Sequence to Function

Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification
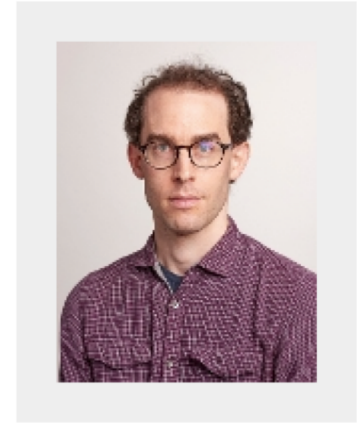
Bingbing

Nucleic Acids Research

Volume 47, Issue 20

18 November 2019

Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification 🔓

Gabriel E Hoffman ✉, Jaroslav Bendl, Kiran Girdhar, Eric E Schadt, Panos Roussos

Gabriel E. Hoffman

ASSOCIATE PROFESSOR | Genetics and Genomic Sciences

Icahn School of Medicine at Mount Sinai

Outline

Background

Model

-- DeepFIGV: CNN-based framework ->predict quantitative epigenetic variation

Results (Downstream analysis)

Discussion

Background

- Motivation: There is a widely recognized need for accurate computational predictions of the functional impact of variants in non-coding regulatory regulatory DNA
- Challenge: Linkage disequilibrium (LD) in GWAS and molecular trait QTL studies
- Dnase → chromatin accessibility   Chip-seq -> histone modifications
- SNP(Single nucleotide polymorphisms)

DeepFIGV

Workflow for DeepFIGV

-- Standard molecular trait QTL analyses rely on the correlation between the epigenetic signal and a given genetic variant

-- Deep learning using a convolutional neural network explicitly models the DNA sequence context to train a predictive model

-- Predicted score can integrated with other analyses to refine the mechanism of known functional variants, identify novel risk variants and prioritize downstream experiments.
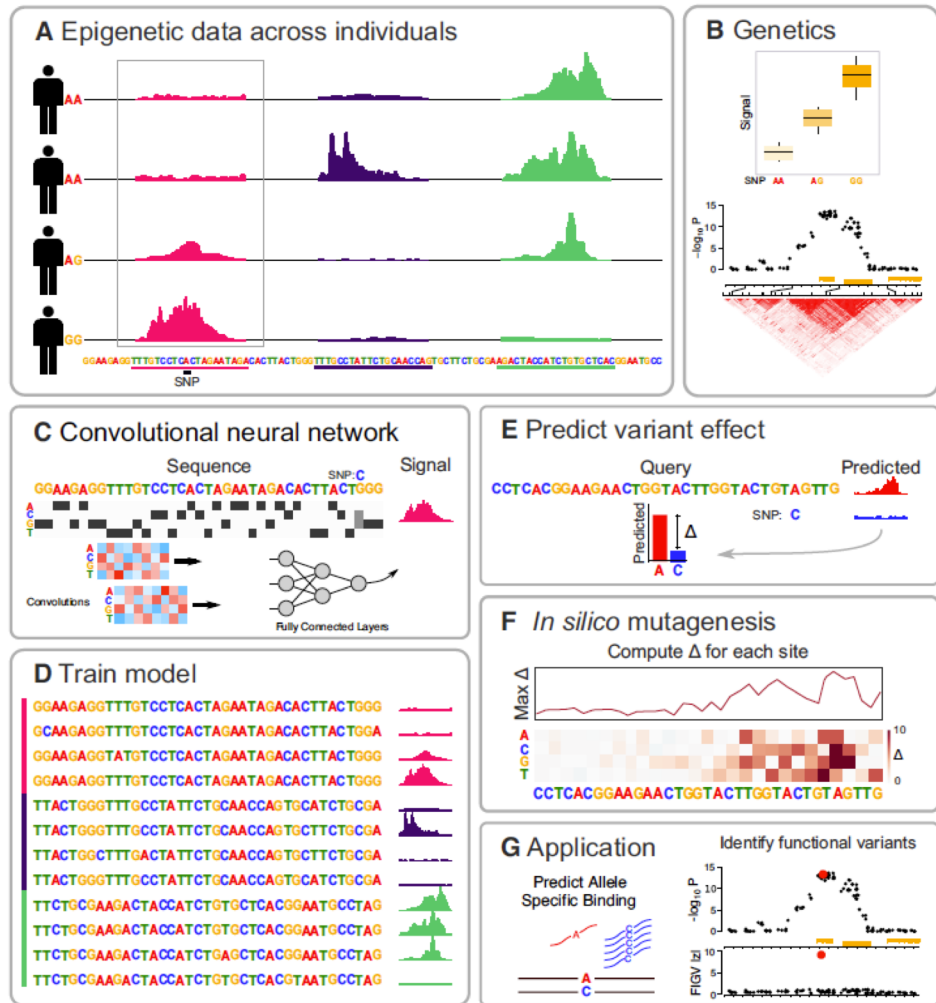
DeepFIGV

Model details

--The dataset comprises ChIP-seq experiments for 3 histone modifications (H3K27ac, H3K4me1 and H3K4me3) for 75 individuals and chromatin accessibility experiments for 69 individuals

Model construction:
--Genome sequence as input to epigenetic signal as output
--A 3 layer deep neural network and 300 convolutional filters each 19 bp wide. The model included rectified linear unit (ReLU) and max pooling.

Model training and testing
--A 30% dropout was applied to avoid overfitting. All training was performed on NVIDIA Tesla K20X GPU. Training on a single assay took between 14 and 45 GPU hours.
--Hyperparameters
--Evaluating all variants for the four assays took 5929GPU hours using 10 NVIDIA Tesla K20X GPUs. (438 million)
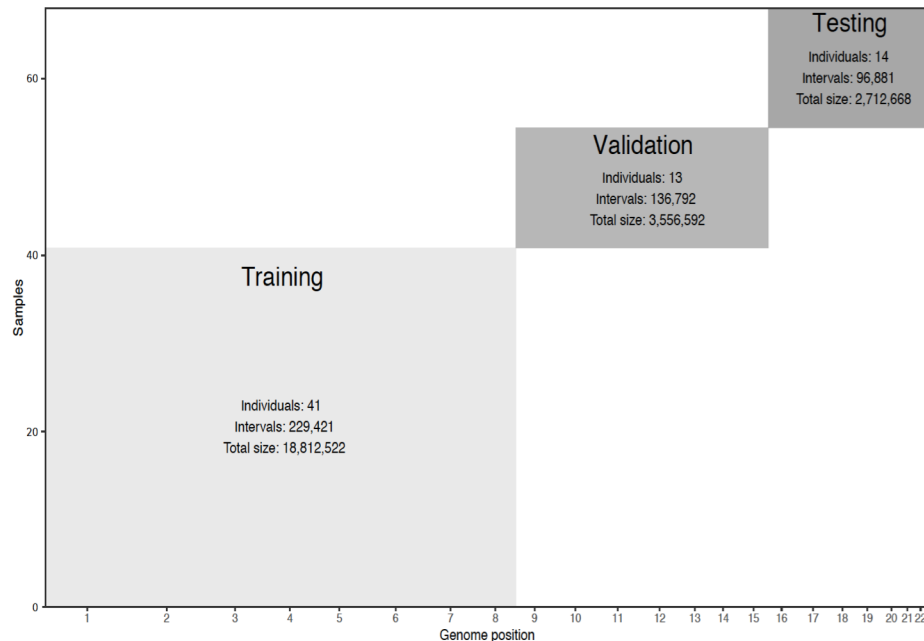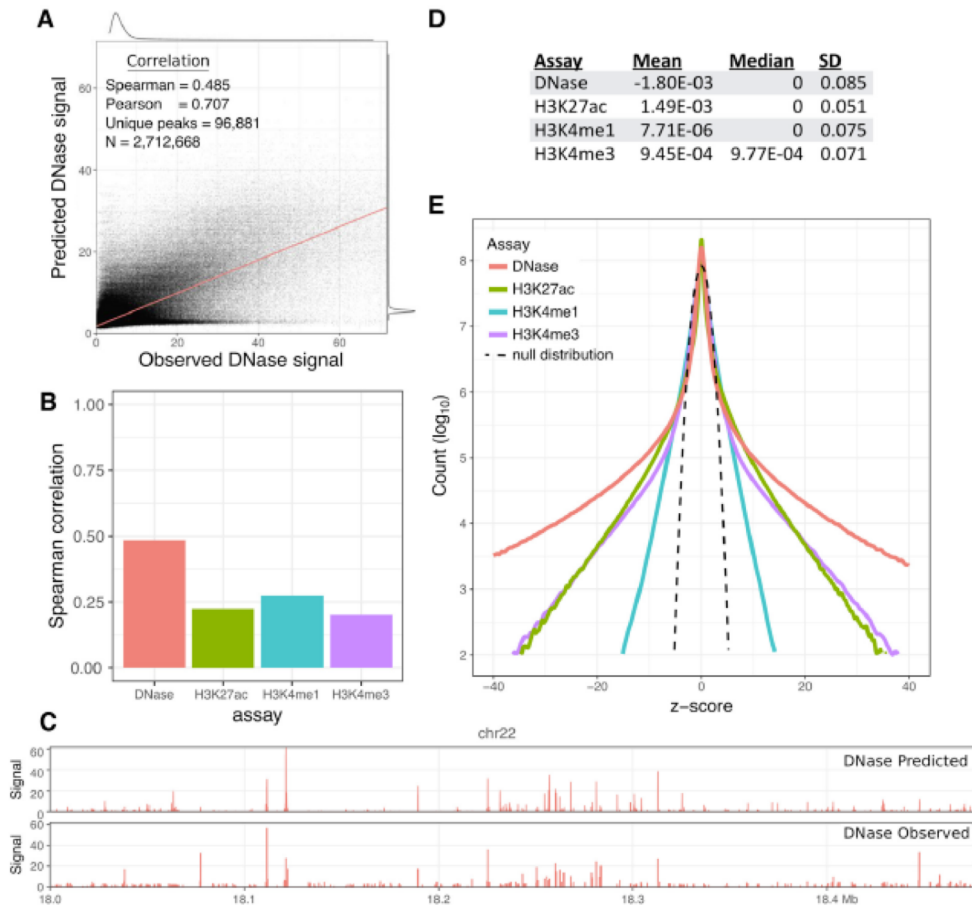
DeepFIGV

Evaluating DeepFIGV model and interpreting variant scores.

(A) The predicted DNase signal shows strong concordance in the test set with the observed signal

(B) Strong Spearman correlation between predicted and observed epigenetic signal on the test set for four assays.

(E) Transforming these delta values to a standard scale (i.e. z-score) by dividing by the standard deviation for each assay shows an excess of variants with scores near zero compared to the standard normal distribution.

Genomic correlates of predicted variant effects

- DeepFIGV recovers multiple aspects of known regulatory biology
- The predictive sequences features extracted by these filters are often similar to known transcription factor bindings site (TFBS) motifs
- TFBS enrichments are consistent with the biology of these assays: variants predicted to affect the open chromatin assay DNase are most enriched for TFBS motif
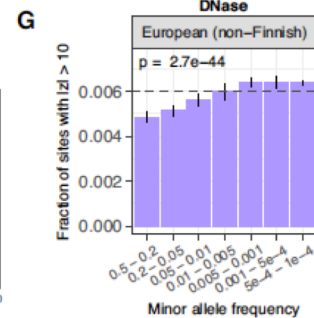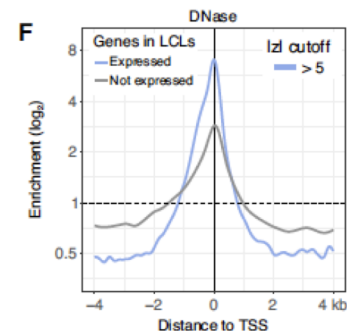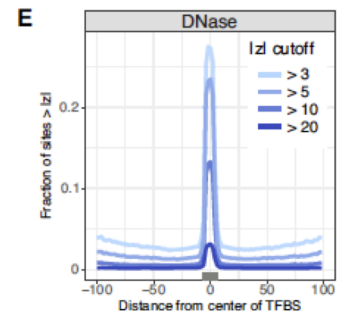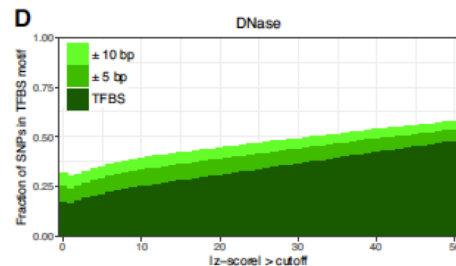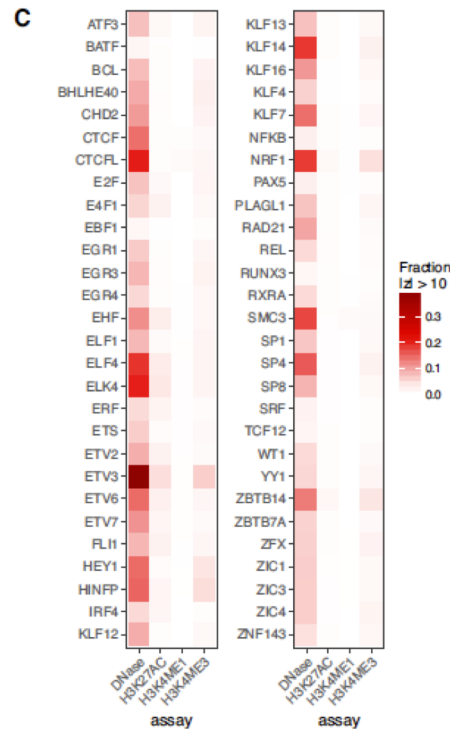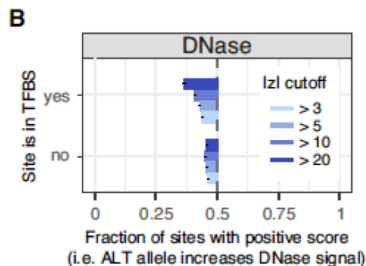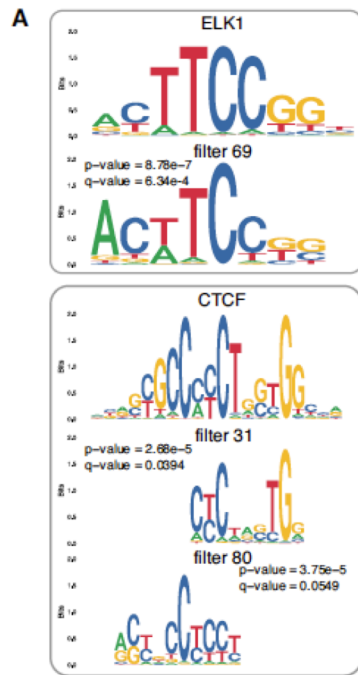
Concordance with molecular trait QTLs

Lead cis-QTL variants (i.e. the local variant with the smallest P-value) for multiple assays are enriched for having strong predicted effect on the epigenome
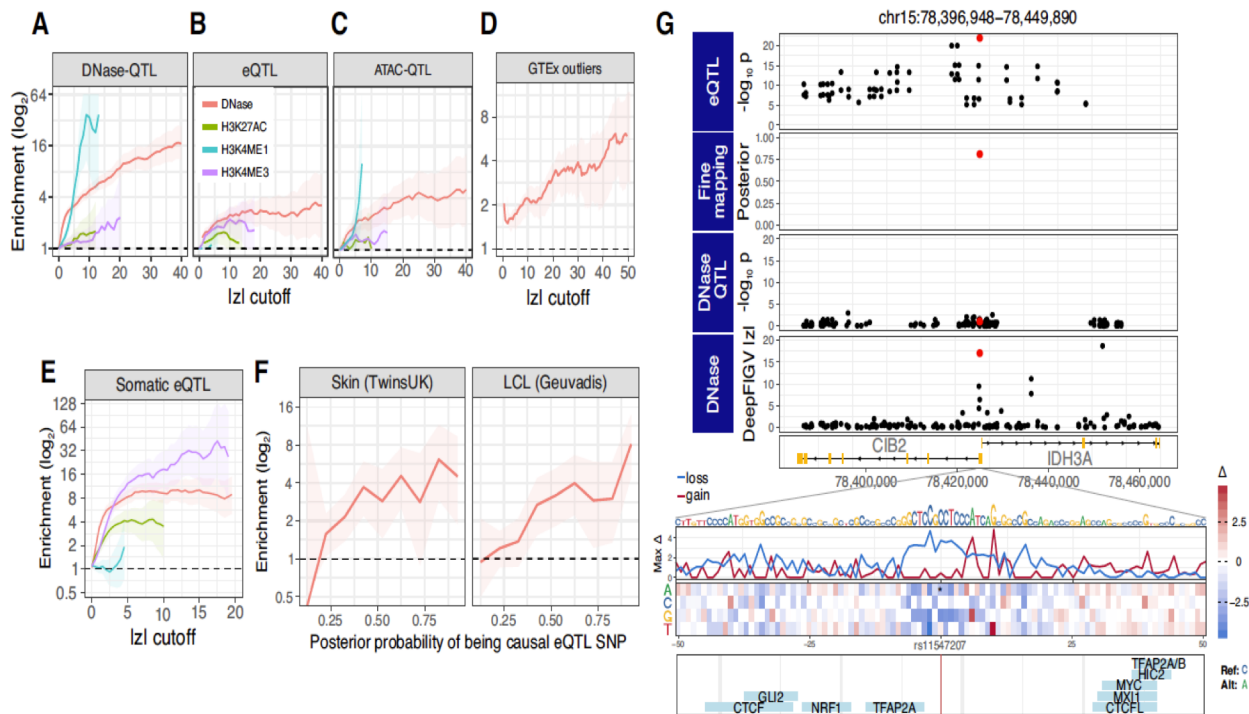
DeepFIGV variant scores predict allele specific binding

The predicted functional effect of genetic variants on each of the four epigenetic assays analyzed in DeepFIGV can identify allele-specific binding (ASB) of TFs in independent ChIP-seq experiments in LCLs.
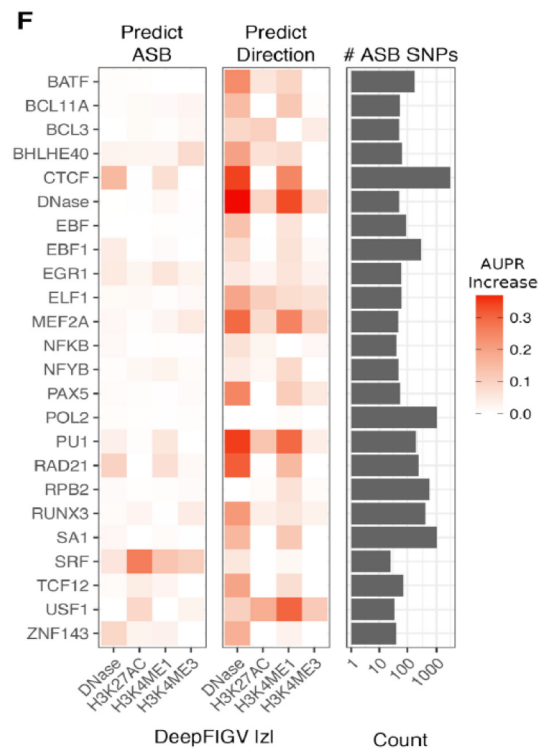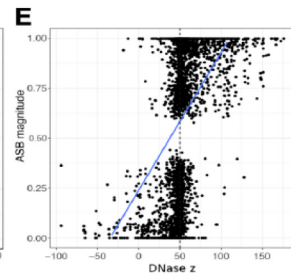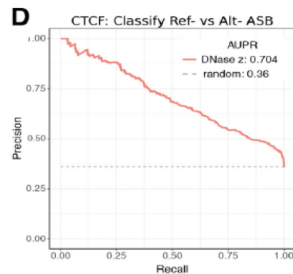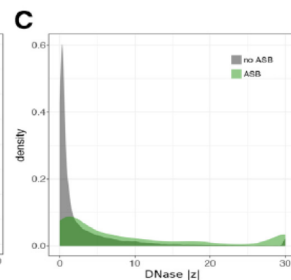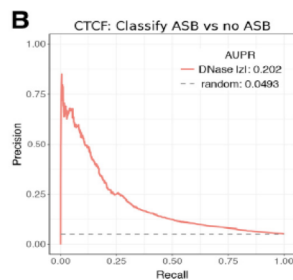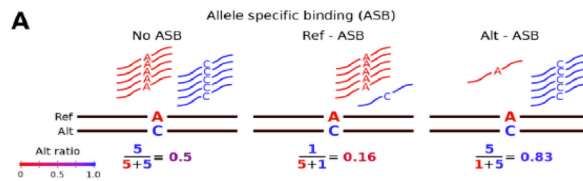
# Concordance with large-scale functional experiments of variant impact

(A) Variants that modulate gene expression in this assay are enriched for having large DeepFIGV z-scores compared to variants in sequences that do not drive expression.

(B) (B) Variants that modulate gene expression in this assay are enriched for having large DeepFIGV z-scores for DNase compared to variants that have no allelic effect in this experiment.
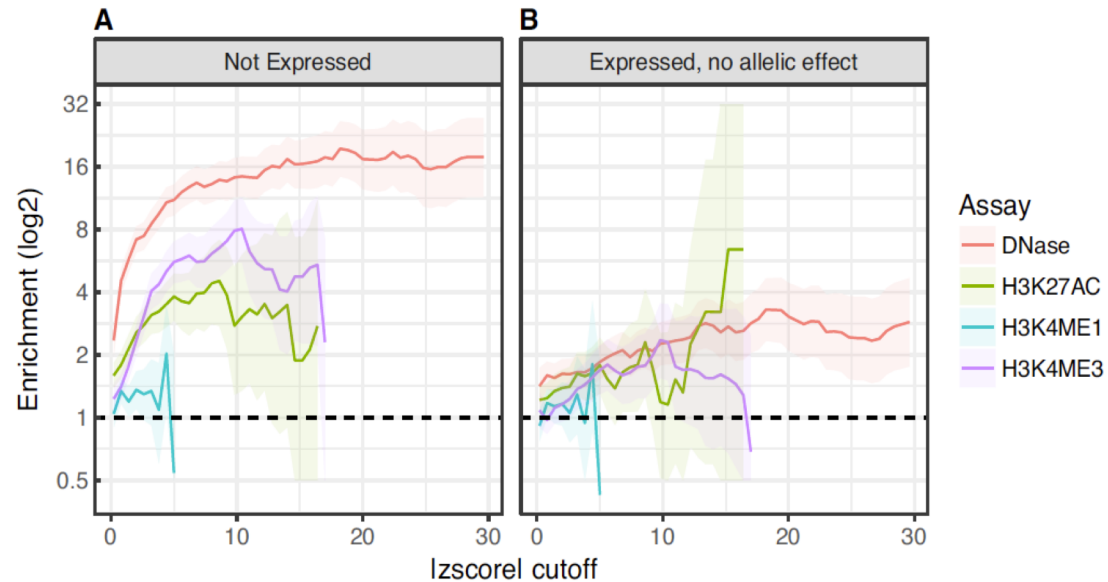
Enrichment for disease risk variants and interpreting causal variants

Integrating DeepFIGV scores with large-scale genome wide association studies shows that risk variants for common disease are enriched for variants predicted to impact the epigenome.
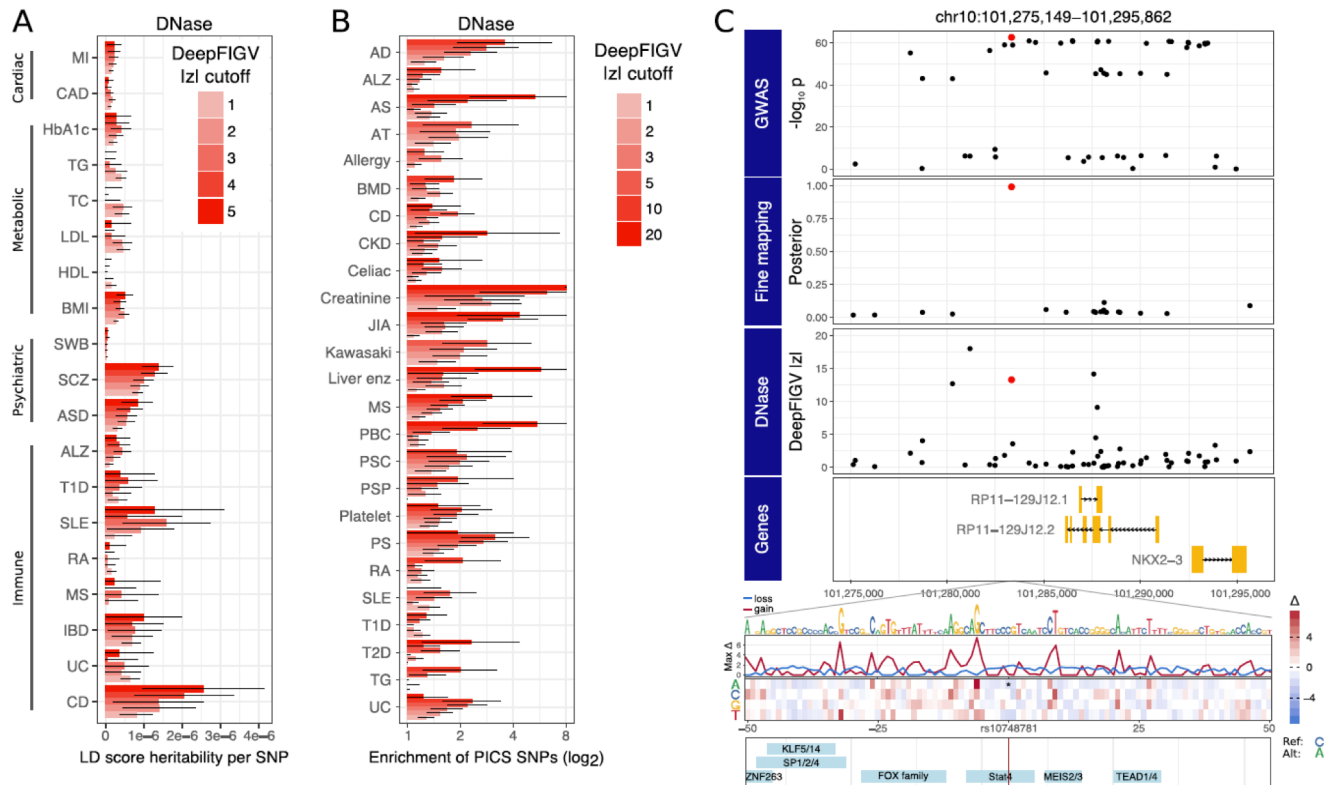
Some take-aways

we introduce a deep learning framework for functional interpretation of genetic variants (DeepFIGV).
(i) performing model training on many epigenetic experiments for a particular assay and cell type instead just a single representative sample.
(ii) integrating whole genome sequencing to create a personalized genome sequence for each individual.
(iii) modeling quantitative variation in the epigenetic readout rather than dividing the genome into two classes.

Discussion

-- CNN-> transformer?

-- Any confounding factors when calculating the correlation?