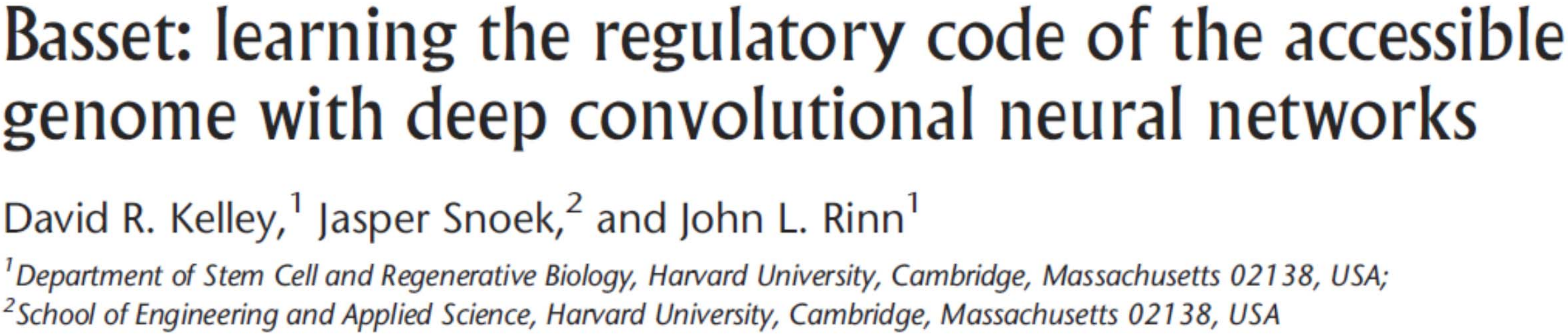


# Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

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*

AH

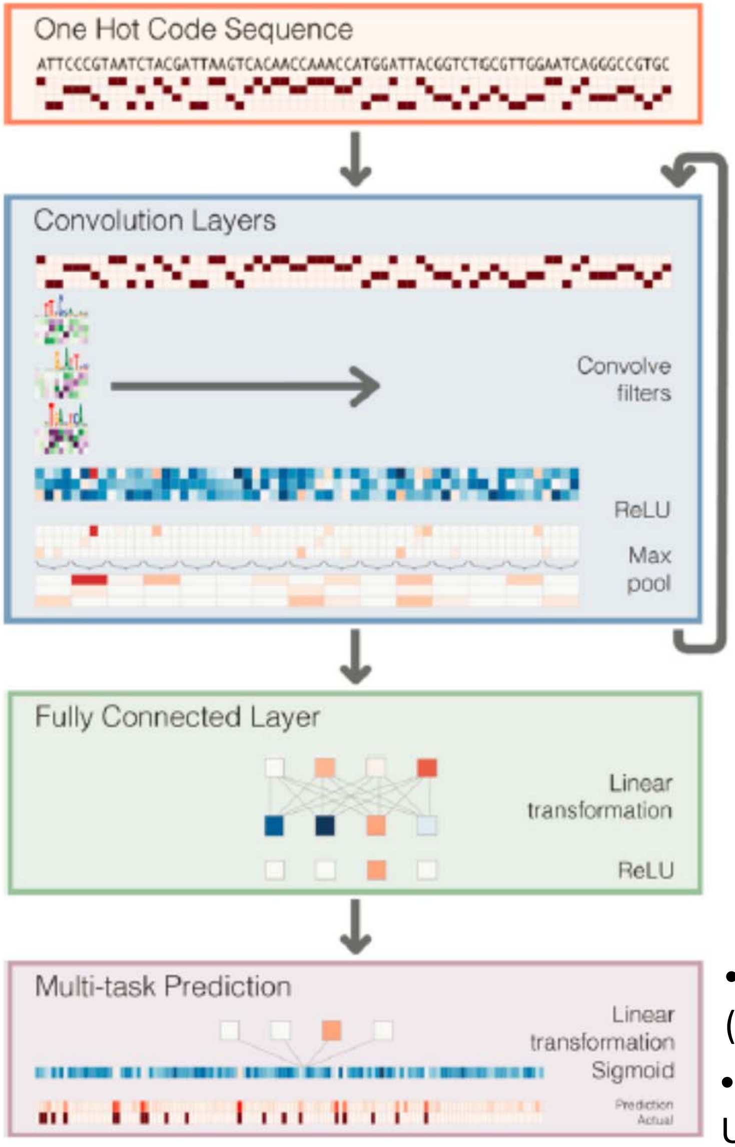
# Bassett

- A machine learning model for learning the regulatory code of sequence
  - Focus is on open and closed chromatin
- The model aims to learn the DNA sequence signals for the cell specific signatures of open and closed chromatin
- The model is a ***convolutional neural network*** (CNN)
- Current state-of-the-art method is called gkm-SVM, which is an SVM method.

# ENCODE and RMEC DNase-Seq Datasets

- The ENCODE Project Consortium: DNase-seq on 125 cell types (Thurman et al. 2012)
- The Roadmap Epigenomics Consortium: Additional 39 (Roadmap Epigenomics Consortium et al. 2015).
- Pooling of these generates 2 million sites across all cells
- The GENCODE v18 reference annotation
  - 17% promoters,
  - 47% intragenic,
  - 36% intergenic,
  - 4.1%–19.0% (median 8.2%), are accessible in any individual cell type,
  - 3.8% constitutively open in >50% of the cells.
- For each DHS site, authors extracted 600 bp from the hg19 reference genome around the midpoint as input to the model.

# Bassett



Use the current filter in a sliding window manner (There are in total 300 Filters)

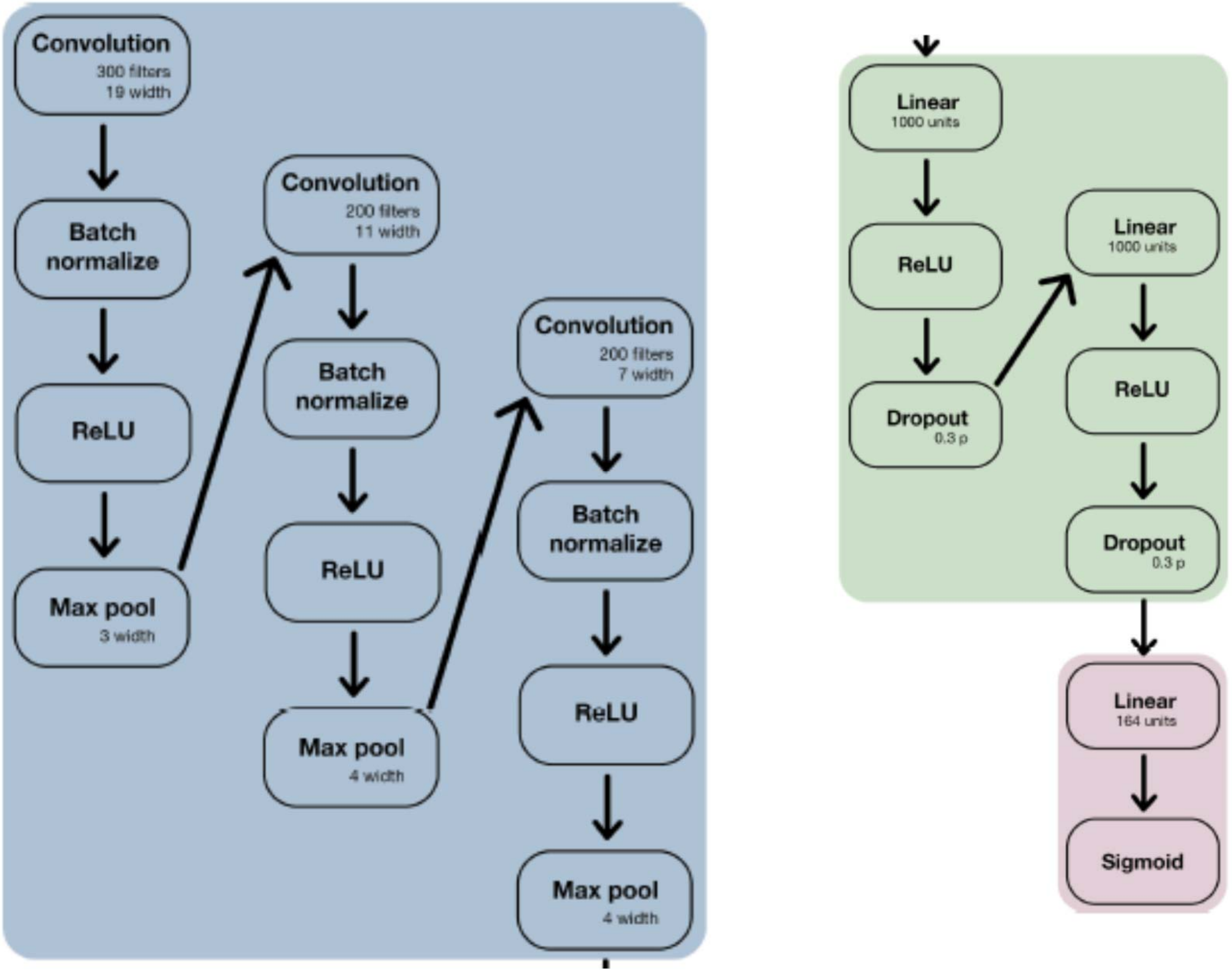
Each layer is a position weight matrix corresponding to the output of the previous layer

- First layer is the PWM on the DNA sequence
- Later layers are higher order PWMs: Subsequent convolution layers consider the orientations and spatial distances between patterns recognized in the previous layer.

There are in total 3 convolution layers.

- Sigmoid function generates scores between 0 and 1 (probabilities of open chromatin)
- Training involves updating the filter weights using the DNase-seq peaks Utilizes backpropogation for training

# Bassett



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

Linear 1000 units

ReLU

Dropout 0.3 p

Linear 1000 units

ReLU

Dropout 0.3 p

Linear 164 units

Sigmoid

# Accuracy of Bassett

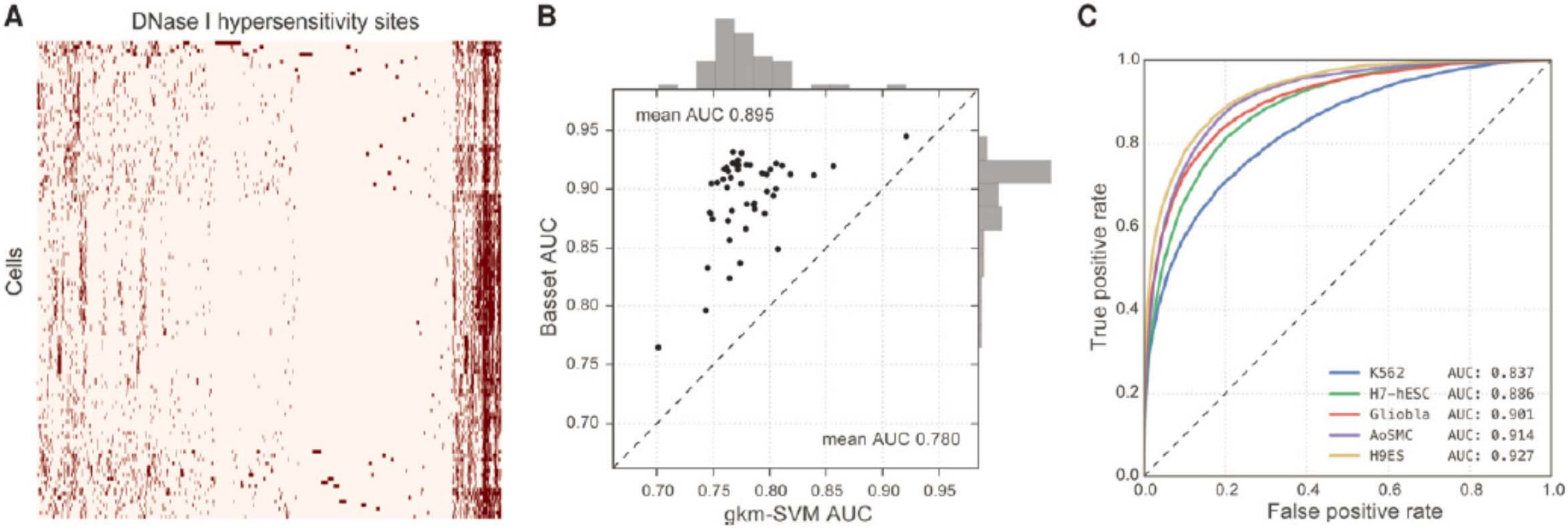
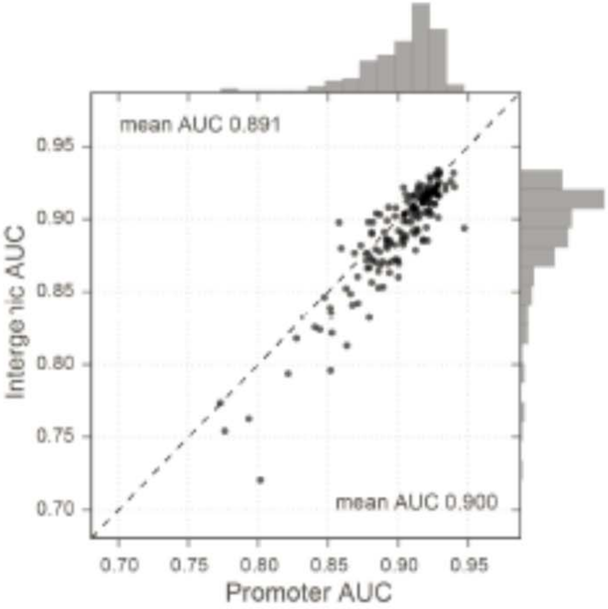
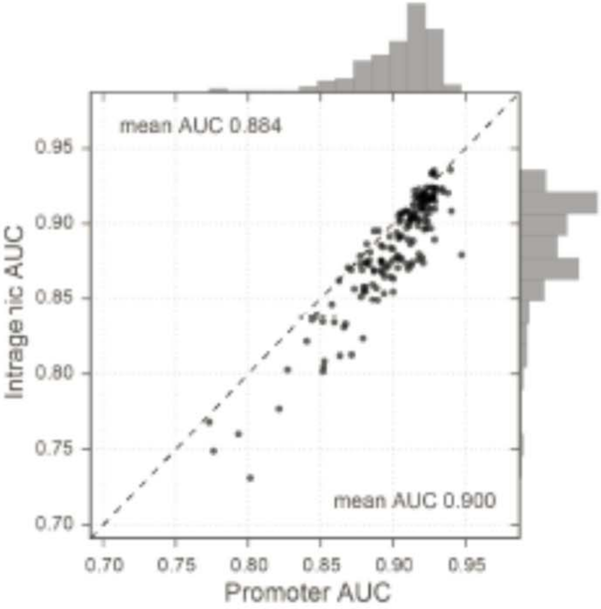
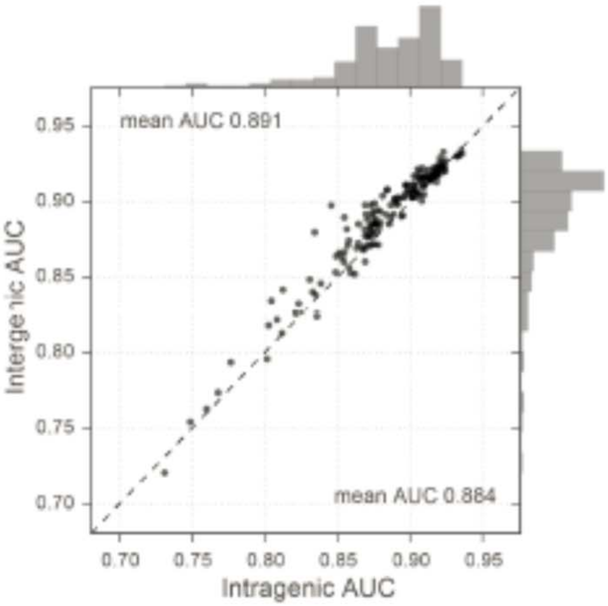


**Figure 2.** Basset accurately predicts cell-specific DNA accessibility. (*A*) The heat map displays hypersensitivity of 2 million DNase I hypersensitive sites (DHSs) mapped across 164 cell types. We performed average linkage hierarchical clustering using Euclidean distance to both cells and sites. (*B*) The scatter plot displays AUC for 50 randomly selected cell types achieved by Basset and the state-of-the-art approach gkm-SVM, which uses support vector machines. (*C*) The ROC curves display the Basset false-positive rate versus true-positive rate for five cells, selected to represent the 0.05, 0.33, 0.50, 0.67, and 0.95 quantiles of the AUC distribution.

# Accuracy of Bassett: Different Annotations





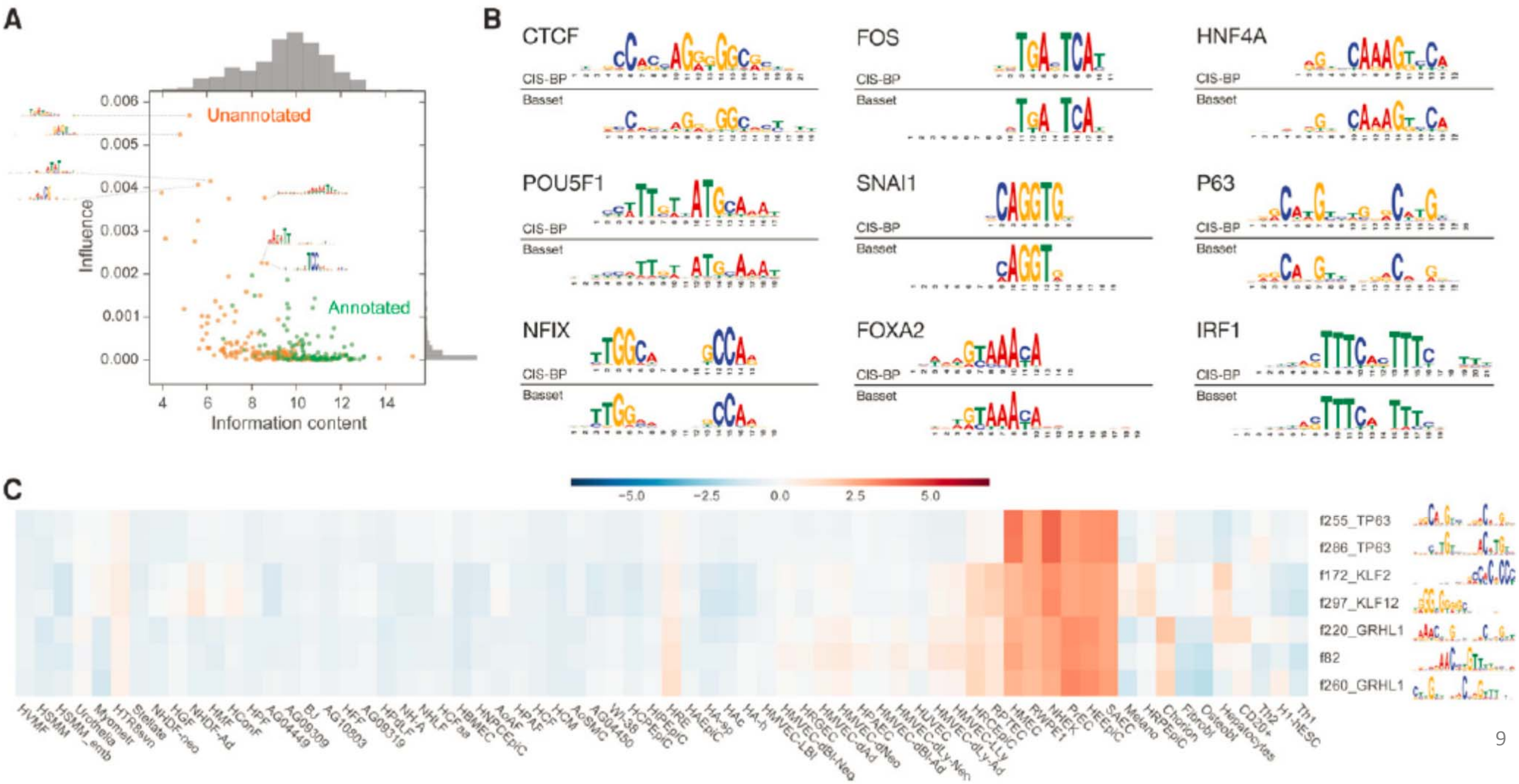


Promoter > Intergenic > Intragenic

# PWMs can be extracted from the model

- To assess the contribution of a filter, force the filter’s output to a constant value
- The change in the predicted accessibility probability is an estimate of how much the filter contributes.
- CTCF was the most predictive of the DNA accessibility.
- Model “allocates” 12 filters to 19-bp long CTCF motif
- Each filter is a variant of the motif: Different nucleotide content or a translation of the motif
- At q-value threshold of 0.1, 45% of the filters align significantly to proteins motifs in CIS-BP

# Basset recovers known protein binding motifs



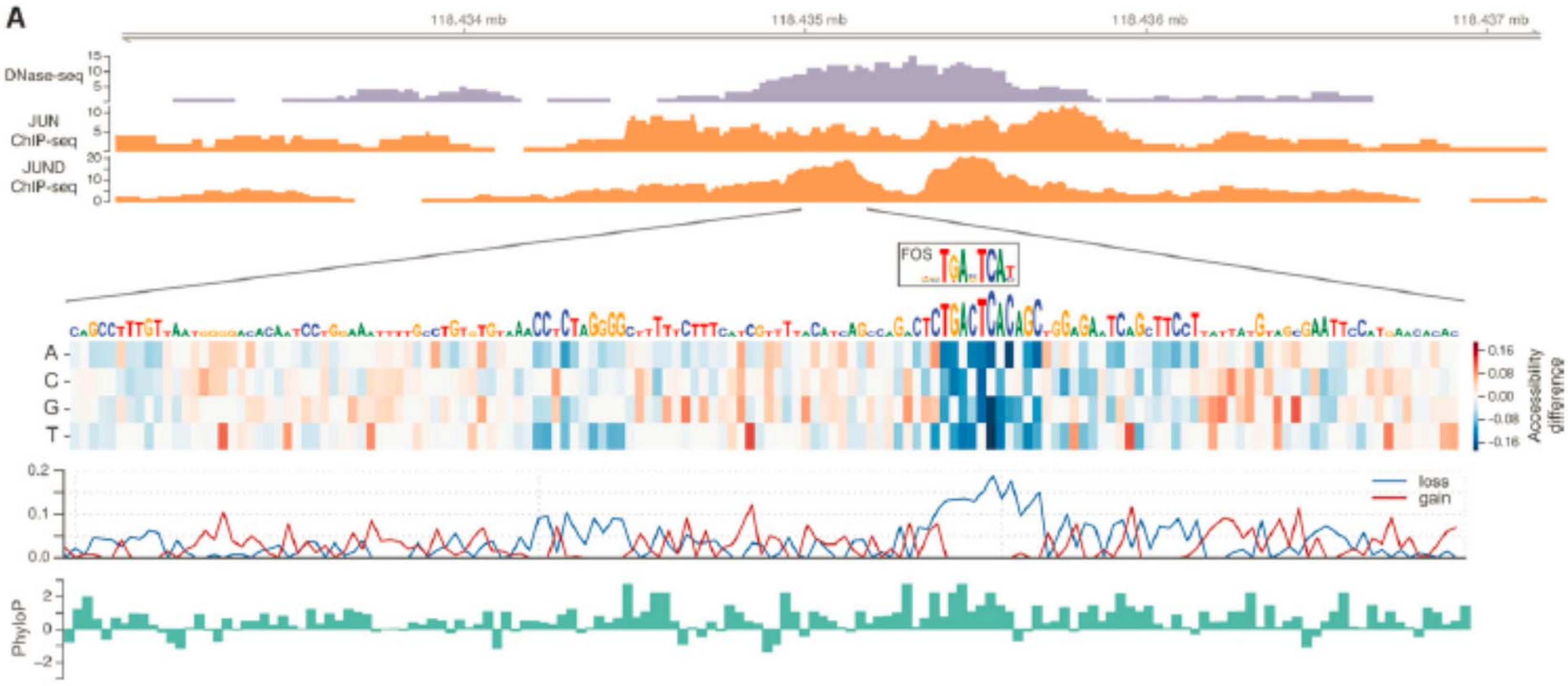
# Unrecognized filters reveal other features with lower complexity sequence content

- Poly AT stretches
- High GC content
- CpG islands
- This shows importance of these low complexity regions in determining DNA accessibility
- Also 100 nt flanking sequences affect the predictability of DNA accessibility
- Some unrecognized filters contributions match to those that regulate development
  - Future work is necessary to determine their exact role

# In silico saturation mutagenesis

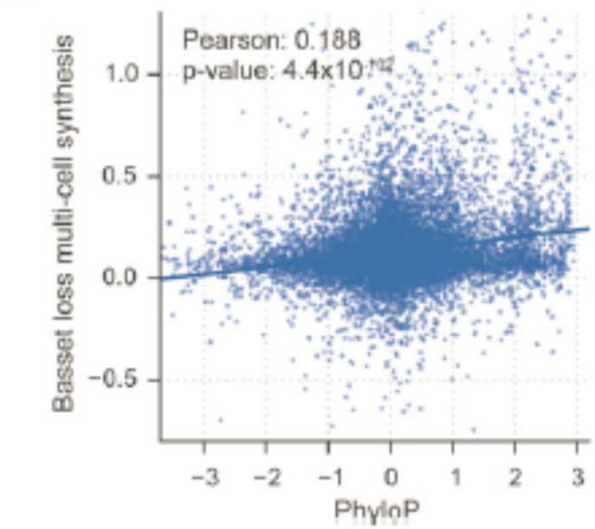
- Every mutation in the sequence is tested to estimate its effect on DNA accessibility
- The authors changed every nucleotide and computed the change in the predicted accessibility at each position
- Two scores are assigned to each *in silico* mutation:
  - Loss score: Largest possible decrease in predicted accessibility
  - Gain score: Largest possible increase in predicted accessibility
- High loss scores indicate possible functional mutations
- High gain scores indicate positions with possible motif gain
  - They call this latent
- Authors also compare the scores with PhyloP

# In silico saturation mutagenesis pinpoints nucleotides driving accessibility

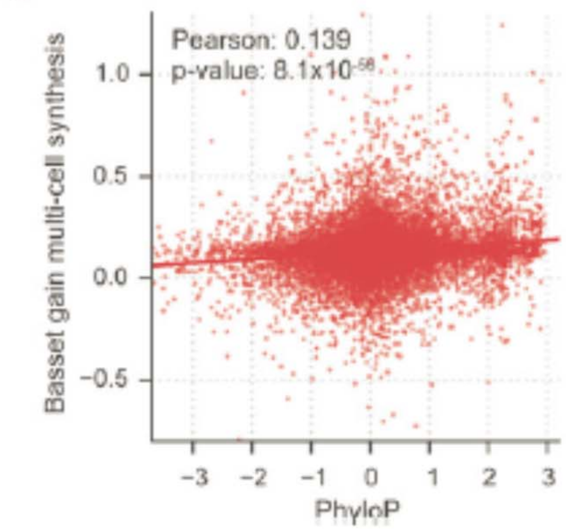


# Comparison with PhyloP

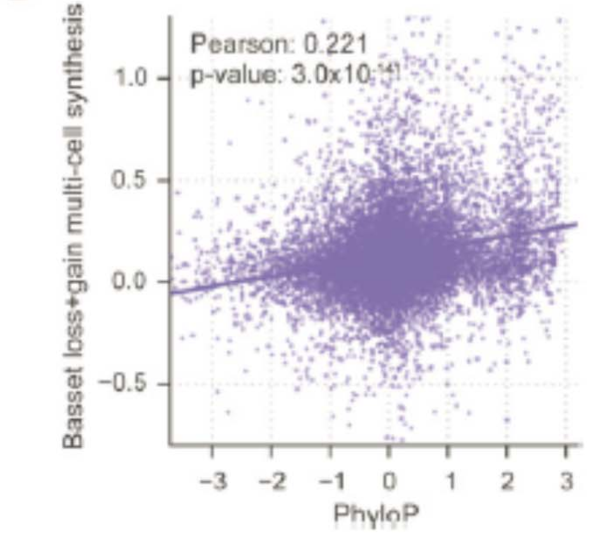
Correlation with Loss Mutations



Correlation with Gain Mutations



Correlation with Gain and Loss Mutations

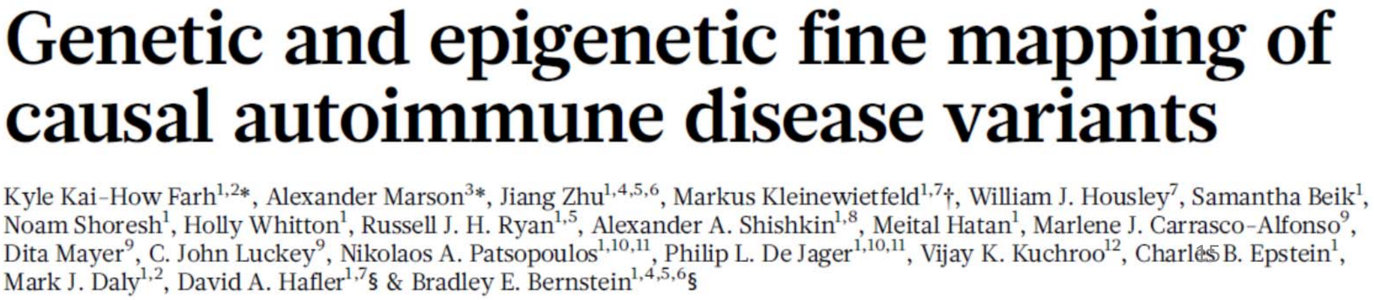


# Relation to GWAS SNPs

- GWAS SNPs are enriched within DHSs
- The authors hypothesize that Bassett can be used for prioritizing non-coding variants
- For this, they define SNP Accessibility Difference (SAD): Difference in predicted accessibility across cell types between two alleles of the SNP
- Since there is very limited amount of confirmed positive examples, the authors compare the SAD scores with an orthogonal method named PICS used on 8,741 GWAS SNPs
    - *My comment: It is kind of ad-hoc to use another tool to verify SAD scores*
    - *There is also no citation for PICS paper*

# Relation to GWAS SNPs

- Farh et al 2015: 8741 GWAS SNPs associated with autoimmune diseases
- Most of the time PICS algorithm identified causal SNP with high probability
- Authors focus on 7252 SNPs from this study where no SNP is in LD with a SNP in a protein coding gene
- 235 high PICS score SNPs
- 3004 low PICS score SNPs

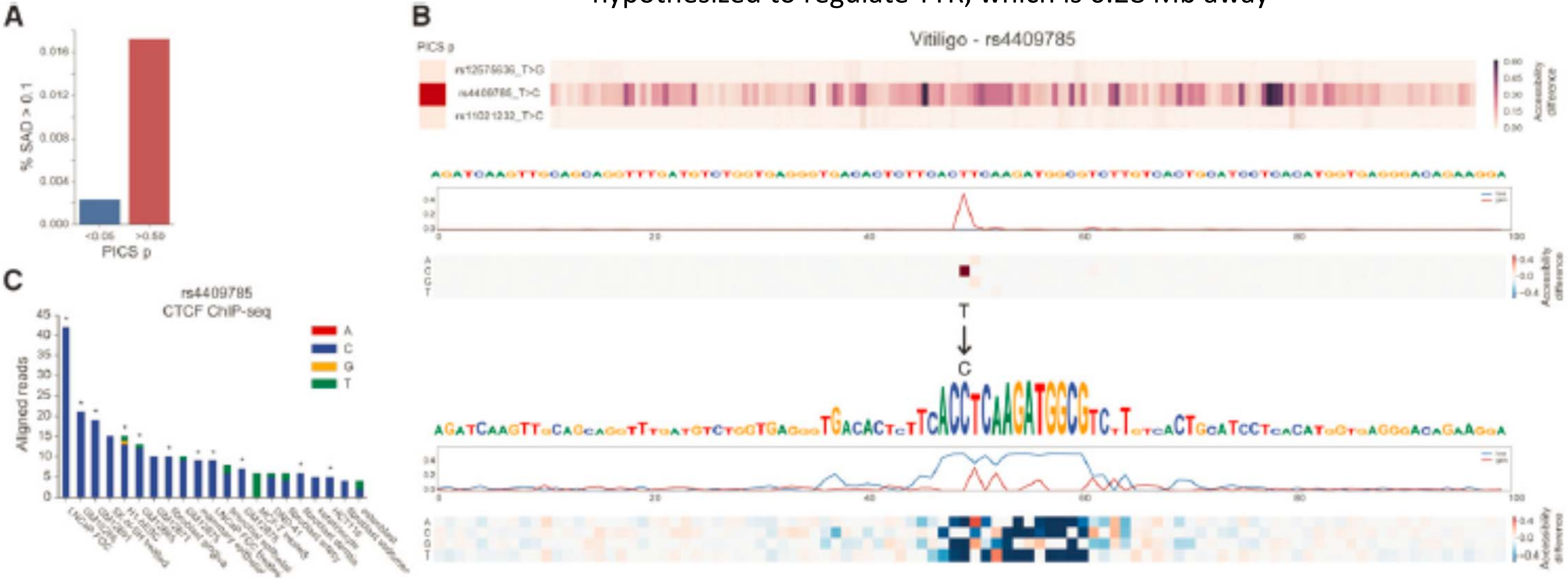


**Genetic and epigenetic fine mapping of causal autoimmune disease variants**

Kyle Kai-How Farh[1,2]*, Alexander Marson[3]*, Jiang Zhu[1,4,5,6], Markus Kleinewietfeld[1,7]†, William J. Housley[7], Samantha Beik[1], Noam Shoresh[1], Holly Whitton[1], Russell J. H. Ryan[1,5], Alexander A. Shishkin[1,8], Meital Hatan[1], Marlene J. Carrasco-Alfonso[9], Dita Mayer[9], C. John Luckey[9], Nikolaos A. Patsopoulos[1,10,11], Philip L. De Jager[1,10,11], Vijay K. Kuchroo[12], Charles B. Epstein[1], Mark J. Daly[1,2], David A. Hafler[1,7]§ & Bradley E. Bernstein[1,4,5,6]§

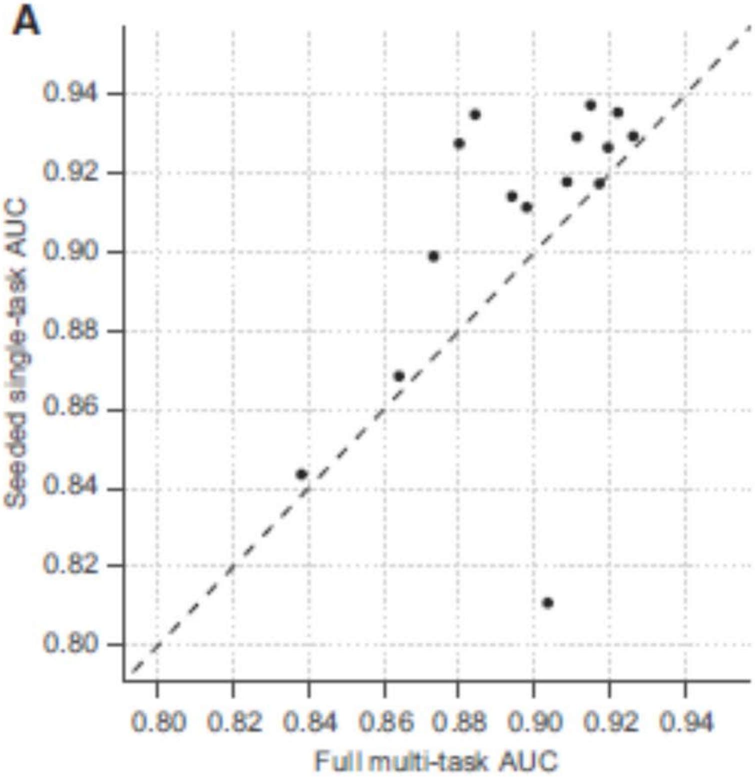
# Basset predicts greater accessibility changes for likely causal GWAS SNPs

rs4409785 is located in a 559-kb gene desert. However, it has been hypothesized to regulate TYR, which is 6.28 Mb away

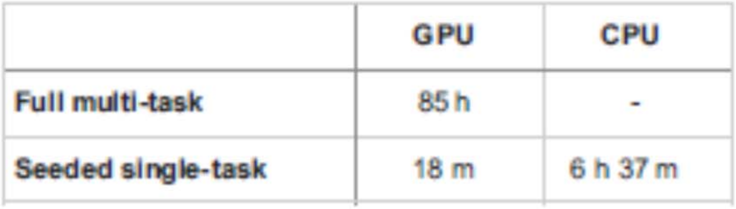


SNP creates a CTCF binding site, authors hypothesize that this may affect the chromatin structure

# Leveraging large-scale models allows accurate and efficient prediction of new data sets



B



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