Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou & Olga G Troyanskaya

Presented by Michaela Ennis, 4/6/2017

Computational Systems Biology
Deep Learning in the Life Sciences

Massachusetts Institute of Technology

http://mit6874.github.io

Vision

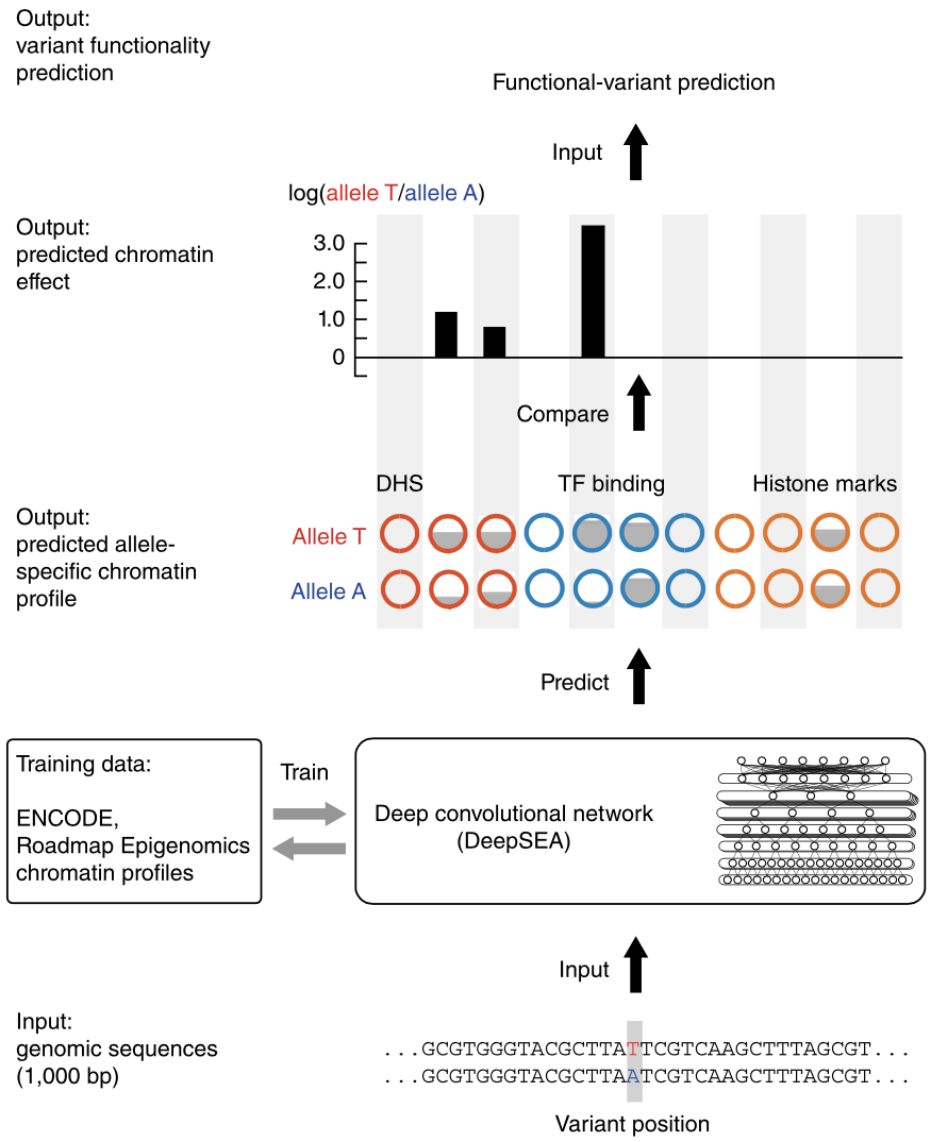
- Genome Wide Association Studies for most diseases have found many SNPs, mostly non-coding
  - How do we determine the functional significance of these SNPs?
- Existing work had used information such as chromatin marks and evolutionary conservation
  - But there wasn’t yet a good approach to predict based purely on genomic sequence
  - Evaluating just motifs misses capturing complex interactions

Overview

- Key Claim
  - ***DeepSEA accurately predicts binding of chromatin proteins and histone marks with high nucleotide resolution, using only DNA sequence as an input***
- Importance
  - ***Ability to obtain quick predictions about the functional role of non-coding SNPs can help to guide GWAS follow ups***
- Issues
  - ***More work needs to be done for using this technique to screen a large number of non-coding SNPs in order to gain higher level insight into a particular disease***

The Model

- DeepSEA predicts chromatin features for an input sequence
- This can be used in a further pipeline to assess functional roles of variants



Data Used

- Split genome into 200 bp bins, for each bin compute label for all 919 chromatin features (1 if more than half of bin is in peak region)
  - ***Used bins only w/ at least 1 TF binding event, so 17% of genome total***
  - ***Input was 1000 bp chunks (one hot encoding), w/ label corresponding to chromatin feature for 200 bp center***
  - ***Chromosomes 8 and 9 were set aside, w/ ~5MB set aside for hyper parameter validation, and the rest for testing***
  - ***Data from Roadmap and ENCODE databases***
- To evaluate variants, used non-coding data from HGMD, GRASP eQTL, and the GWAS catalog
  - ***Control against 1000 Genomes***
  - ***Include evolutionary conservation data in addition to network output***

CNN

Model Architecture:

1. Convolution layer (320 kernels. Window size: 8. Step size: 1.)
2. Pooling layer (Window size: 4. Step size: 4.)
3. Convolution layer (480 kernels. Window size: 8. Step size: 1.)
4. Pooling layer (Window size: 4. Step size: 4.)
5. Convolution layer (960 kernels. Window size: 8. Step size: 1.)
6. Fully connected layer (925 neurons)
7. Sigmoid output layer

Regularization Parameters:

Dropout proportion (proportion of outputs randomly set to 0):

- Layer 2: 20%
- Layer 4: 20%
- Layer 5: 50%
- All other layers: 0%

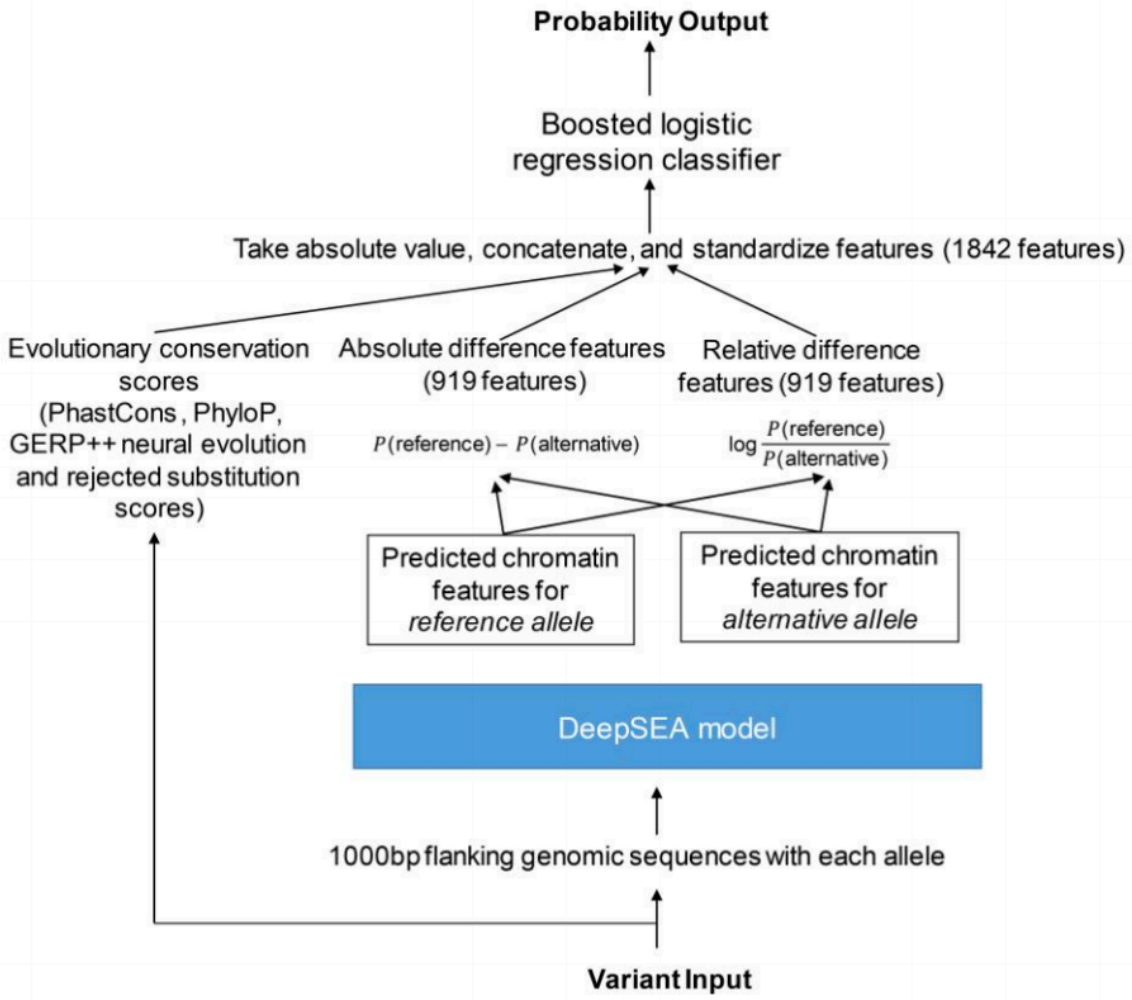
L2 regularization (λ_1): 5e-07

L1 sparsity (λ_2): 1e-08

Max kernel norm (λ_3): 0.9

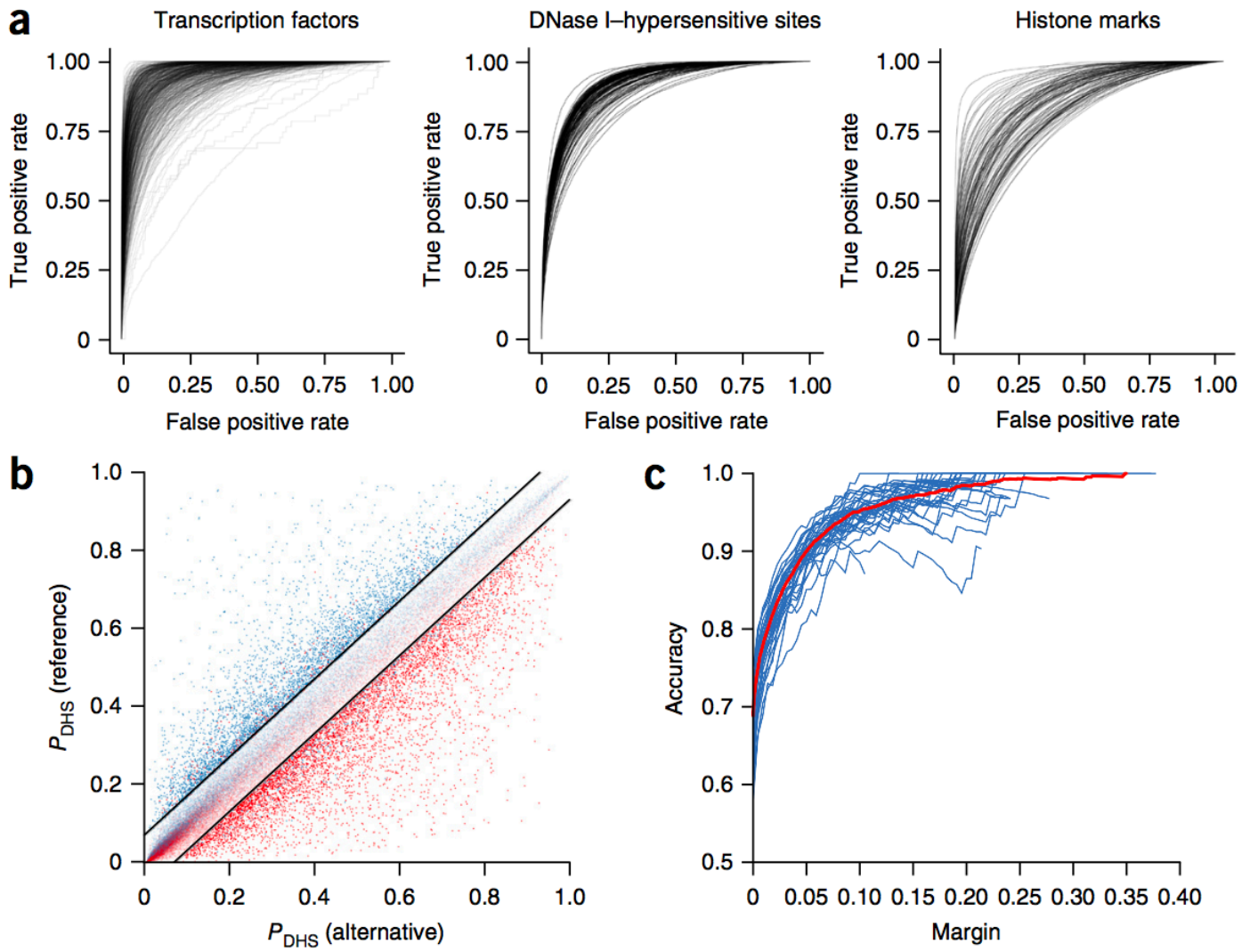
Functional Significance Score

- Input a sequence and a variant to get predicted differences in chromatin marks
- Use output of DeepSEA plus evolutionary conservation scores as input to a classifier that gives probability of functional significance



Results

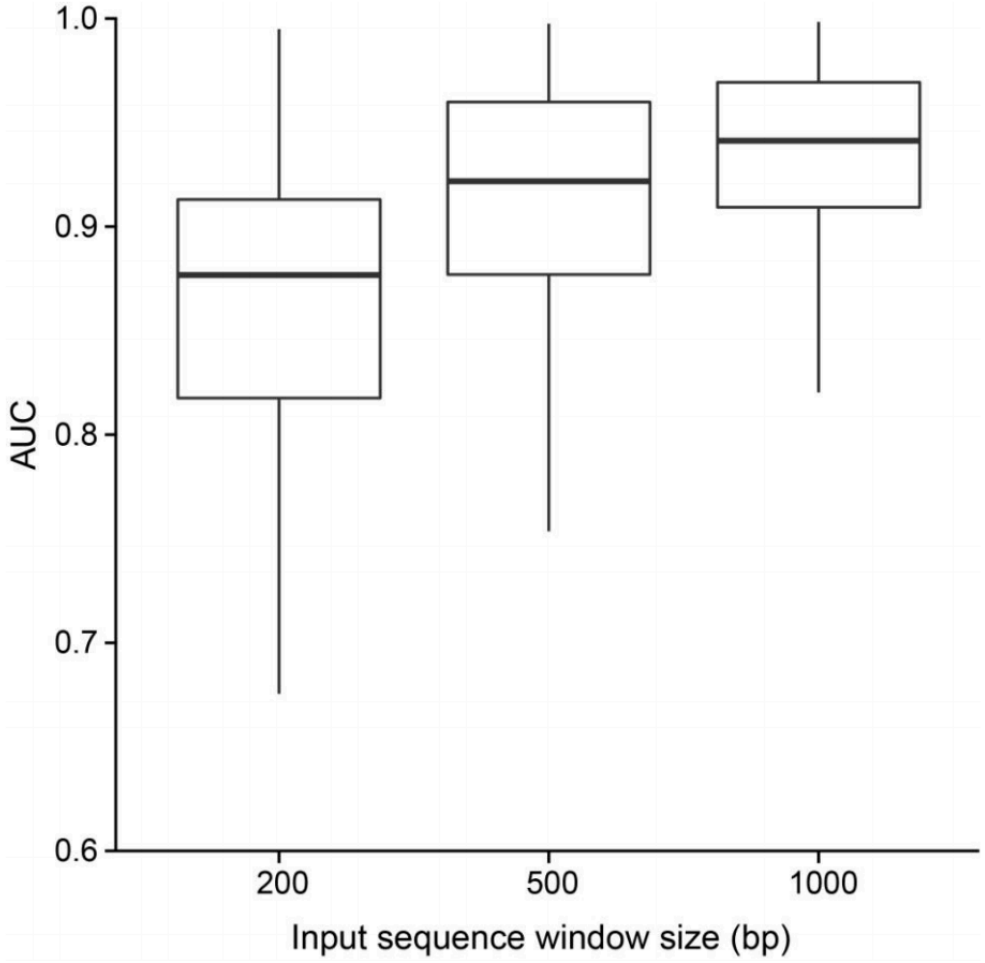
- DeepSEA predicted TF binding sites with high accuracy, with median area under the curve (AUC) of 0.958
  - The previous best was an SVM architecture that had median AUC = 0.896)
- DeepSEA also predicted DHSs and histone modifications, with median AUC of 0.923 and 0.856 respectively
- DeepSEA was then used to further evaluate functional significance of non-coding SNPs, with results better than previous work


a
Transcription factors
DNase I–hypersensitive sites
Histone marks
True positive rate
False positive rate
0
0.25
0.50
0.75
1.00
b
P_{DHS} (reference)
P_{DHS} (alternative)
0
0.2
0.4
0.6
0.8
1.0
c
Accuracy
Margin
0.5
0.6
0.7
0.8
0.9
1.0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

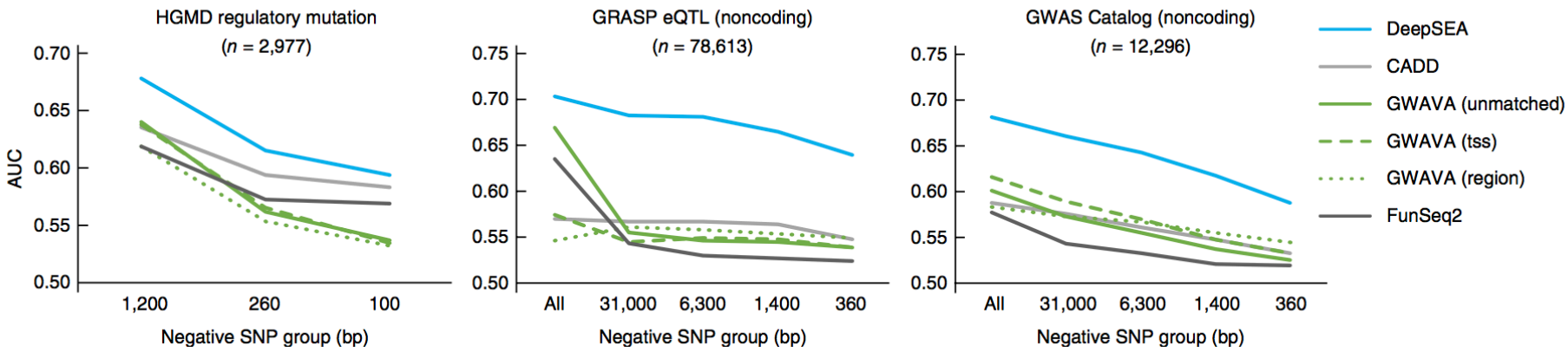
Results

A larger input window improves results when predicting chromatin features for the center 200 bp

1000 bp input window was used for the rest of the results



Results

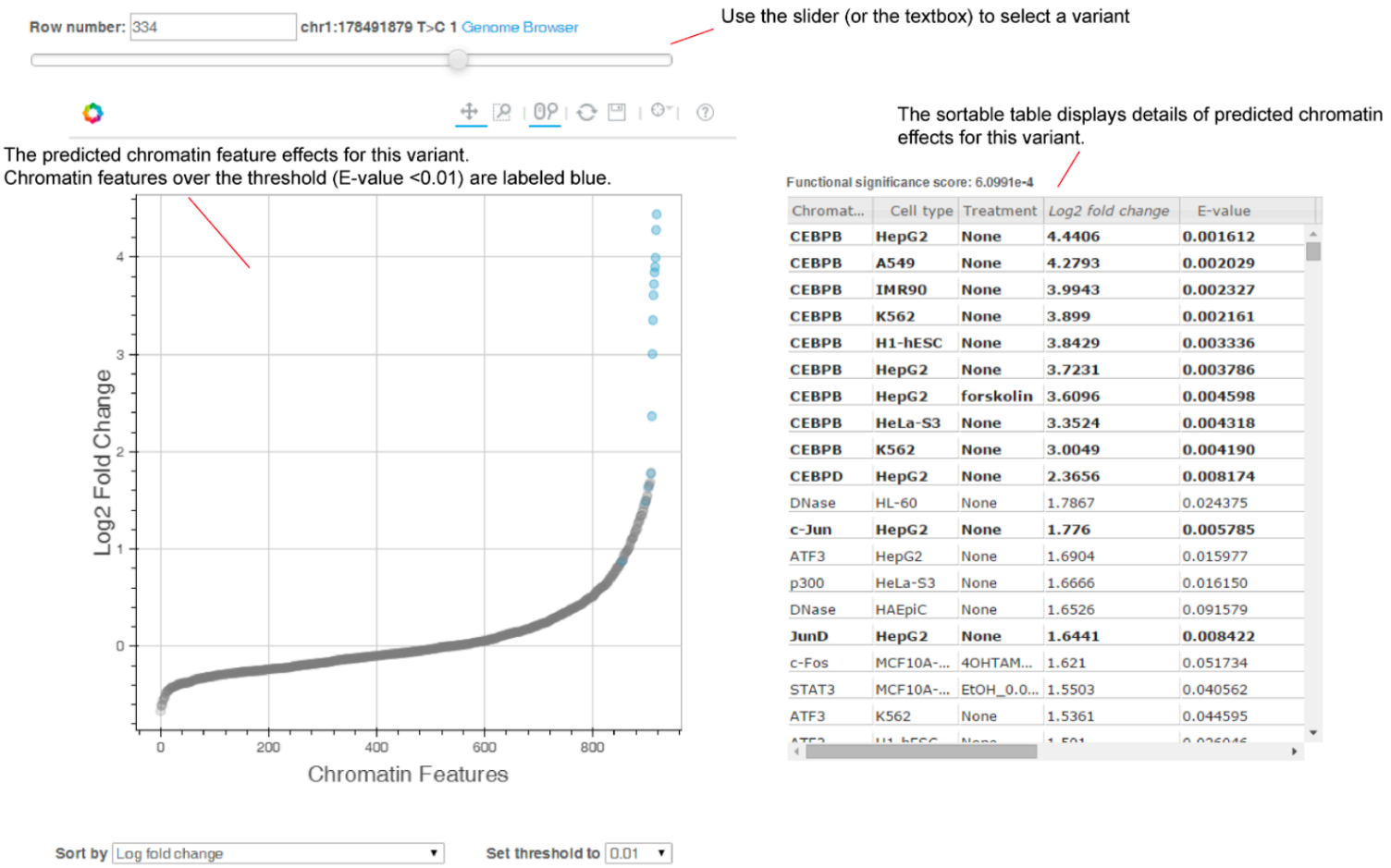


Comparing functional significance predictions using DeepSEA's results with previous work

In particular, the DeepSEA method is able to give decent results even when the negative sample SNPs are chosen very close to the positive sample SNPs

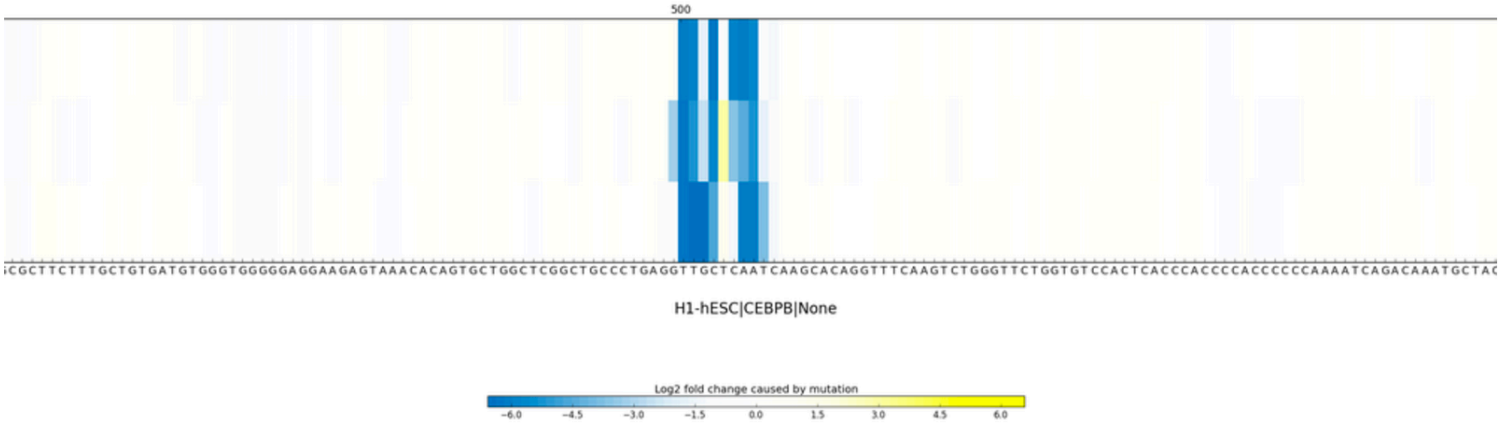
Evolutionary conservation does a better job predicting HGMD mutations, which tend to be more explicitly deleterious, and therefore should undergo a higher selective pressure

Online Tool



DeepSEA can give results for a particular 1000 bp input sequence, or it can give comparative results for a particular variant- the variant can even be a (small) insertion or deletion, not just SNPs!

In Silico Mutagenesis



If you've pinpointed a particular chromatin feature that you would like to evaluate for your sequence, you can use this simulation tool, which runs DeepSEA on all 3000 possible individual SNP variants, and notes which mutation sites caused the greatest change for that feature

Key Claims

- DeepSEA can accurately predict TF binding sites, DNAse hypersensitivity sites, and histone marks, with high nucleotide resolution, using only DNA sequence as an input
- DeepSEA can be used in a further pipeline to understand functional significance of non-coding SNPs
  - Compare variants to find chromatin features most effected
  - Determine functional risk score for large number of non-coding variants
  - Simulate mutagenesis on a particular sequence to find sites that will most effect a particular chromatin feature

Analysis

- The results where DeepSEA predicts chromatin features for a given input sequence I thought were pretty impressive
- I would have liked to see more work on the functional significance side of things
  - The AUCs didn't seem particularly significant to me, although an improvement
  - Not training on a particular disease- how do we even know what is more 'functional'
  - Not as much cell type information in the original training data as I would have liked
- Perhaps beyond the scope of this, but it would be cool to see something like % of epigenetic variance explained by genetic differences for a disease like Schizophrenia
  - How does environmental component factor in?

Impact

- Ability to obtain quick predictions about the functional role of non-coding SNPs can help to guide GWAS follow ups
- In particular, DeepSEA is available not only as code, but also as an online tool that researchers in a variety of fields can take advantage of
- Makes it accessible for follow up computational studies as well!