A machine-learning based pairwise phenotypic similarity score for disease-associated genetic variants

Aahna Rathod, Luke Li, and Jason Ernst

Ernst Laboratory, Department of Biological Chemistry, David Geffen School of Medicine, UCLA

Abstract

Genome-wide association studies (GWAS) investigate the association between specific phenotypic traits and common variants in the human genome. These genetic variants can facilitate changes in gene expression through histone modifications, changes that can be reflected in the characterization of chromatin states. We reason that variants associated with the same trait should share functional similarities. Therefore, we propose a machine-learning based framework that computes a phenotypic similarity score for a pair of variants based on their functional annotations. We took variants from the EMBL-EBI GWAS catalog and annotated them with the recently developed universal ChromHMM state segmentations. We trained the model to distinguish between pairs of variants associated with the same trait ("positive pairs") and pairs associated with different traits ("negative pairs"), using the pair's correspondence of ChromHMM states as input features. Preliminary analysis indicates that positive and negative pairs have significantly different distributions of training features. Our results will offer insight into if epigenomic features can be predictive of variants' shared phenotypic association.

Background

- Genome-wide association studies (**GWAS**) : sequences genomes from large populations for single nucleotide polymorphisms (**SNPs**)
- Some SNPs are part of the non-coding genome and drive changes in gene expression through histone modifications
- Histone modifications can alter chromatin structure, creating possible chromatin states, changing gene expression, and causing phenotypic changes
- Chromatin states capture classes of genomic elements that cause downstream change, such as promoters and enhancers
- Recognized chromatin states give us an annotation of DNA elements that can be applied to GWAS

Methods

- Annotated variants in EMBL-EBI GWAS catalog with 24 stacked universal ChromHMM state segmentations
- Create positive and negative training pairs of variants
- Create training features for pairs based on the overlap between the chromatin states for each of the ChromHMM models
- **Goal**: train a machine-learning based framework that computes a phenotypic similarity score for a pair of variants based on their functional annotations

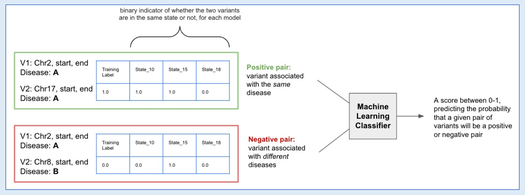

Fig 1. Sampling of states and positive pairs used to make features

Results

- Logistic Regression analysis on 1 million positive and negative pairs (500k each), using a 10-fold cross validation approach
- The AUROC curve predicts the probability of a binary outcome, which here is the difference between a pair being positive or negative. The orange logistic curve is shown to deviate slightly from the no skill curve
- Precision recall curves evaluate the skill of our prediction model. It is calculated from the machine-learning framework, and this figure shows the leftmost 3% section of the curve
- Heatmap measures the probability that chromatin state pairs, ranging from 1-100, will be predicted as a positive or negative pair. Values above 0.5 indicate a preference for positive pair prediction, and probabilities were calculated from the mentioned machine learning framework in the 10-fold testing process
- Possible chromatin state pairs with no pairing in 1m sample were given a score of zero, as indicated by strips of white space

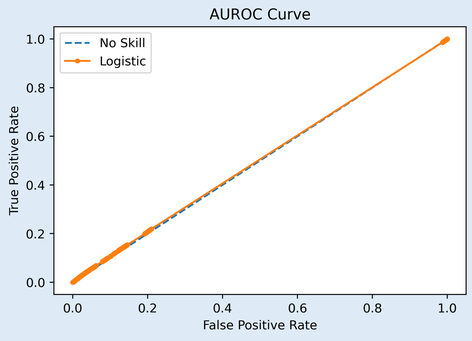

Fig 2. AUROC Curve from Logistic Regression classifier

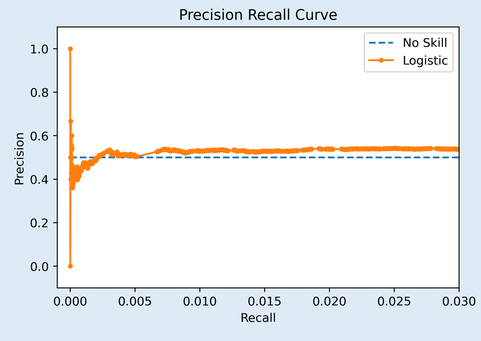

Fig 3. Precision Recall Curve from Logistic Regression classifier

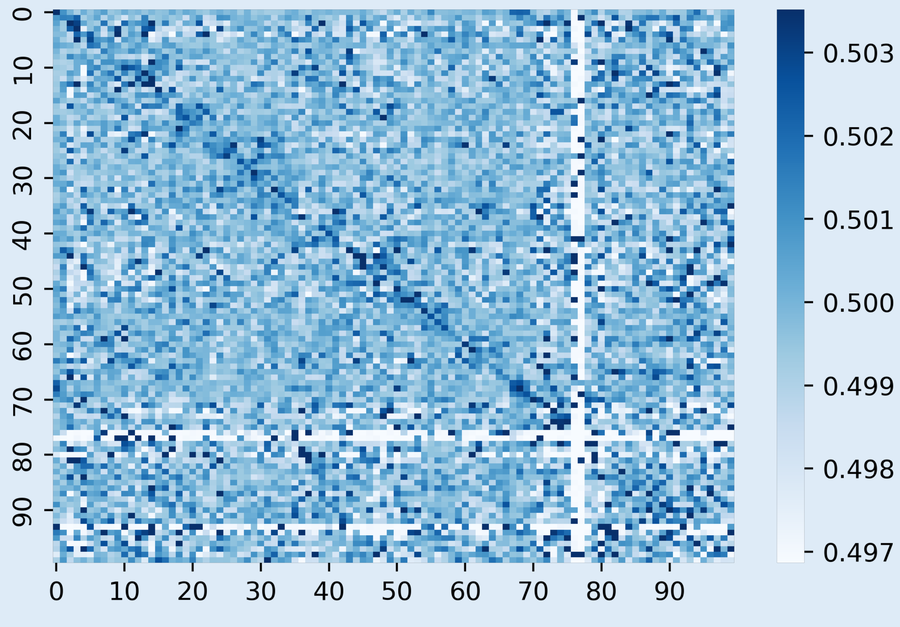

Fig 4. Heatmap measuring positive and negative pair prediction given two chromatin states

Results

- Training features for each pair were summed, and comparison between positive and negative pairs yielded significant differences → positive pairs averaged slightly higher

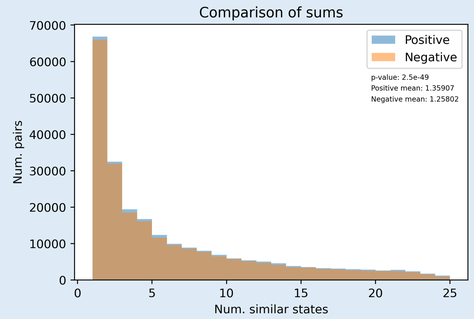

Fig 5. Comparison of training feature sums, p-value calculated using Mann-Whitney U-test

Discussion

- When comparing the chromatin state similarities between positive variant pairs and negative variant pairs, we found that there is a significant difference between positive and negative pairs. Sums for positive pairs averaged higher than sums for negative pairs
- Machine learning with logistic regression showed a recall curve with points slightly above the 'No Skill' line, which supports the model's performance being slightly above random
- The heatmap generated shows clustering of predicted positive pairs across the diagonal of chromatin states. This shows some indication that variants in similar states (for example, a pair of variants in the E5 state) are more likely to be predicted as a positive pair in our model. Future work may include analyzing if two variants in different states of the same group of chromatin markers also have a higher probability of being a positive pair
- About 55 million positive and negative pairs each were calculated; future works may include expanding the amount of data used in training and testing

Acknowledgements

We would like to thank Ha Vu for providing chromatin state data and additional mentorship on this project

Ernst, J., Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12,** 2478–2492 (2017). https://doi.org/10.1038/nprot.2017.124

Vu, H., Ernst, J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol* **23,** 9 (2022). https://doi.org/10.1186/s13059-021-02572-z