Comparing Variant Pathogenicity Scores to Understand Patterns in Scoring Methods via ScoreHMM

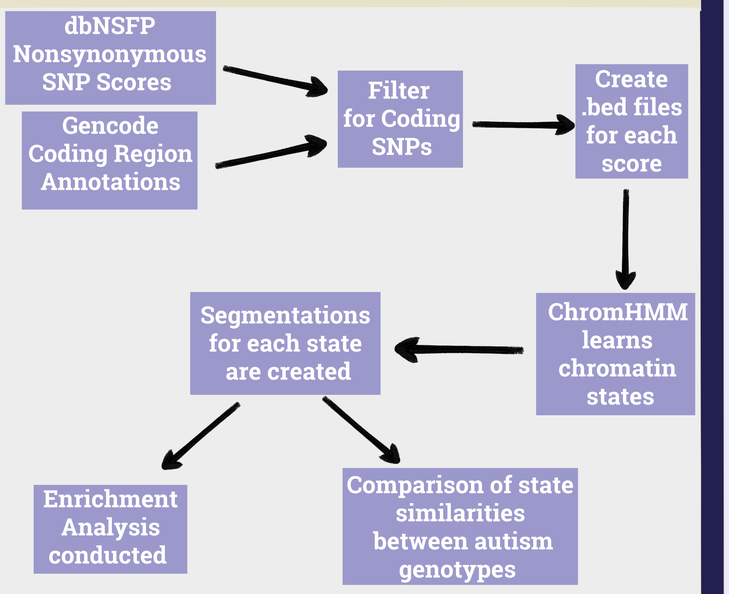
SIDDHARTH NAIDU[1,4], **Luke Li**[2,3], **Jason Ernst**[2,3]

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA. [2] Bioinformatics Interdepartmental Program, UCLA. [3] Department of Biological Chemistry, David Geffen School of Medicine, UCLA. [4] Department of Biology, Department of Computer Science, Duke University

Introduction

- A multitude of scoring methods have been developed to determine the pathogenicity of genetic variants, using structural, probabilistic, and evolutionary considerations to generate a score that quantifies the impact of variants on complex diseases.
- As different scoring methods may represent diverse aspects of the variant, we attempted to develop an approach that captures the combinatorial patterns of these scores.
- Here we present ScoreHMM, a hidden Markov model that takes genomic tracks of multiple scores as input and learns states that summarize their patterns.
- 51 scoring methods were used from dbNSFP, a database developed for prediction and annotation of nonsynonymous SNPs across the human genome.
- Different scoring distributions across the genome account for variation of which qualities are taken into account when scoring an variant.

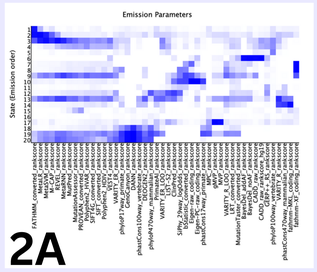
Pipeline

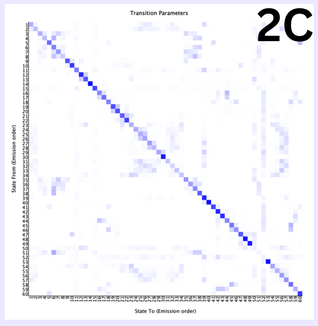


Improvements

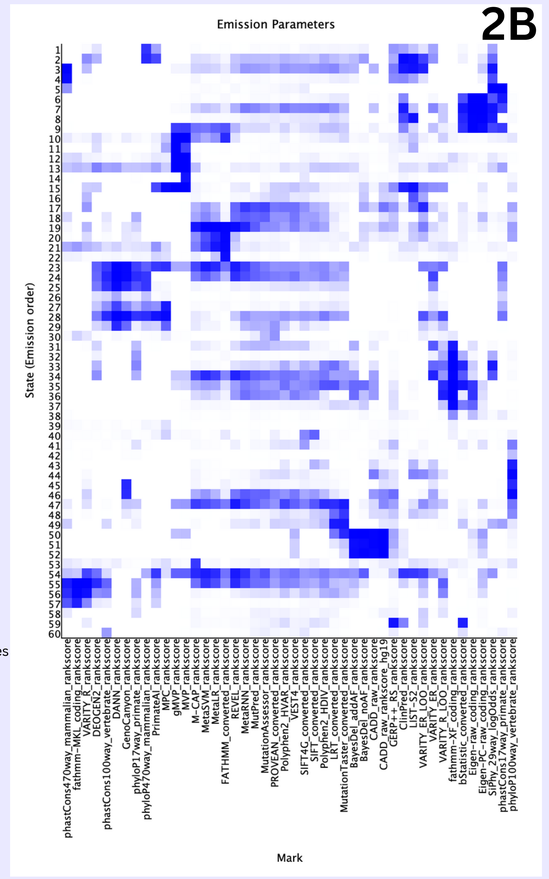
- Regions of analysis can be expanded to other chromosomes, noncoding regions, and other types of variants.
- Find better ways to binarize due to different distributions amongst scores; find ways to arbitrarily break ties within the top 10% of scores
- Segmentation can be used to study other conditions and diseases besides autism such as significant GWAS variants and other potential genetic disorder

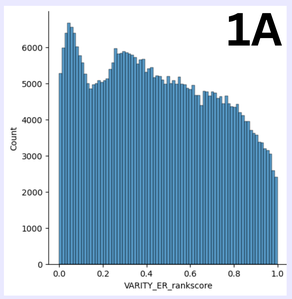
Figures

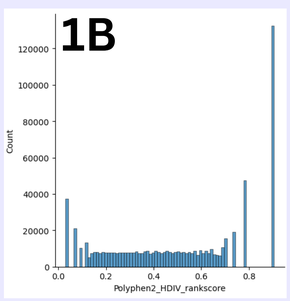




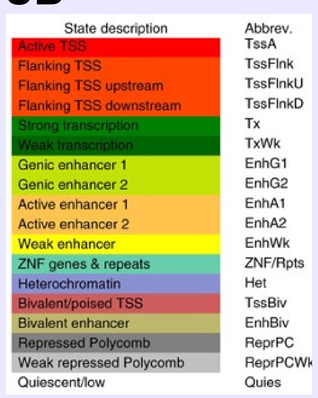
2A shows the emission graph for 20 states.
2B shows the emission graph for 60 states
2C shows the transition parameters for 60 states

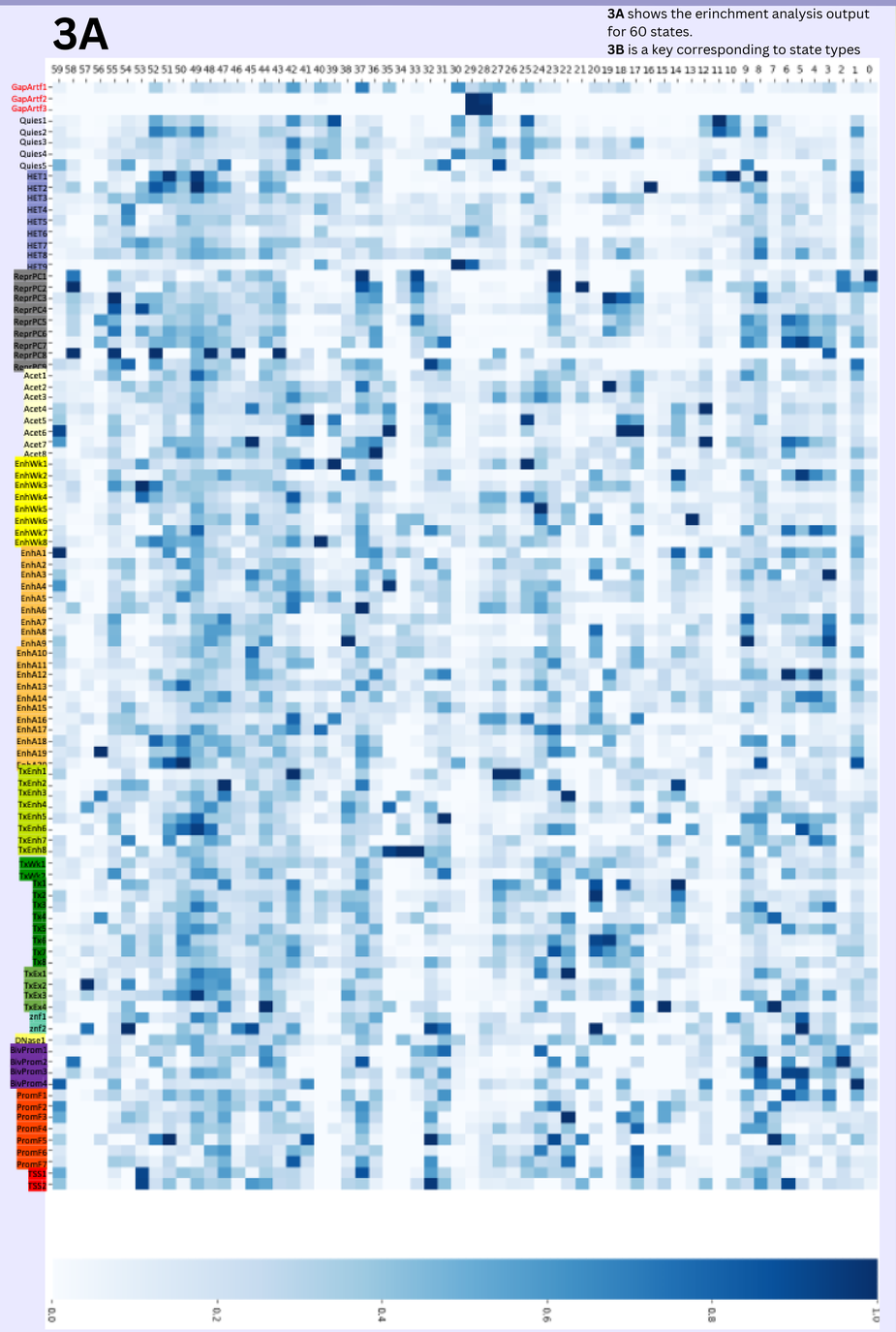






1A shows a skewed distribution of Polyphen2 scores.
1B shows a more even distribution of VARITY_ER scores





Discussion

- Clusters in the emission graph confer to sites in the genome that have similar intensity in pathogenicity based on qualities common amongst the related scores.
- The enrichment analysis shows some common states that are overexpressed across chromosome 21
- Future steps must be taken to find more commonalities among chromosomes and states.
- More enrichment analyses can be conducted with other disease phenotypes

Acknowledgements

The Ernst Lab
ChromHMM
Gencode
dbNSFP
BIG Summer Program