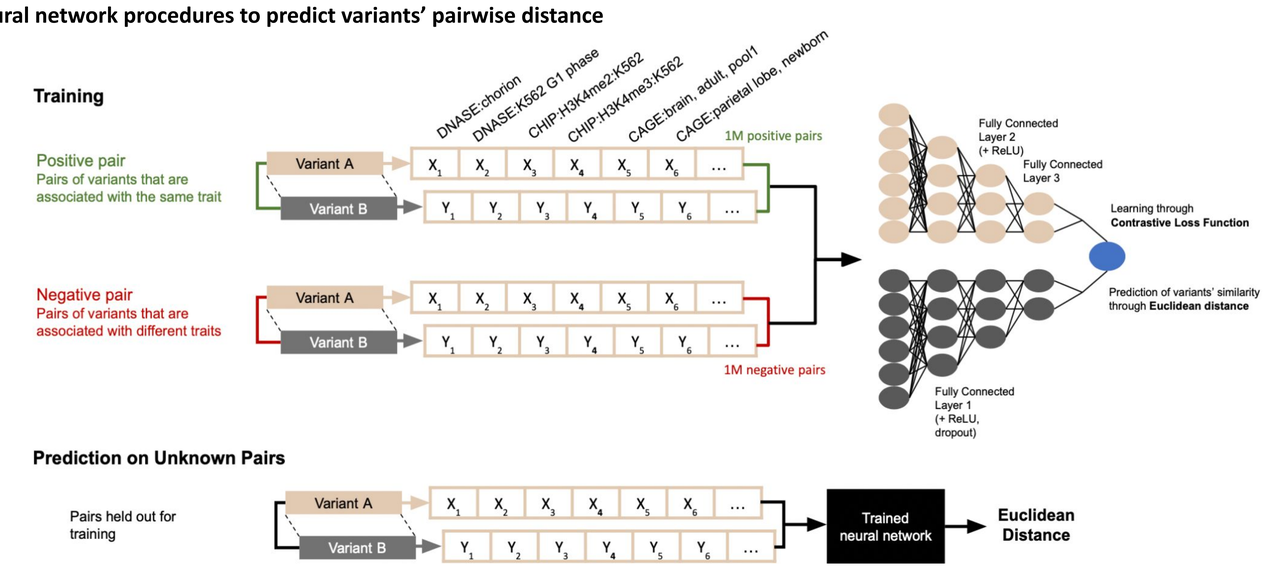

Figure 3: Neural network procedures to predict variants' pairwise distance. Neural network learning procedure. For each pair of variants, the two corresponding vectors of 5,313 features predicted by Enformer are given to the model. The model is trained to differentiate positive variant pairs, which are associated with the same trait, and negative variant pairs, which are associated with different traits. The training set consisted of approximately 1 million positive and 1 million negative variant pairs, all of which came from different chromosomes.

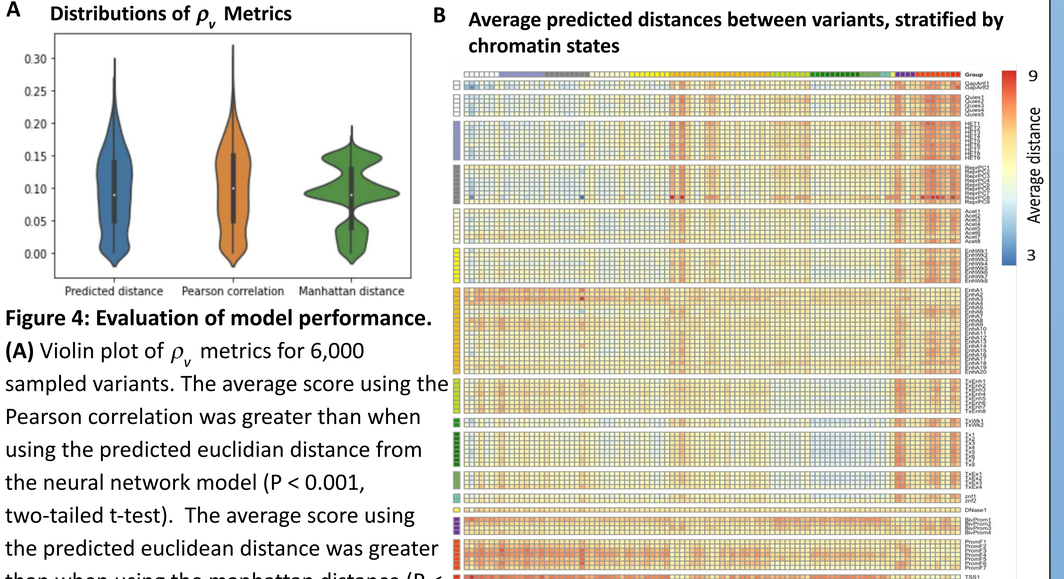

Figure 4: Evaluation of model performance. **(A)** Violin plot of ρ_v metrics for 6,000 sampled variants. The average score using the Pearson correlation was greater than when using the predicted euclidian distance from the neural network model ($P < 0.001$, two-tailed t-test). The average score using the predicted euclidean distance was greater than when using the manhattan distance ($P < 0.001$, two-tailed t-test). **(B)** Average predicted distances between variants, stratified by chromatin states.

(4) Conclusions and Applications

- This study offers evidence that epigenetic signals present at each variant in a pair can be predictive of shared associations with phenotypes.
- The pairwise score of the epigenetic distance between a pair of variants may be used as a reference in studies of human variants and their associated phenotypes.

(5) Acknowledgements and References

We thank Soo Bin Kwon, Alec Chiu, and Michael Kleinman for their help on this project.

Wang, J., Huang, D., Zhou, Y., et al. C *Nucleic Acids Res*. 48(D1):D807-D816 (2020)

Avsec, Ž., Agarwal, V., Visentin, D. et al. *Nat Methods* 18, 1196–1203 (2021)

Vu, H., Ernst, J. *Genome Biol* 23, 9 (2022)

Benhur S. J., (2020), GitHub repository, https://github.com/seanbenhur/siamese_net