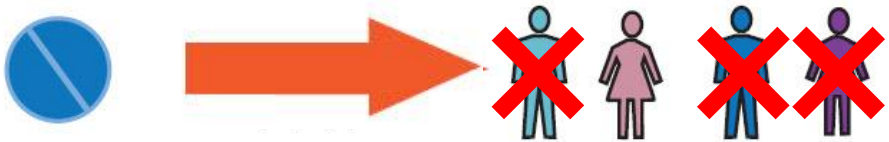
Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

David R. Kelley,[1] Jasper Snoek,[2] and John L. Rinn[1]

[1]*Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA;*
[2]*School of Engineering and Applied Science, Harvard University, Cambridge, Massachusetts 02138, USA*

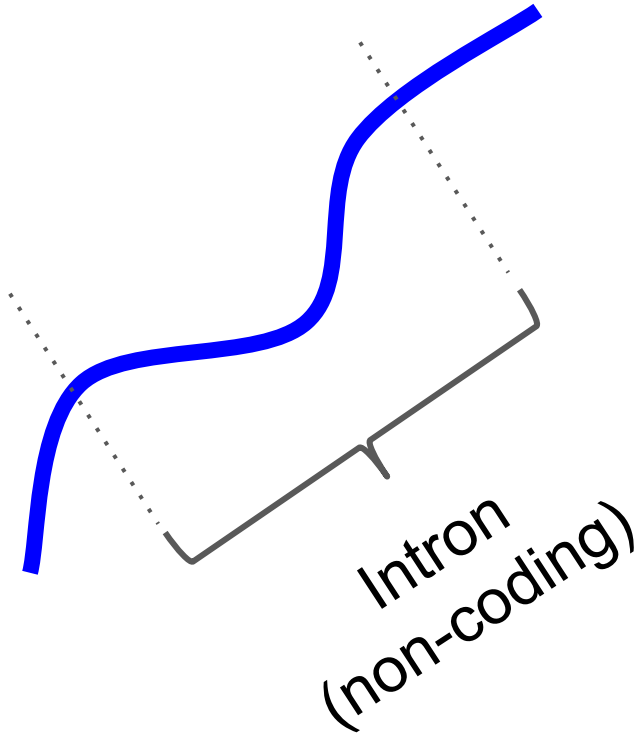
CS273B Presentation
10/19/2016
Amr Mohamed, Wisam Reid, Irán Román

Towards Personalized Medicine



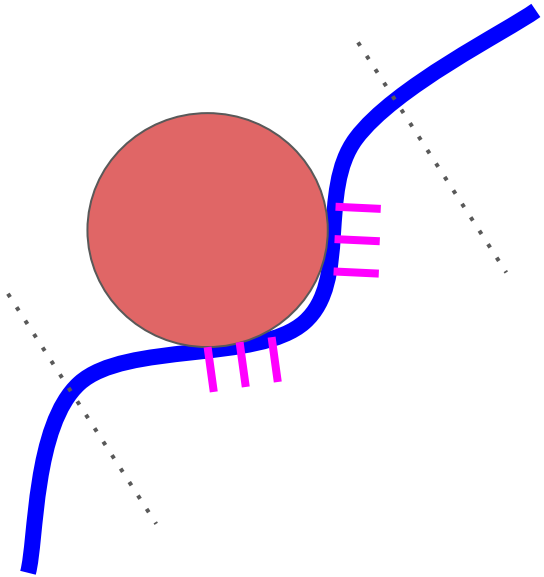
http://medicine.iupui.edu/IIPM

Non-coding DNA and personalized Medicine



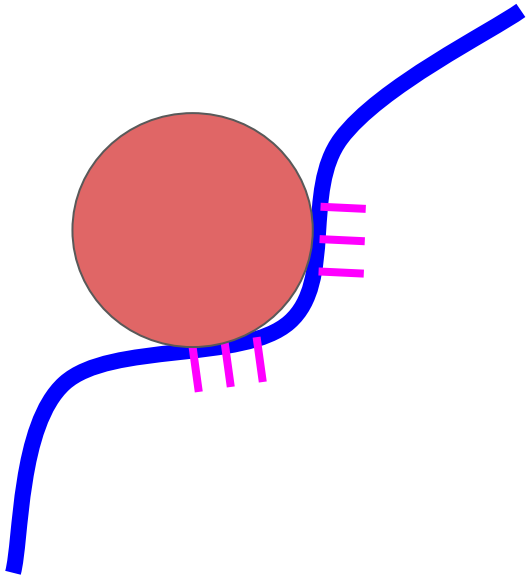
Can we relate non-coding DNA with phenotypes?

Non-coding DNA and personalized Medicine



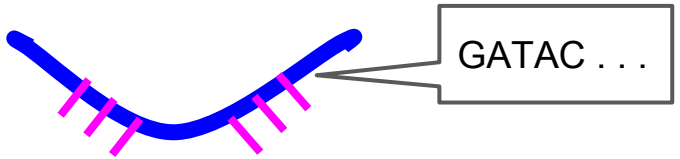
Large surveys indicate that these modifications are statistically related to phenotypes (ENCONDE, 2012)

Taking full advantage of these annotations

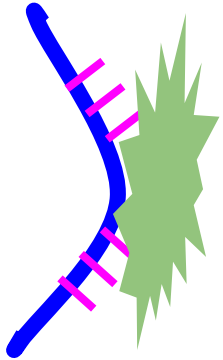


Two possibilities:

1)

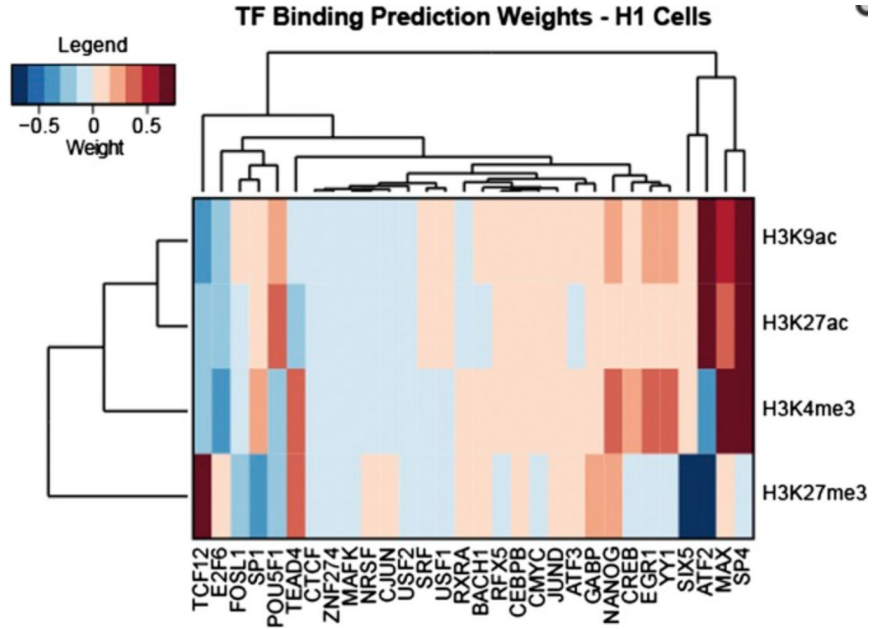


2)



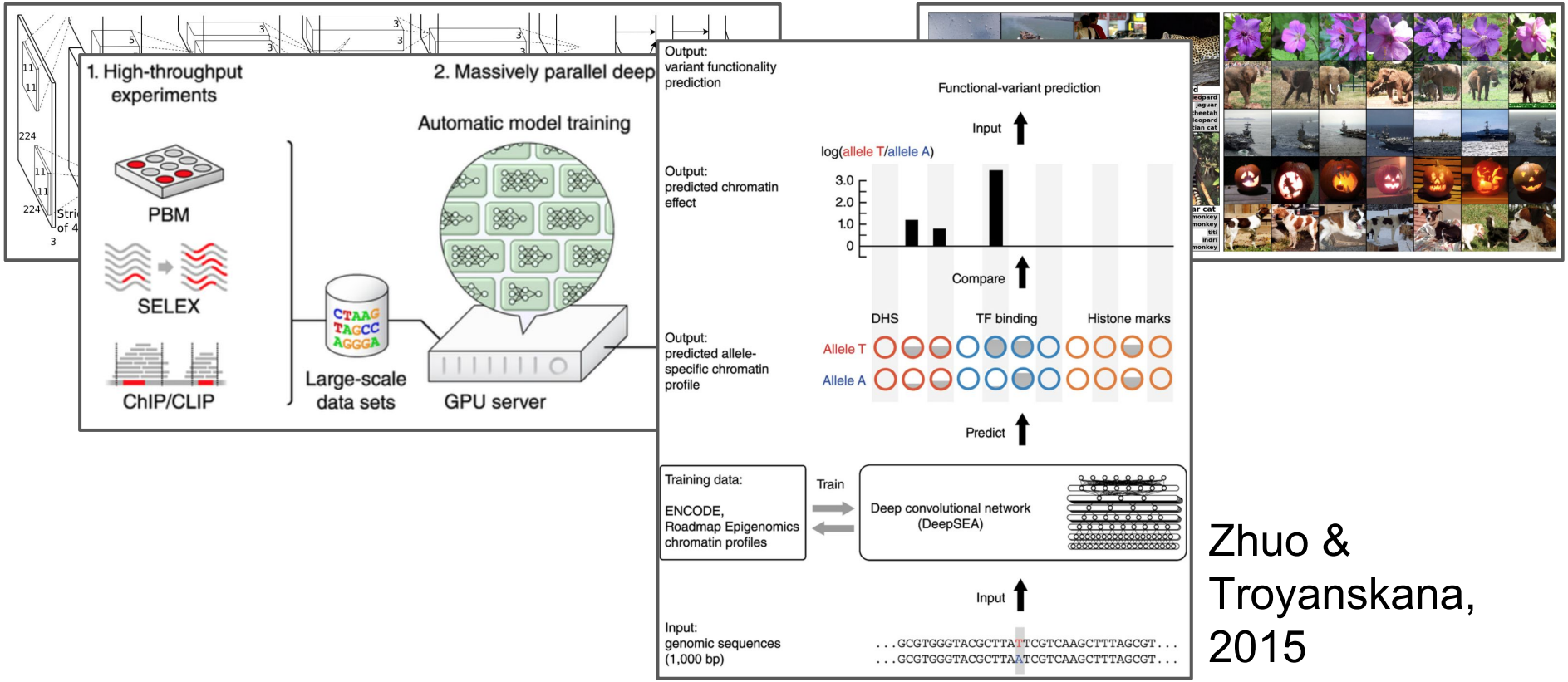
Machine Learning identifies DNA interactions

Pinello et al., 2014



Beneviste et al., 2014

Using CNNs to Advance Genomics



Zhuo & Troyanskana, 2015

“It is essential that [these techniques] are technically and conceptually accessible to the researchers who can take advantage of their potential.”

Saving the day …

- Deep CNNs

Basset

- Open Source

- Functional assessment of DNA

- Tailored to the Biosciences community

Benchmarks Using Basset:

- Predict the accessibility of DNA sequences in 164 cell types, as mapped by DNase-seq.
- Learn the relevant sequence motifs and the regulatory logic with which they are combined to determine cell-specific DNA accessibility.

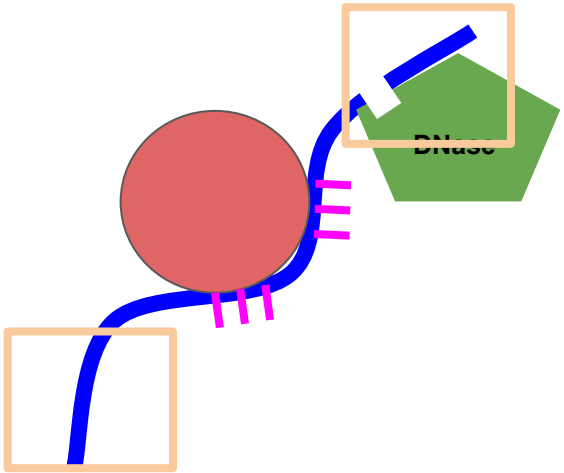
Significance:

- Meaningful, nucleotide-precision measurements.
- Scores that reflect the accessibility difference predicted by the model between two alleles.
- Highly predictive of the causal SNP among sets of linked variants.

The DNase I Hypersensitivity Dataset

- Merge of two datasets DNase-Seq datasets from 164 cell types.

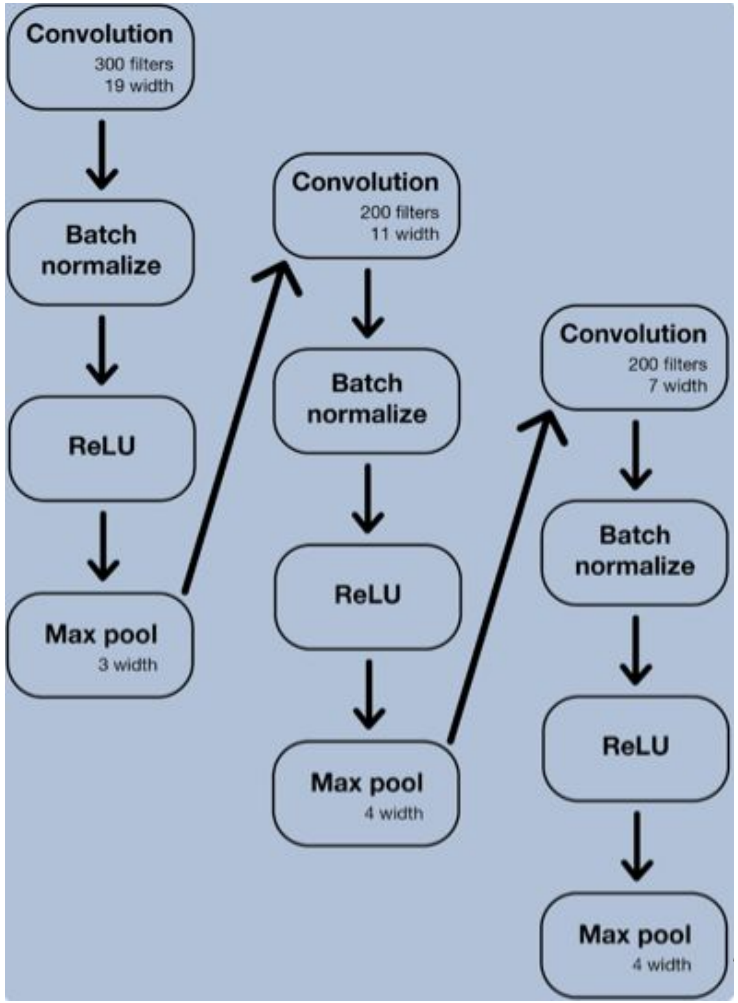
- DNase-Seq



Methods

Neural Network Architecture

- Convert to one hot code sequence
- 3x [Convolutional layer with PWMs as filters, ReLU, Max Pool]
- 2 standard fully connected layers
- 1 fully connected sigmoid to 164 outputs, representing probabilities for eh cell type.


Convolution
300 filters
19 width
Batch normalize
ReLU
Max pool
3 width
Convolution
200 filters
11 width
Batch normalize
ReLU
Max pool
4 width
Convolution
200 filters
7 width
Batch normalize
ReLU
Max pool
4 width

Data, Loss Function, and Optimizations

- Example dataset: about 2 million examples total
- about 70,000 reserved for testing, another 70,000 for validation

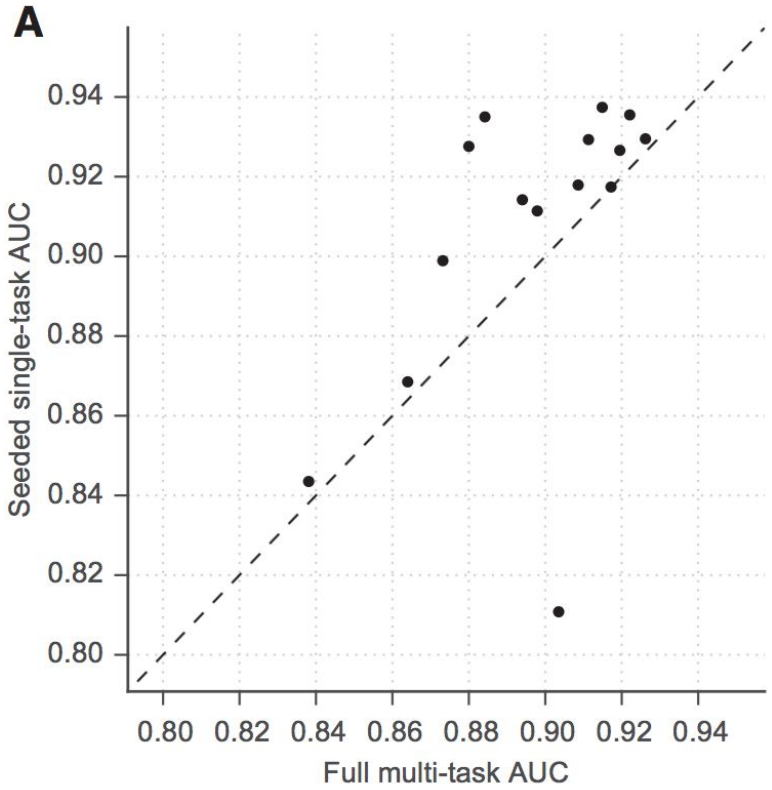
- Cross entropy loss function
- Initialization
- Stochastic gradient descent for all parameters
- RMSprop updates with mini-batches
- Dropout regularization
- Early stopping

Results

Deep CNNs predict genome accessibility

Efficient Prediction using Pretraining

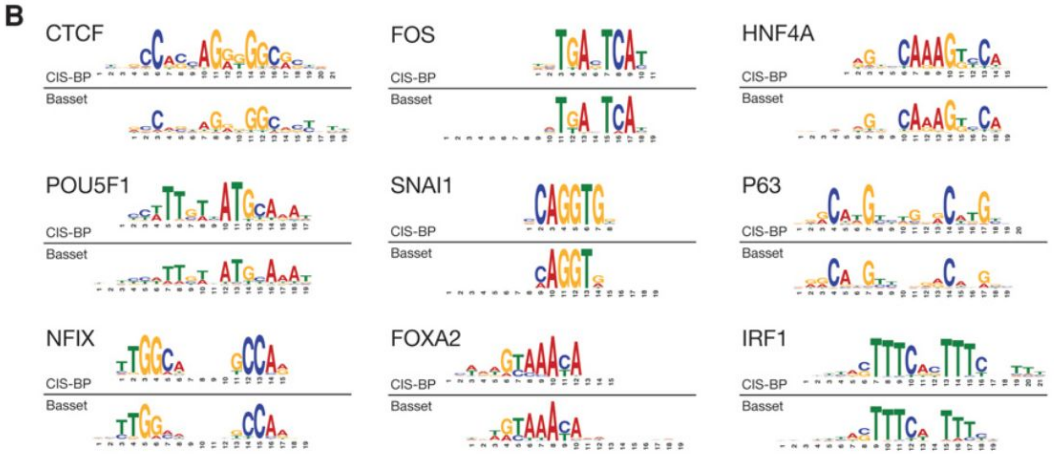
Efficient Prediction using Pretraining



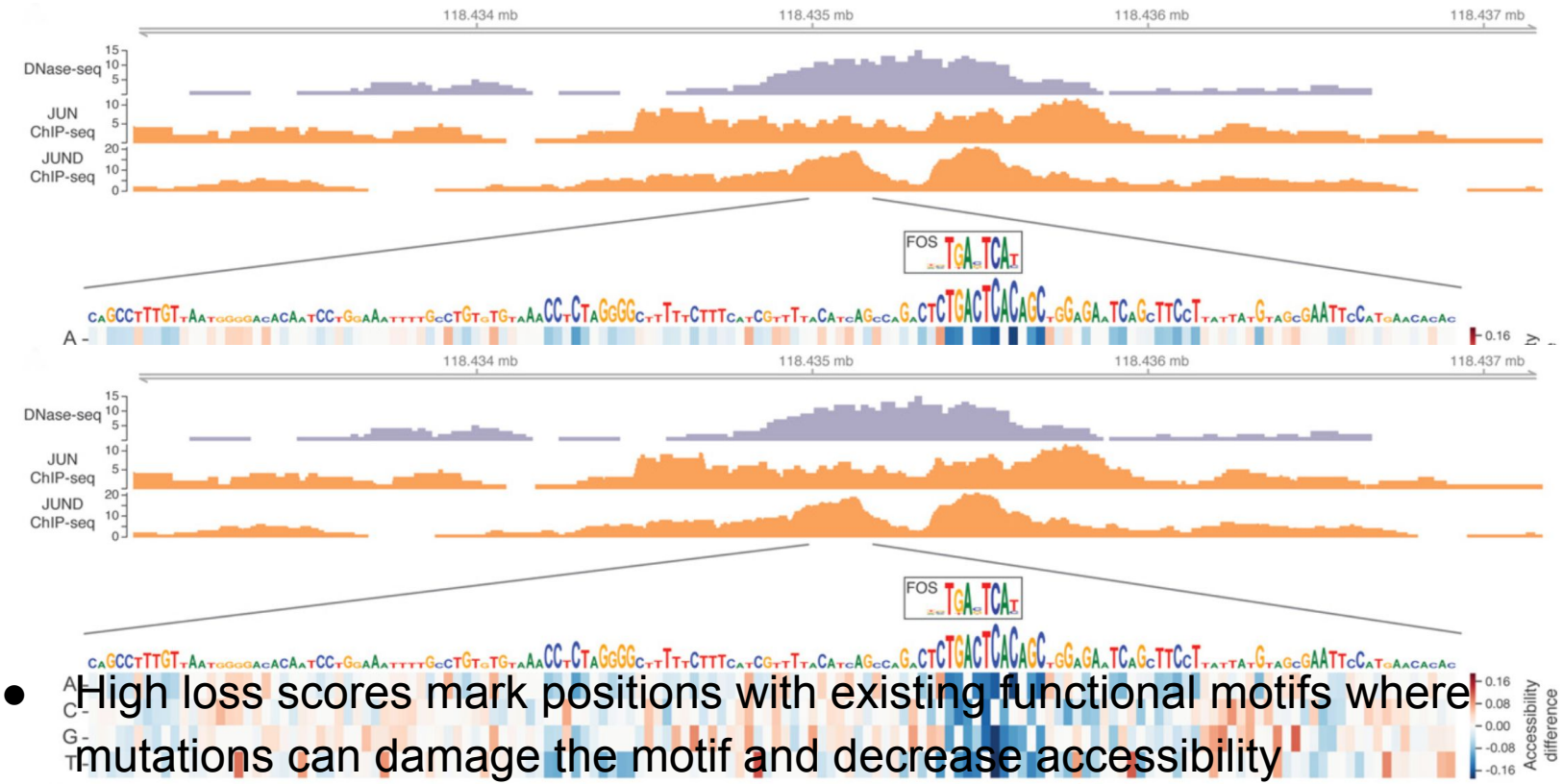
B

	GPU	CPU
Full multi-task	85 h	-
Seeded single-task	18 m	6 h 37 m

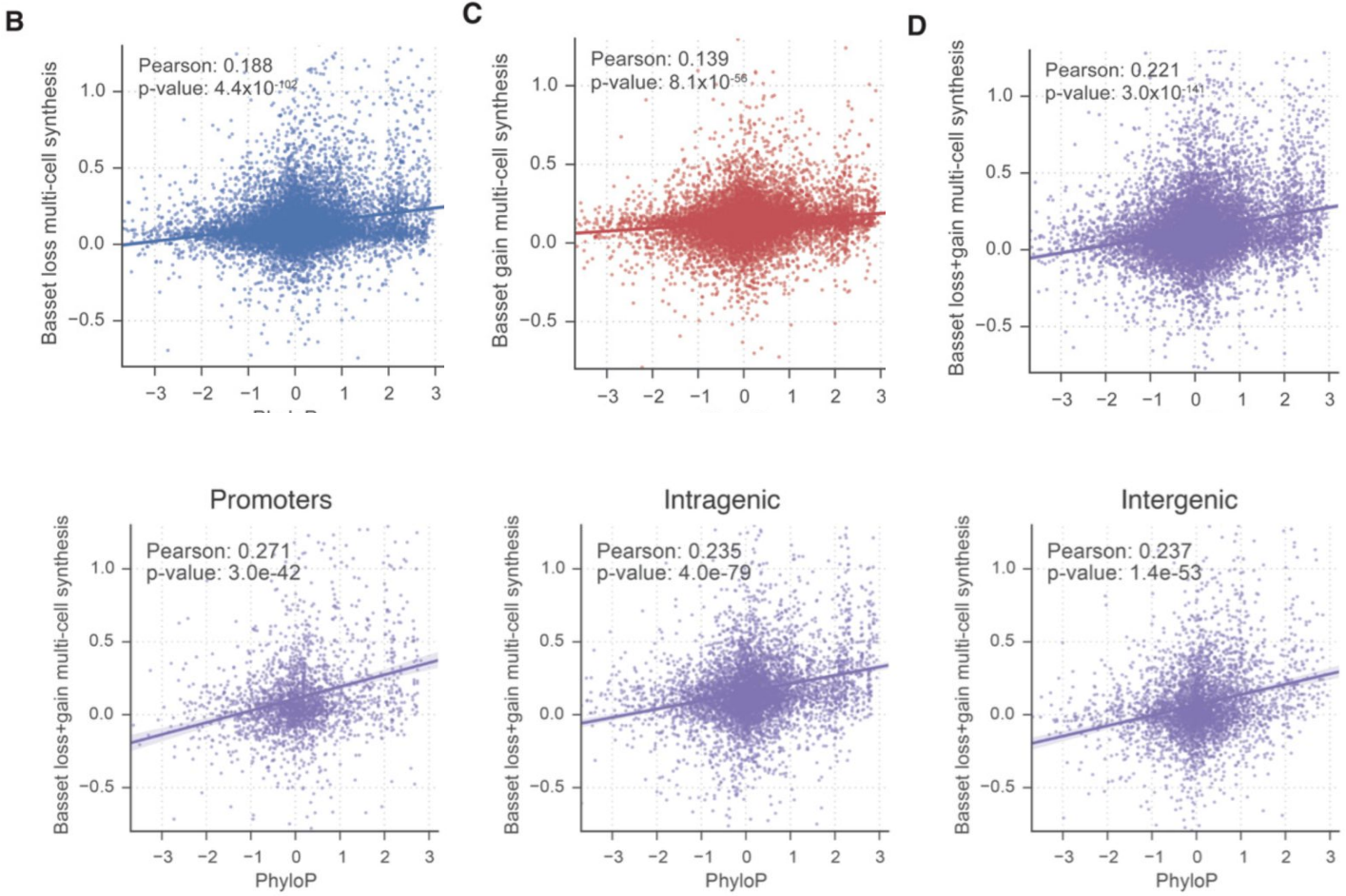
Recovery of protein binding motifs



In silico saturation mutagenesis (ISSM) pinpoints nucleotides driving accessibility



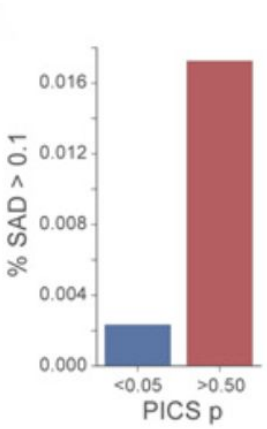
- High loss scores mark positions with existing functional motifs where mutations can damage the motif and decrease accessibility


B
C
D
Basset loss multi-cell synthesis
Pearson: 0.188
p-value: 4.4x10-102
Basset gain multi-cell synthesis
Pearson: 0.139
p-value: 8.1x10-56
Basset loss+gain multi-cell synthesis
Pearson: 0.221
p-value: 3.0x10-141
Promoters
Pearson: 0.271
p-value: 3.0e-42
Intragenic
Pearson: 0.235
p-value: 4.0e-79
Intergenic
Pearson: 0.237
p-value: 1.4e-53
Basset loss+gain multi-cell synthesis
PhyloP

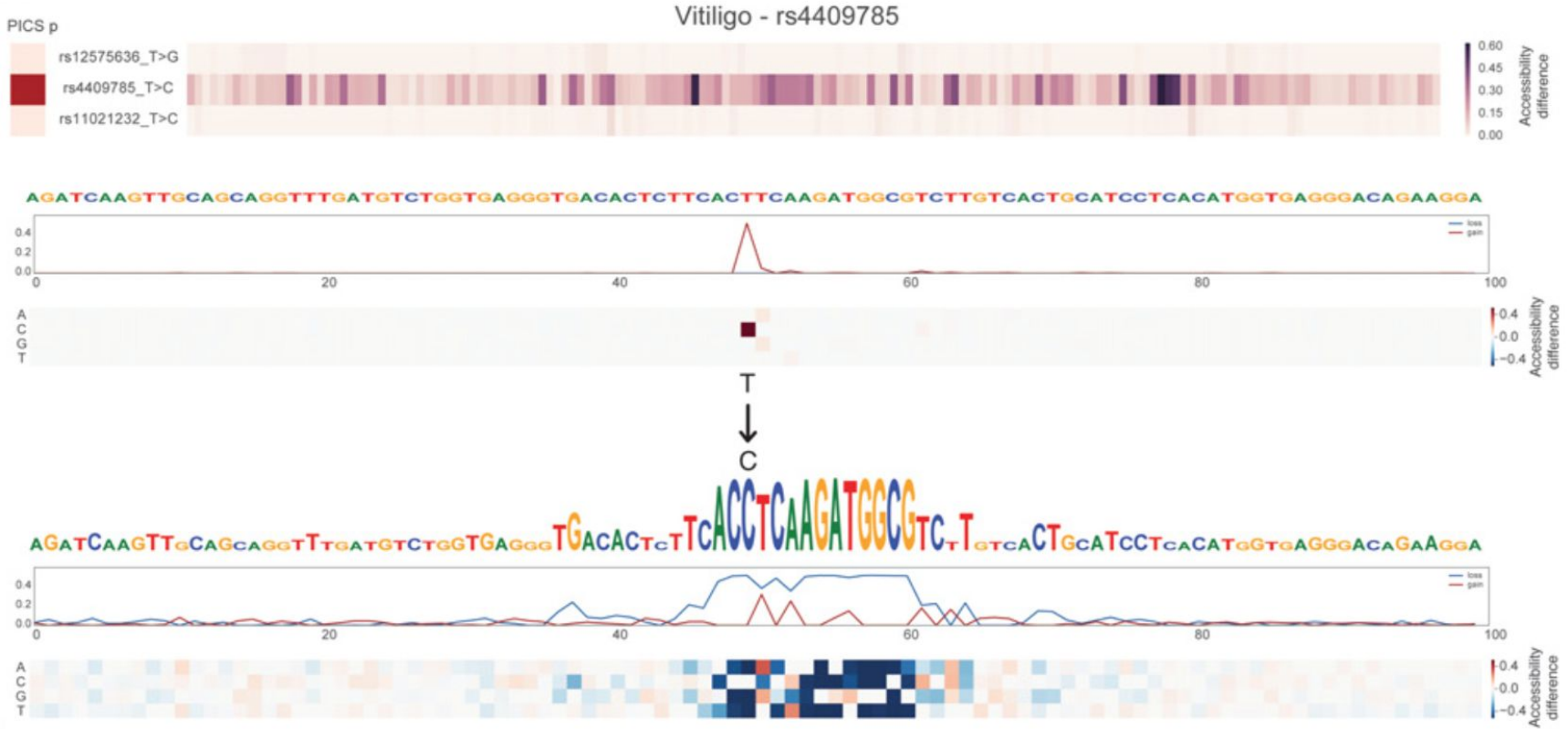
Basset predicts greater accessibility changes for likely causal GWAS SNPs

- Genome-wide association studies (GWAS) have uncovered ample noncoding variants associated with physical traits and disease in human populations.

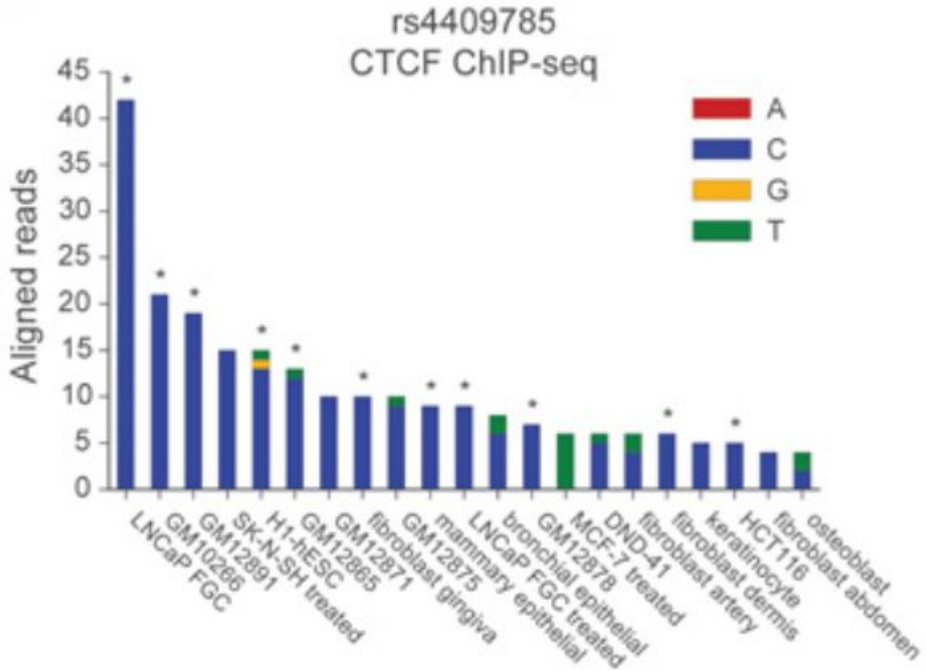
- A set of 7252 non-coding GWAS SNPs associated with auto-immune disease were analyzed with a statistical method called PICS



Basset predicts greater accessibility changes for likely causal GWAS SNPs



Basset predicts greater accessibility changes for likely causal GWAS SNPs



Discussion

- Basset is an open source package to apply deep CNNs to learn DNA sequence activity.
- Effectively learns the complex code of DNA accessibility across many cell types and substantially surpasses the predictive accuracy of the present state of the art.
- NNs trained via stochastic gradient descent scale very well to large data sets, allowing us to learn good parameters within a general and expressive model structure.
- Researcher can learn a cell's chromatin accessibility code and annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility with just a single sequencing assay in their cell type of interest

Caveats

- TensorFlow is becoming the standard for Neural Network development and Basset does not use it.
- Only trained on DNase-seq data which doesn’t capture epigenetic effects
- The paper seems to be very exploratory and its case studies often leave much to be further researched.
- Realistically, the results are still a very far away from achieving informing personalized medicine.

Questions?