Summer Review 7

Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk

Jian Zhou[1,2,3], Chandra L. Theesfeld[1], Kevin Yao[3], Kathleen M. Chen[3], Aaron K. Wong[3] Olga G. Troyanskaya[1,3,4]

Nature Genetics

Paper Link

Reviewed by : Arshdeep Sekhon

[1]Department of Computer Science, University of Virginia
https://qdata.github.io/deep2Read/

Motivation

The task

Sequence $\rightarrow$ Chromatin features $\rightarrow$ expression

- DNA sequence to transcriptional effect of mutations
- precision medicine and evolutionary biology
- what controls gene expression?
- how genome variations(both gene and non coding regions) affect gene transcription?
- very large space of the non coding mutation space

Validation

- four immune related diseases: causal variants for disease and traits
- in silico saturation mutagenesis, for > 140 million promoter proximal mutations
- Uses: evolutionary constraints on gene expression
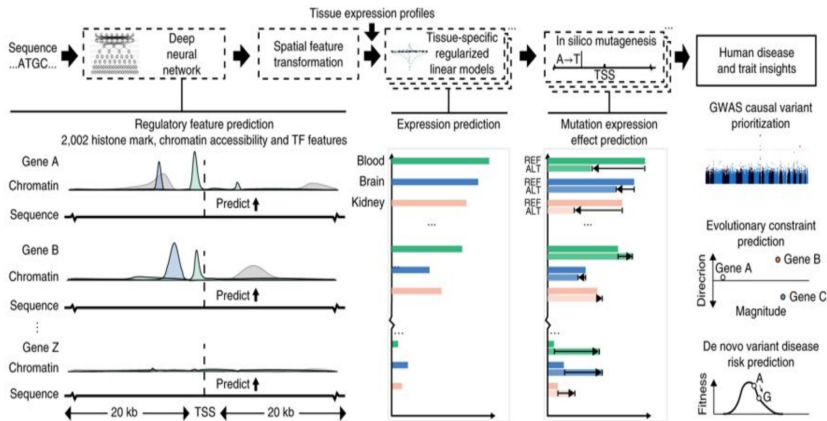- uses: ab initio prediction of disease risk

Expecto

Figure: The pipeline

- Modular framework
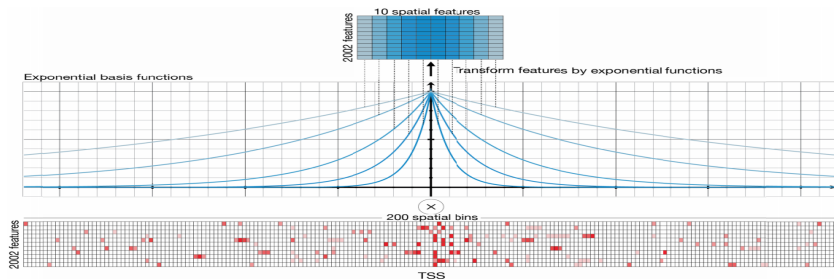- CNN followed by generalized linear model

Model (a)

Figure: CNN followed by spatial transofrmation

- generate probabilities for 2002 HMs/TFs/DNAse per window for $+/-$ 20 kbp
- a moving window with a 200-bp step size
- 200 spatial bins with a total number of 400,400 features.
- Deeper Model from DeepSEA

- reduced the input dimensionality with ten exponential functions
- weights based on relative distance to the TSS (transformed features with a higher decay rate were more concentrated near TSSs).
- reduced the number of features 20-fold to 20,020.
- The exponential functions were prespecified
- log (RPKM) value is target

$$\text{expression} = \sum_{d \in D} \sum_{i} p_{id} \left[\sum_{k} 1(t_d < 0) \beta_{ik}^{\text{up}} e^{-a_k \frac{|t_d|}{200\text{bp}}} + \sum_{k} 1(t_d > 0) \beta_{ik}^{\text{down}} e^{-a_k \frac{|t_d|}{200\text{bp}}} \right]$$

Figure: gene expression computation

- p_{id} is the predicted probabilities for chromatin feature i at region d relative to the TSS
- D represents the set of $200bp \times 200bp$ spatial bins within 20 kb of the TSS
- I represents the indicator function, which equals 1 when the specified condition is satisfied and 0 otherwise.
- t_d represents the mean distance of region d to the TSS.
- β are learned expression linear models: are shared across spatial bins indexed by d due to spatial transformation
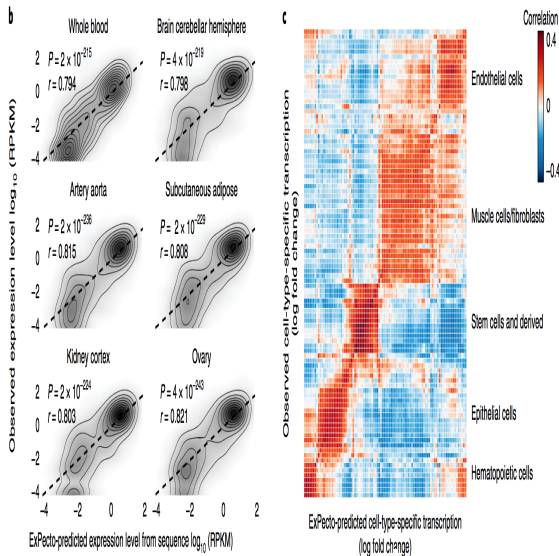
Effect of genomic variants on tissue-specific expression

- prioritizing causal eQTL variants: as it was unconfounded by linkage disequilibrium (LD).
- Even though majority of eQTL are non causal, expected the strong ExPecto-predicted effect variants to be highly enriched in bona fide causal variants
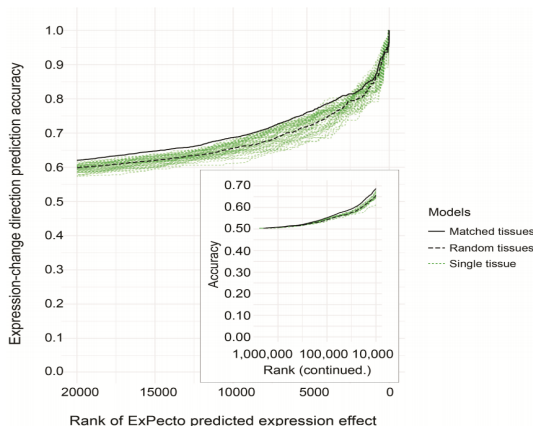
Prioritizing and experimental study of causal GWAS variants

- in silico mutagenesis
- GWAS : identifies a multitude of associated loci
- lacks the resolution to pinpoint causal genomic variants, largely owing to LD.
- loci having variants with stronger predicted effects were significantly more likely to be replicated in a different GWAS [2]

---

[2] GWAS uncover disease-associated loci, but due to sparse genotyping arrays and linkage disequilibrium (LD), identifying the specific SNP driving the association is difficult. Therefore, GWAS usually report the most significant hit as the single lead SNP for a loci, leaving the identification of a causal SNP for later research. Often multiple GWAS of the same disease will identify different lead SNPs in the same region, presumable all tagging the same causal variant. Therefore, around any lead SNP is a region of indeterminationa genomic window in which the SNP driving the association is likely to reside.

Prioritizing and experimental study of causal GWAS variants

- potential of using predicted expression effect information to improve identification of causal associated loci from GWAS.
- experimentally validated GWAS vs Expecto predicted variants for immune diseases:measured the expression alteration effects of the top three ExPecto-prioritized SNPs and compared their allele-specific regulatory potential to that of the lead SNPs from the corresponding GWAS
- LD SNPs prioritized by expression effect, although having no prior evidence of functionality, showed transcriptional regulatory activity, whereas lead GWAS SNPs did not.

Variation potentials and evolutionary constraints of genes

- in silico mutagenesis: systematically predicted all ($>$140 million) possible singlenucleotide substitution variations across all human promoters within 1 kb of the representative TSS on both sides
- Variation Potential
    - directionality : sum of predicted log(fold change) values for all mutations per gene,
    - and magnitude : sum of all absolute predicted log(fold change) values.

Variation potentials and evolutionary constraints of genes

- negative variation potential directionality (i.e., mutations that tended to cause a decrease in tissue-specific expression) were actively expressed in the modeled tissue
- inferred that these genes were under positive evolutionary constraint and thus were vulnerable to inactivating mutations.
- directionality score to measure the tendency of the potential mutation effect to be biased toward positive or negative : propose this indicates negative and positive evolutionary constraints, respectively
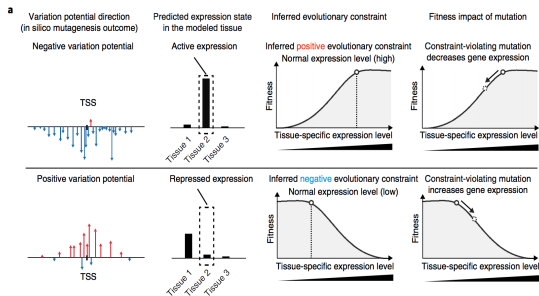
Figure: Gene expression specificity and activation status can be predicted from the magnitude and directionality of gene variation potential. The position of each gene set was computed as the average cumulative mutation effects (directionality) and average cumulative absolute mutation effects (magnitude) across all genes in the set. Each gene set is colored by the directionality of variation potential.
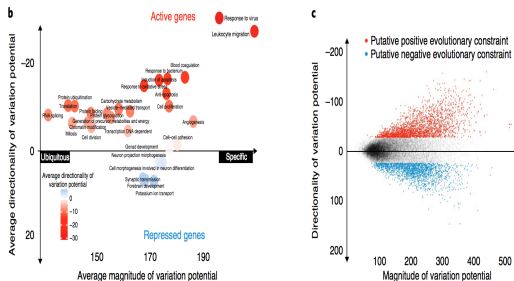
Figure: Inference of genes with putative directional evolutionary constraints from variation potentials. Each dot represents a gene. The x and y axes show the cumulative predicted mutation effects and the log(fold change) values of mutations with positive and negative impact within 1kb of the TSS, respectively.

Conclusion

- provides robust and scalable ab initio, sequence-based prediction of variant effects, enabling genome-wide studies of human genomic variation and disease.
- demonstrated that computational prediction of causal variants in trait-associated loci, including eQTLs and GWAS disease-associated loci, is capable of identifying causal variants and that this can be routinely performed at a whole-genome level.
- also make possible the probing of variation potentials and evolutionary constraints through in silico mutagenesis analysis

NOTES

DeepSEA vs Expecto

- ExPecto predicts variant effects on gene expression
- DeepSEA can identify variant effects that do not lead to significant changes in expression
- wider sequence
- more number of TFs/etc
- deeper architecture

Linkage disequibrium

linkage disequilibrium is the non-random association of alleles at different loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly.

Effect of genomic variants on *tissue* specific expression

- generate in silico mutagenesis variants
- Expecto predicted effect variants
- Rank by maximum effect
- Accuracy with directionality effect of prediction
- Lower rank likely to be less causal: low expression
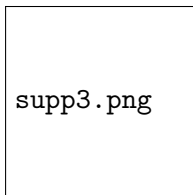- Expecto unconfounded by LD

supp3.png

Figure: Caption

Causal Variants vs Associated variants

In the context of association studies, the genetic variants which are responsible for the association signal at a locus are referred to in the genetics literature as the causal variants. Causal variants have biological effect on the phenotype. Generally, variants can be categorized into three main groups. The first group is the causal variants which have a biological effect on the phenotype and are responsible for the association signal. The second group is the variants which are statistically associated with the phenotype due to LD with a causal variant. Even though association tests for these variants may be statistically significant, under our definition, they are not causal variants. The third group is the variants which are not statistically associated with the phenotype and are not causal.