CS273B: Deep learning for Genomics and Biomedicine

Lecture 8: Genomics applications paper discussions

10/18/2017

Anshul Kundaje, James Zou

Outline

1

BASENJI

2

DeepCpG

3

Evaluation of strategies for training multi-task CNNs for genomics

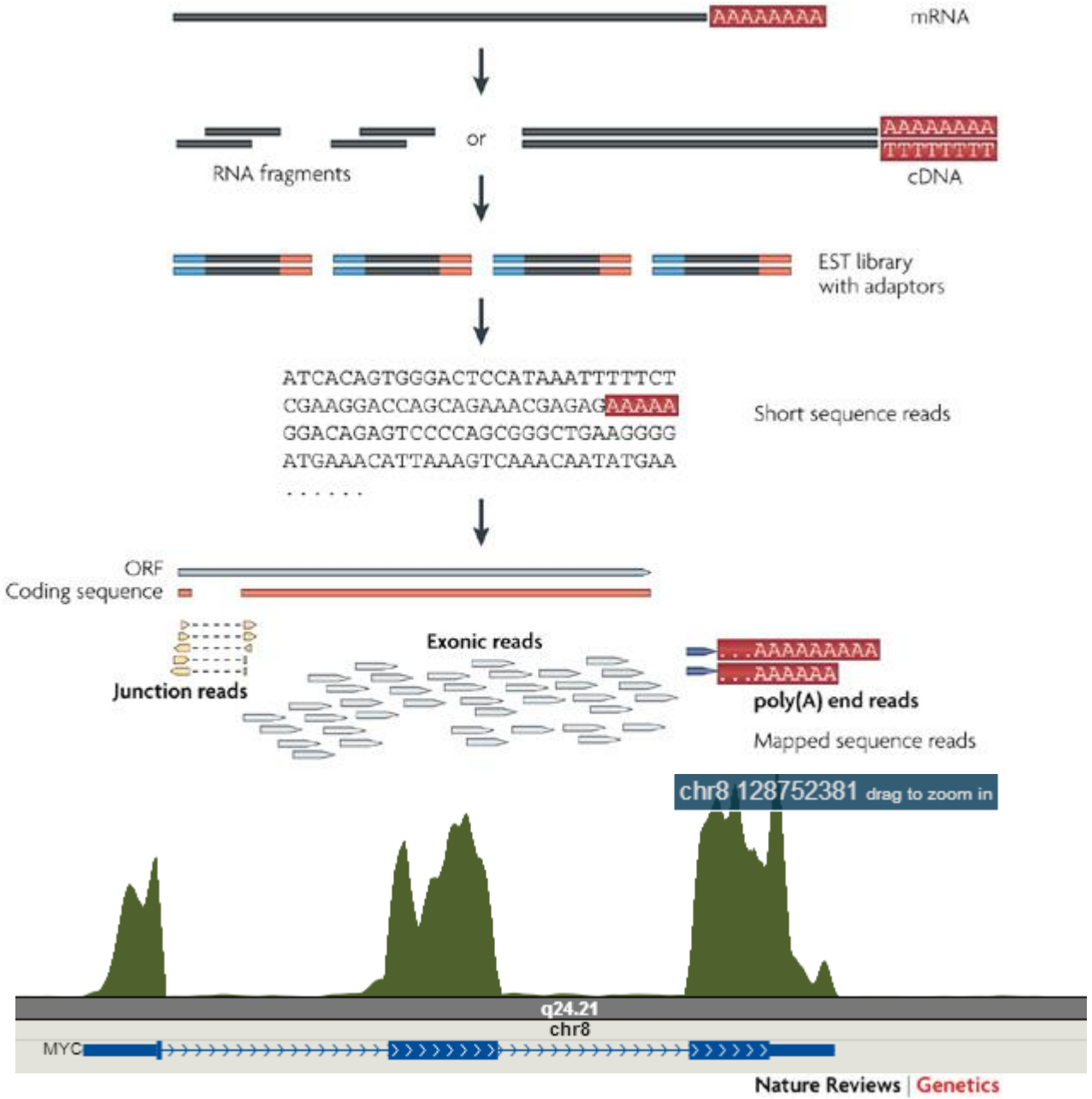
BASENJI

Sequential regulatory activity prediction across chromosomes with convolutional neural networks

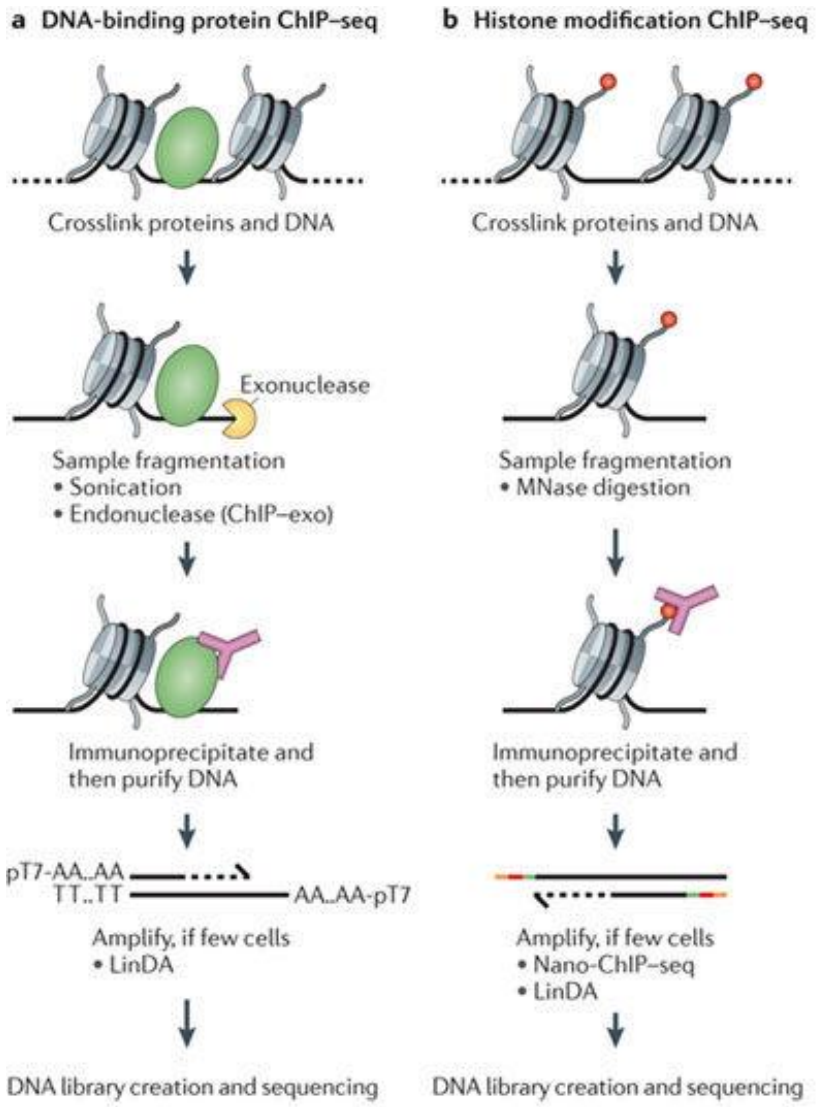
David R. Kelley, Yakir A. Reshef

https://doi.org/10.1101/161851

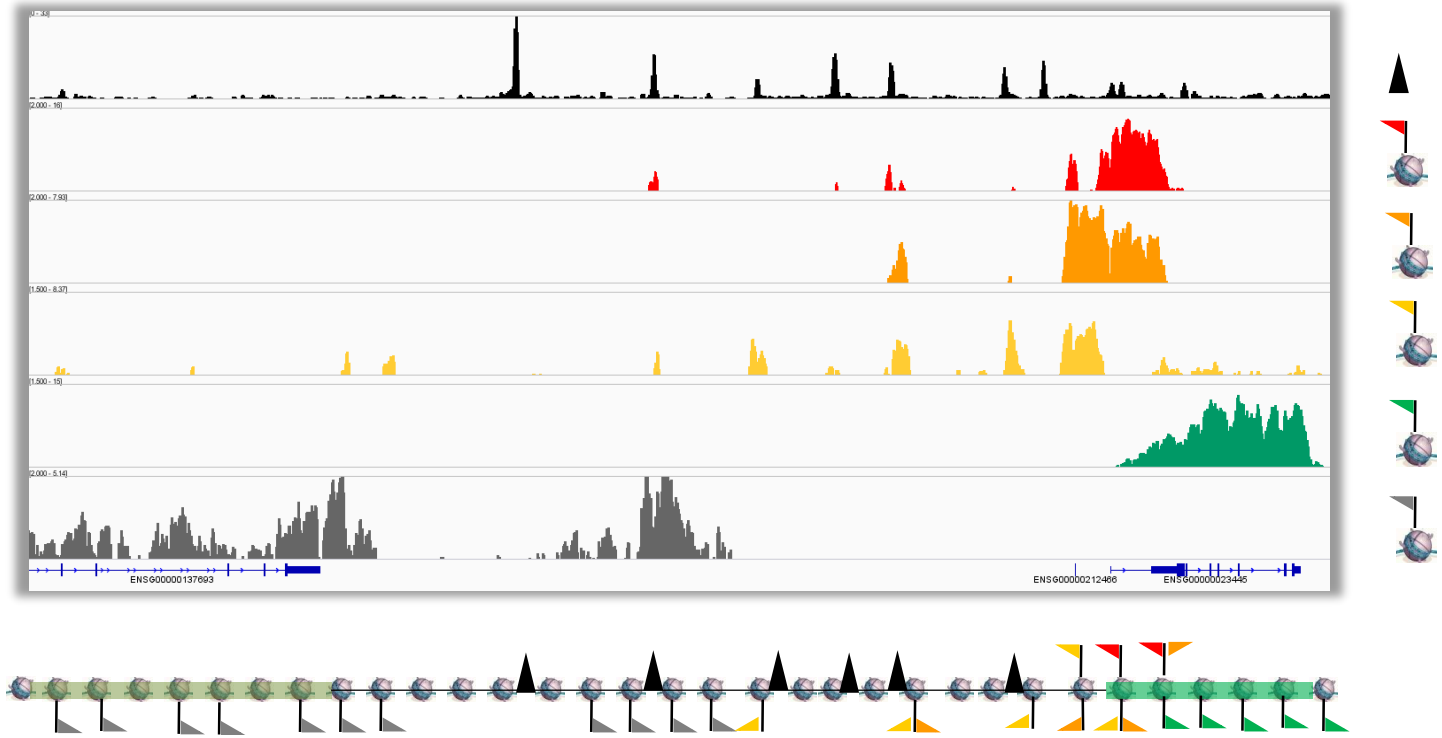
RNA sequencing (RNA-seq)



Chromatin immunoprecipitation (ChIP-seq)



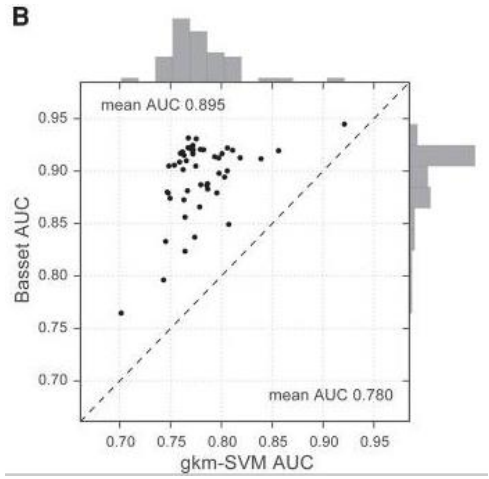
Protein-DNA binding maps
Maps of epigenomic modifications



Genome Res. 2016 Jul;26(7):990-9. doi: 10.1101/gr.200535.115. Epub 2016 May 3.

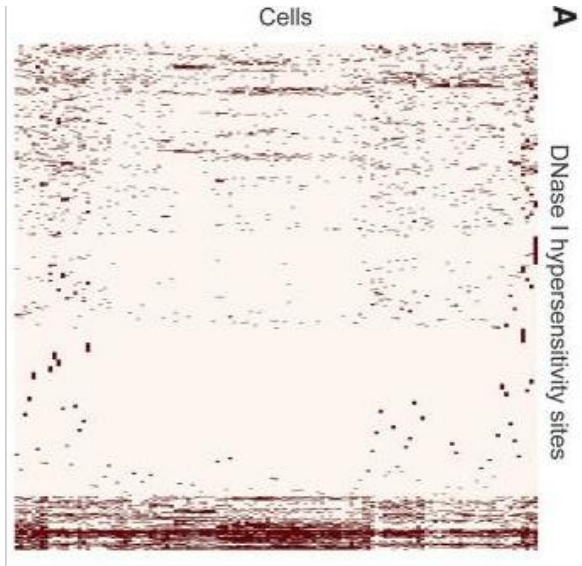
Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks.

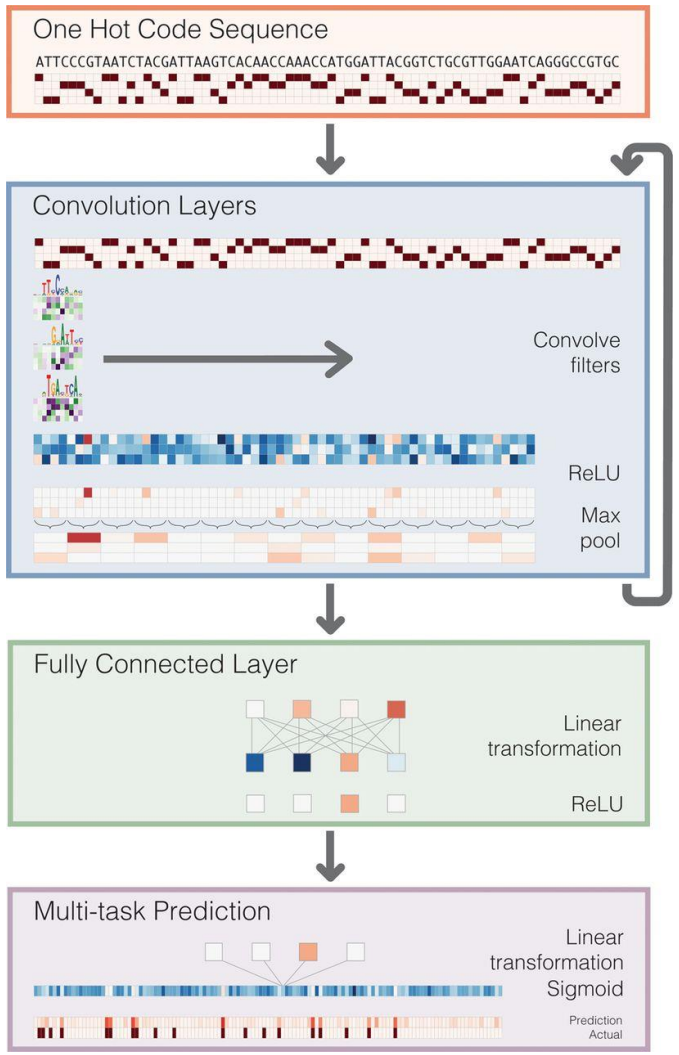
Kelley DR[1], Snoek J[2], Rinn JL[1].

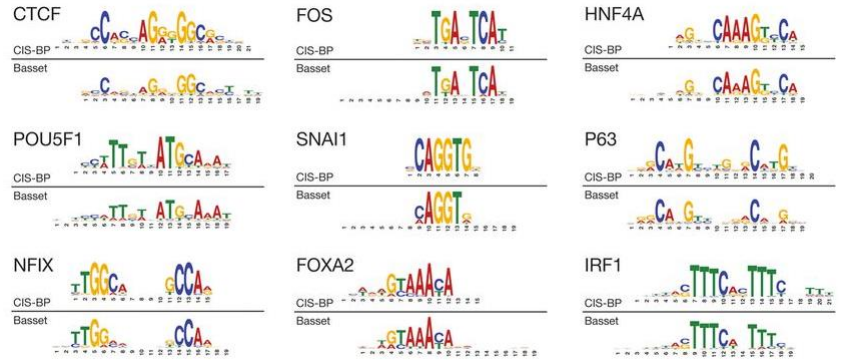


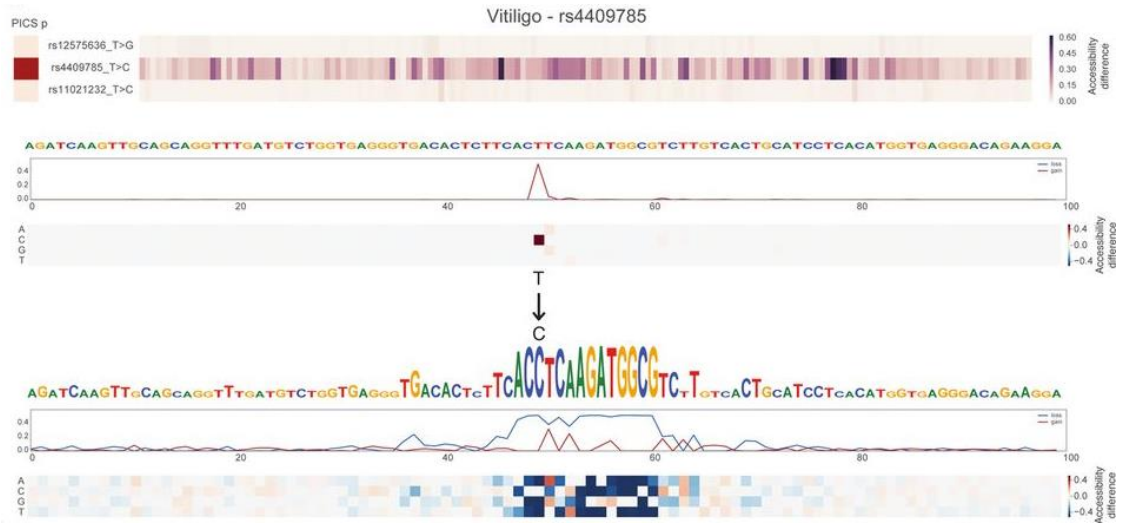
164 Tasks (cell types)

2 million examples (DHS)









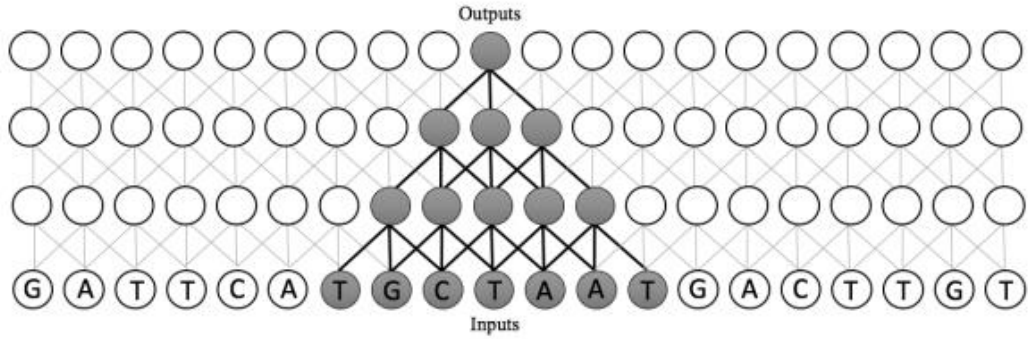
Drawbacks of Basset

- Each bin in the genome is a training/test example. Local sequence around that bin is used to predict the output of the bin.

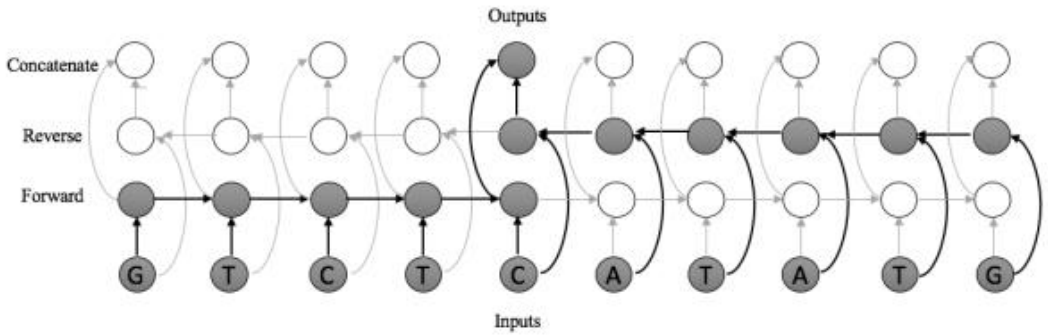
- This may be ok for chromatin based outputs where it is mostly local sequence that affects it. Although 3D genome interactions can cause distal sequence to affect label of a bin.

- This is especially true for expression prediction of genes (i.e. promoters of genes). Local sequence of promoter is only one part of the regulatory input. Distal enhancers are primary drivers of expression in human genome.

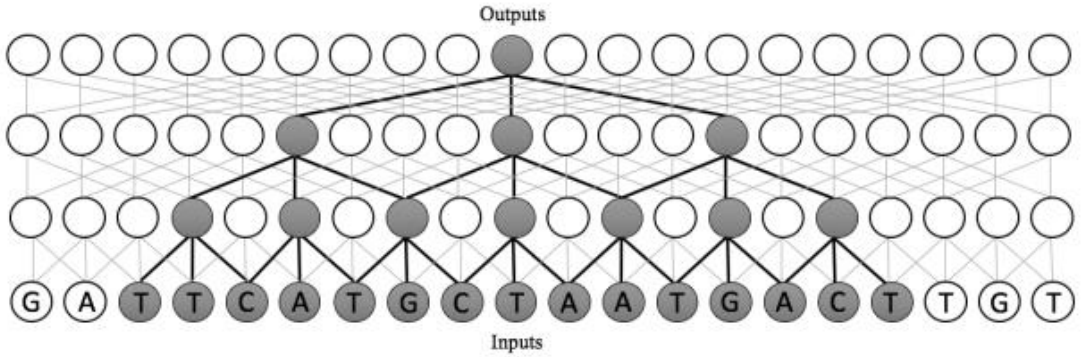
Dilated convolutions instead of convolutions



(a) Convolution



(b) Bidirectional LSTM



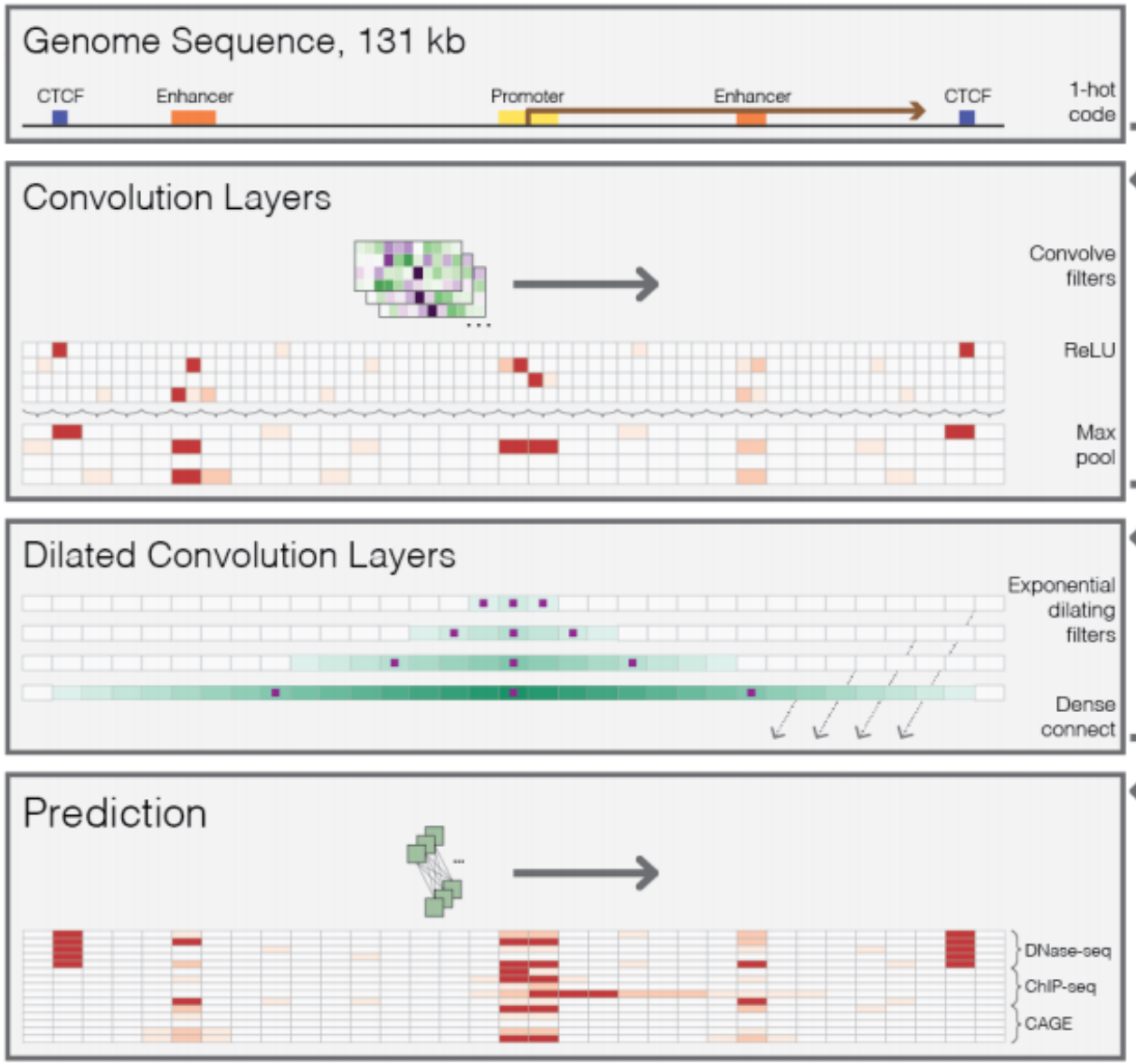
(c) Dilated Convolution

Larger receptive field but few parameters and short distance gradient propagation

For CNNs, the size of the receptive field is linear in the number of layers and the kernel width. Thus, scaling the receptive field to incorporate a large input introduces more layers, making training more difficult. Bidirectional RNNs on the other hand have a receptive field of the whole input, but require gradients to travel long-distances over time. LSTMs still have trouble learning very long-distance relationships.

https://arxiv.org/pdf/1710.01278.pdf

Basenji architecture



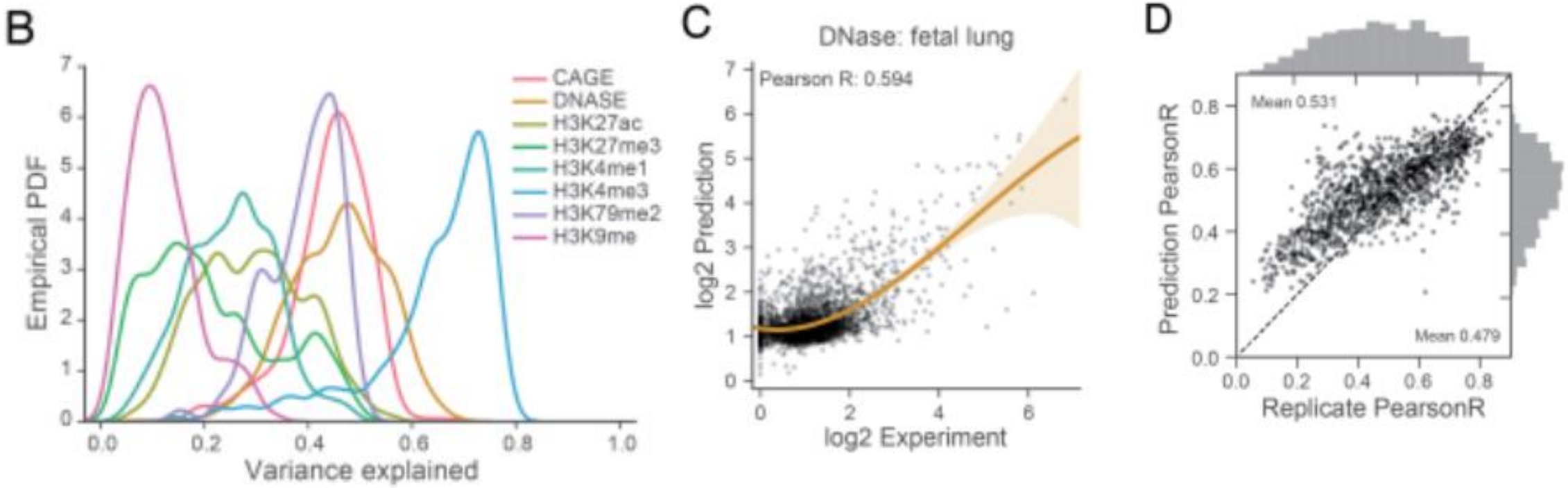
131 KB sequence

131 KB / 128 bp

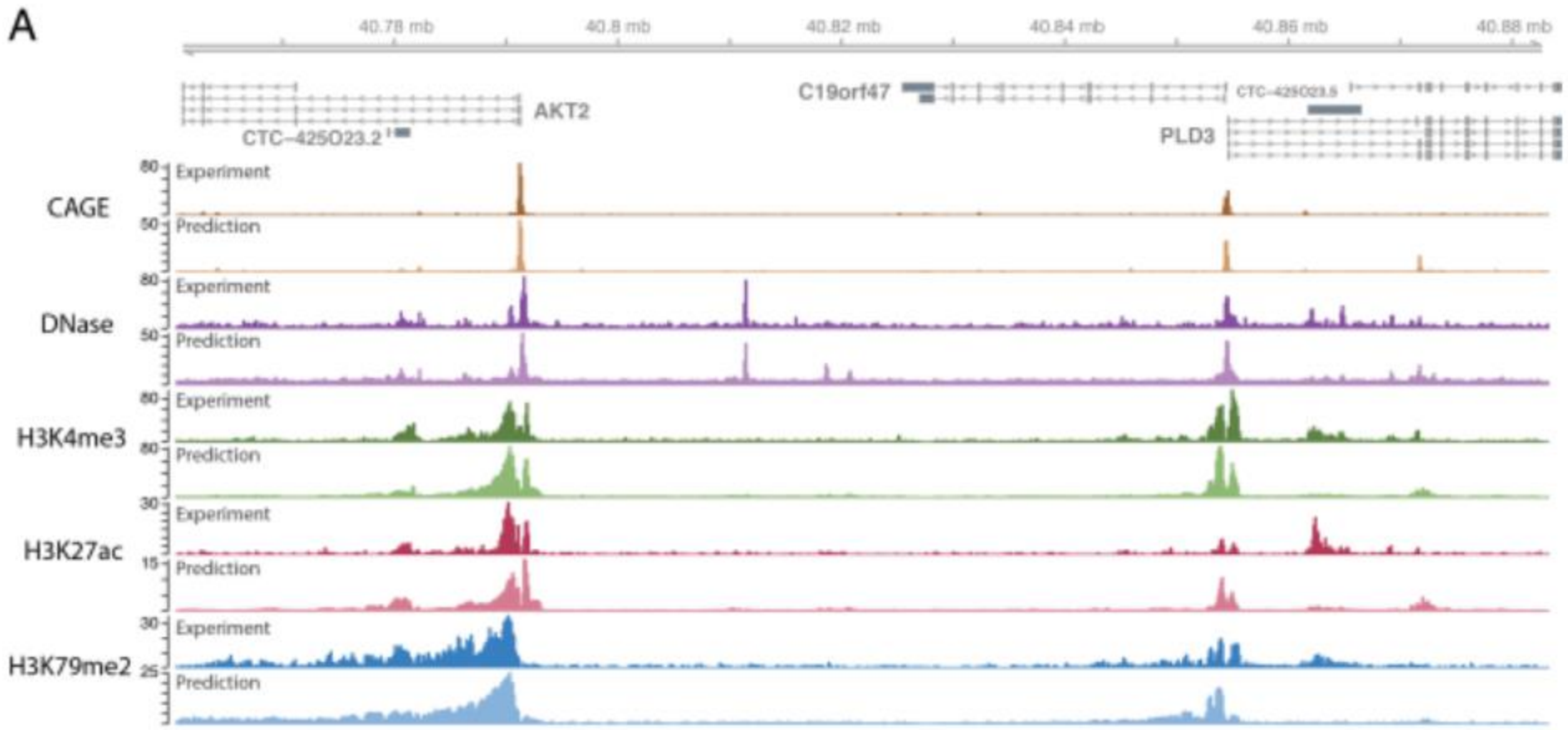
Multi-task Poisson regression output

131 KB / 128 bp outputs

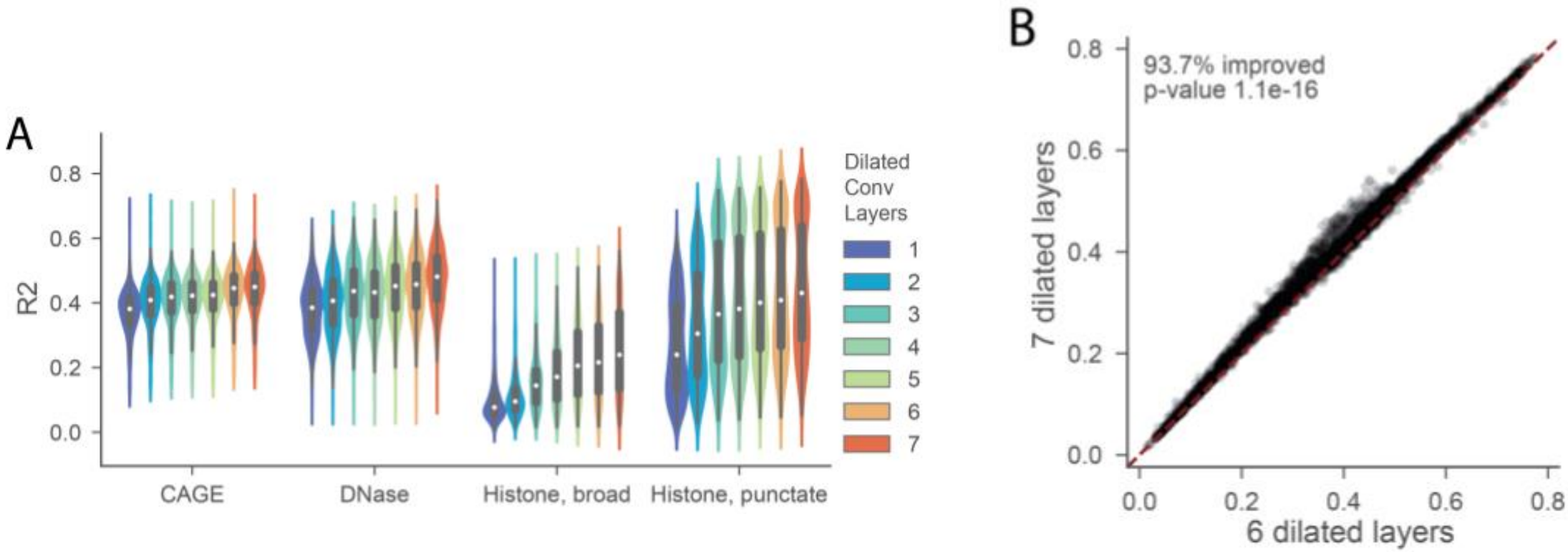
Prediction performance

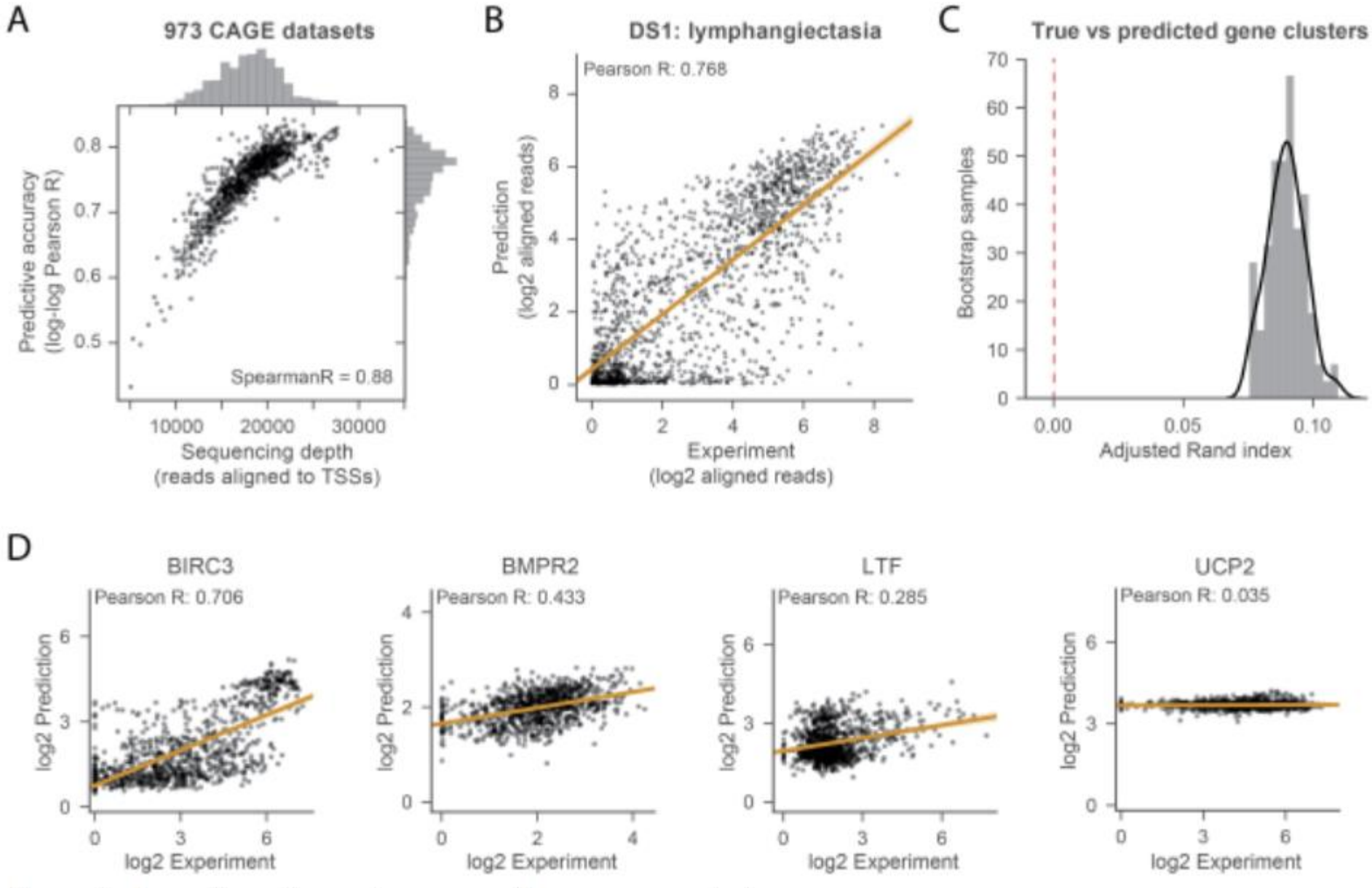


Predictions at an example locus

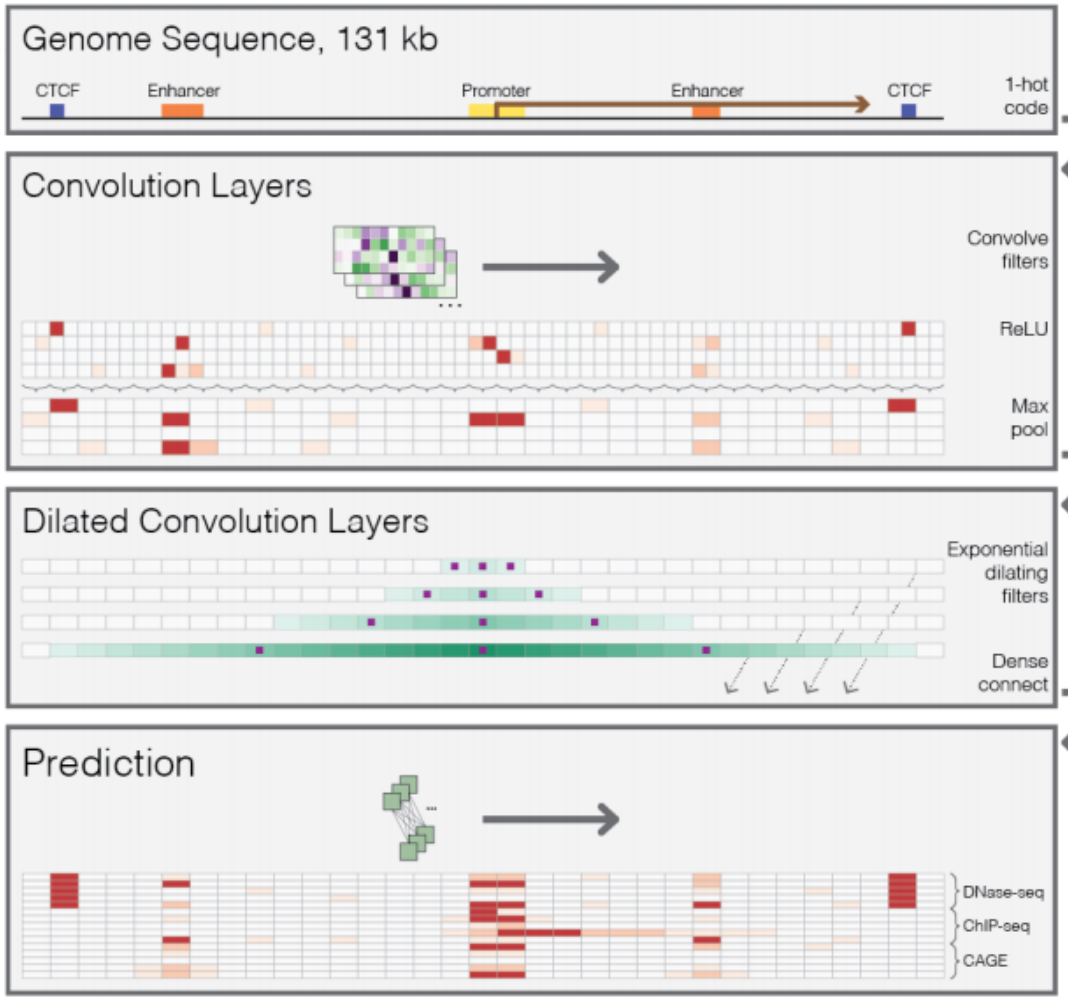


Improvements in performance with dilated layers




A
973 CAGE datasets
Predictive accuracy
(log-log Pearson R)
SpearmanR = 0.88
Sequencing depth
(reads aligned to TSSs)
B
DS1: lymphangiectasia
Pearson R: 0.768
Prediction
(log2 aligned reads)
Experiment
(log2 aligned reads)
C
True vs predicted gene clusters
Bootstrap samples
Adjusted Rand index
D
BIRC3
Pearson R: 0.706
BMPR2
Pearson R: 0.433
LTF
Pearson R: 0.285
UCP2
Pearson R: 0.035
log2 Prediction
log2 Experiment

Saliency maps (gradients*input) to identify predictive distal bins



Gradient dy/dx tells u how infinitesimal change of a specific input x affects output y i.e. sensitivity of output to that input keeping all other inputs constant

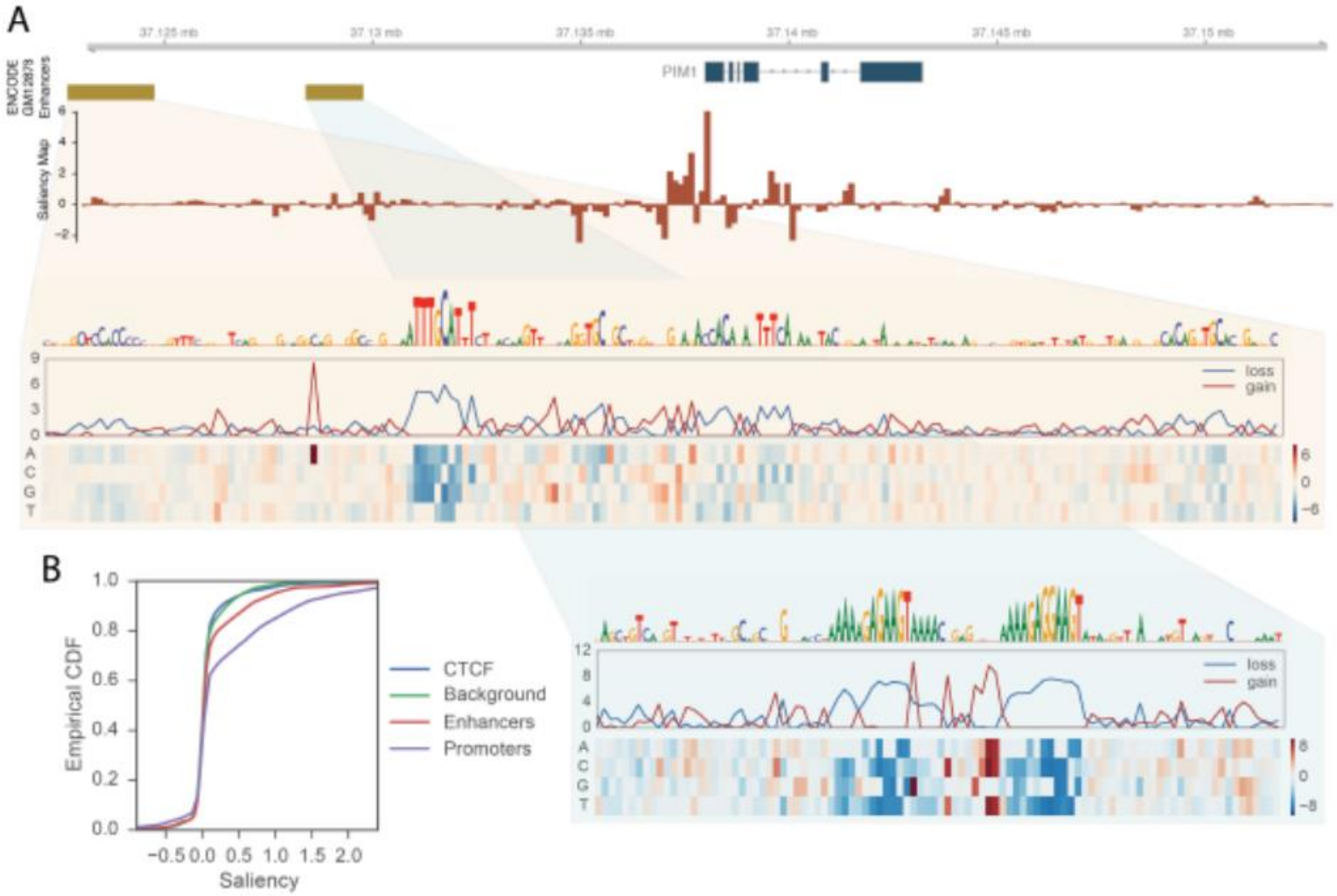
Can be used as a measure of importance/contribution of input to that output

For a specific training/test example which of the 131 KB / 128 bp inputs are important for a specific output

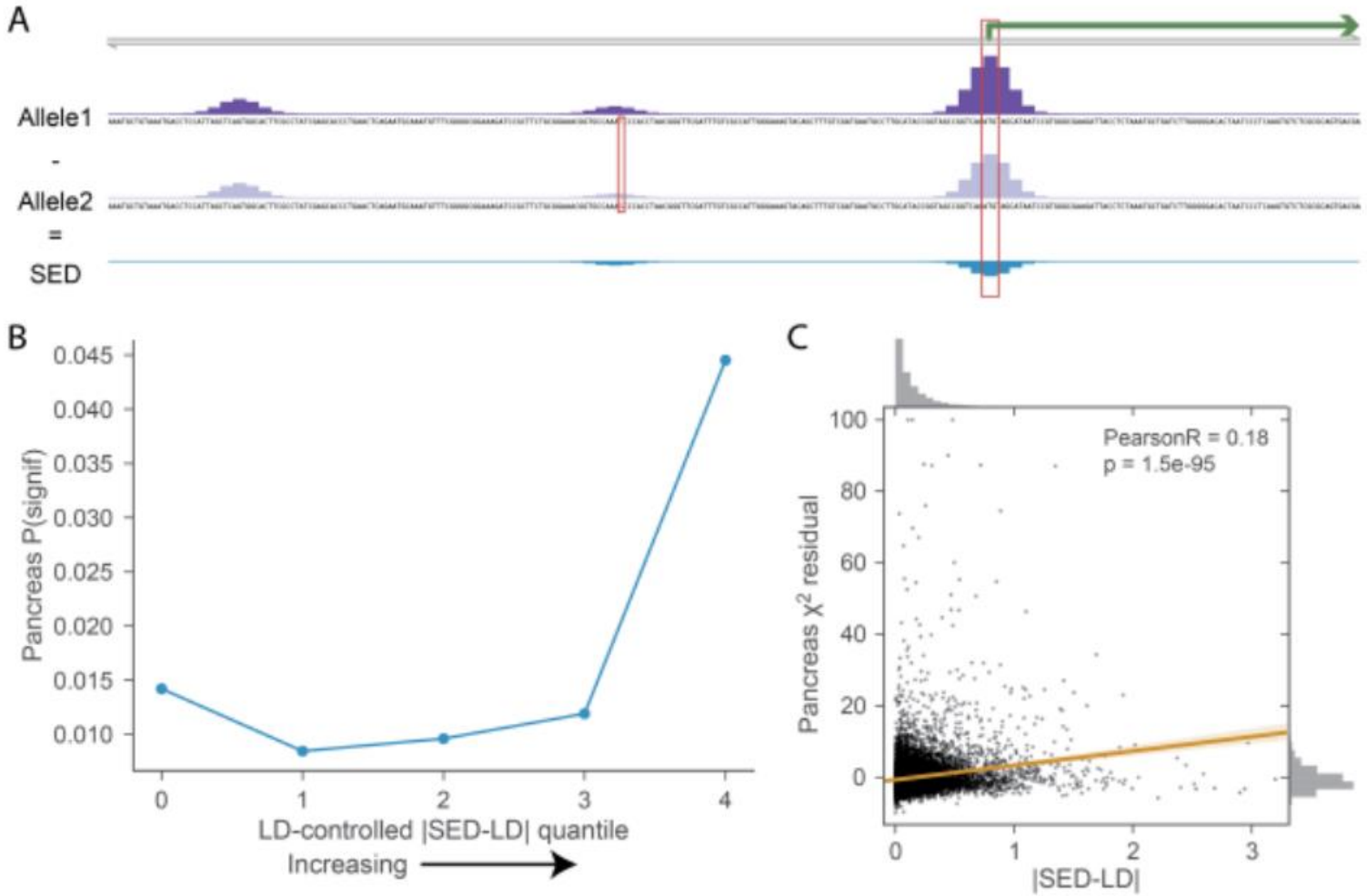
Compute gradient of output bin that matches TSS of a gene wrt. Each of the 131Kb / 128 bp input bins

131 KB / 128 bp outputs

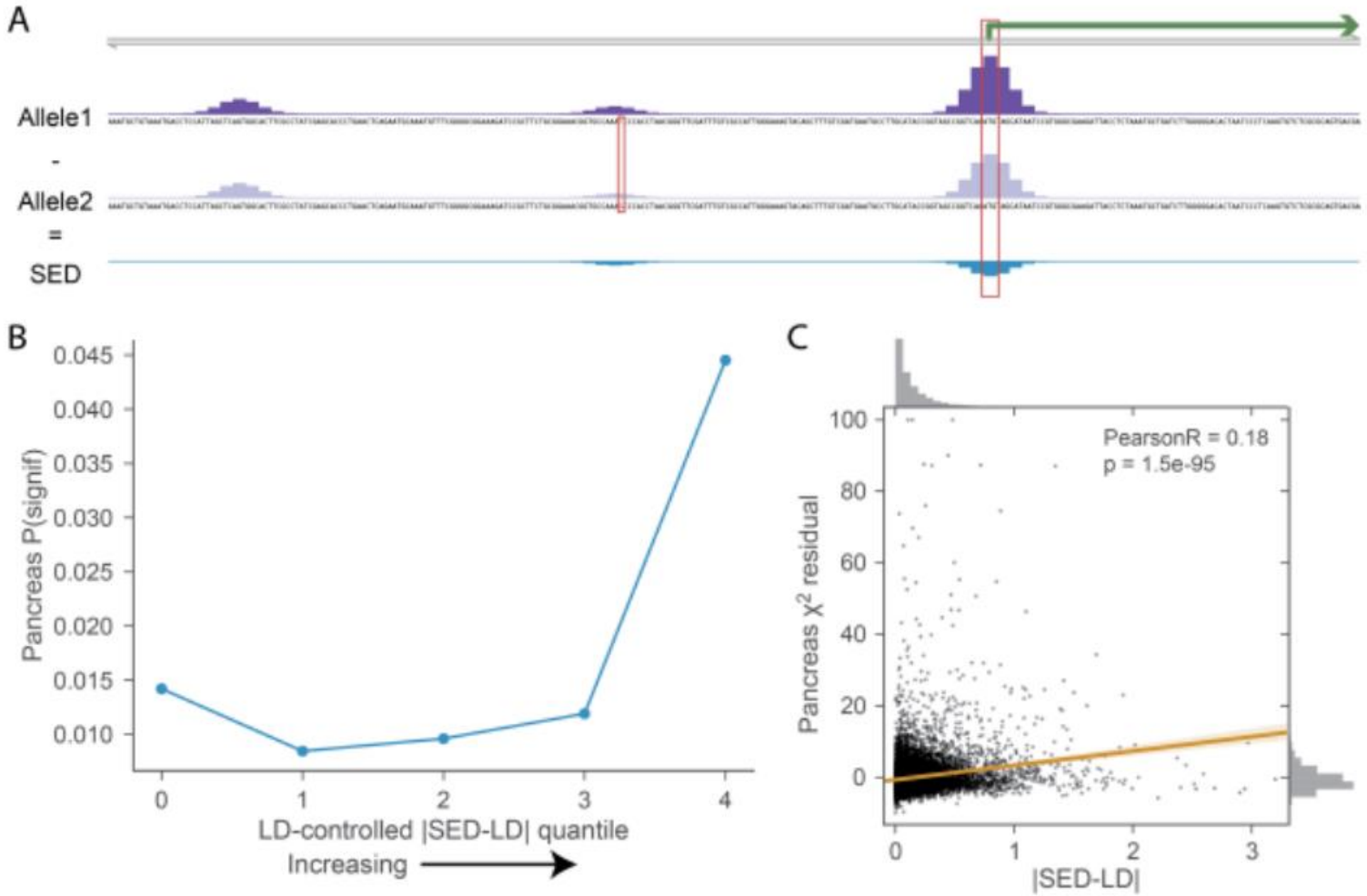
In-silico mutagenesis within each important bin to identify important nucleotides



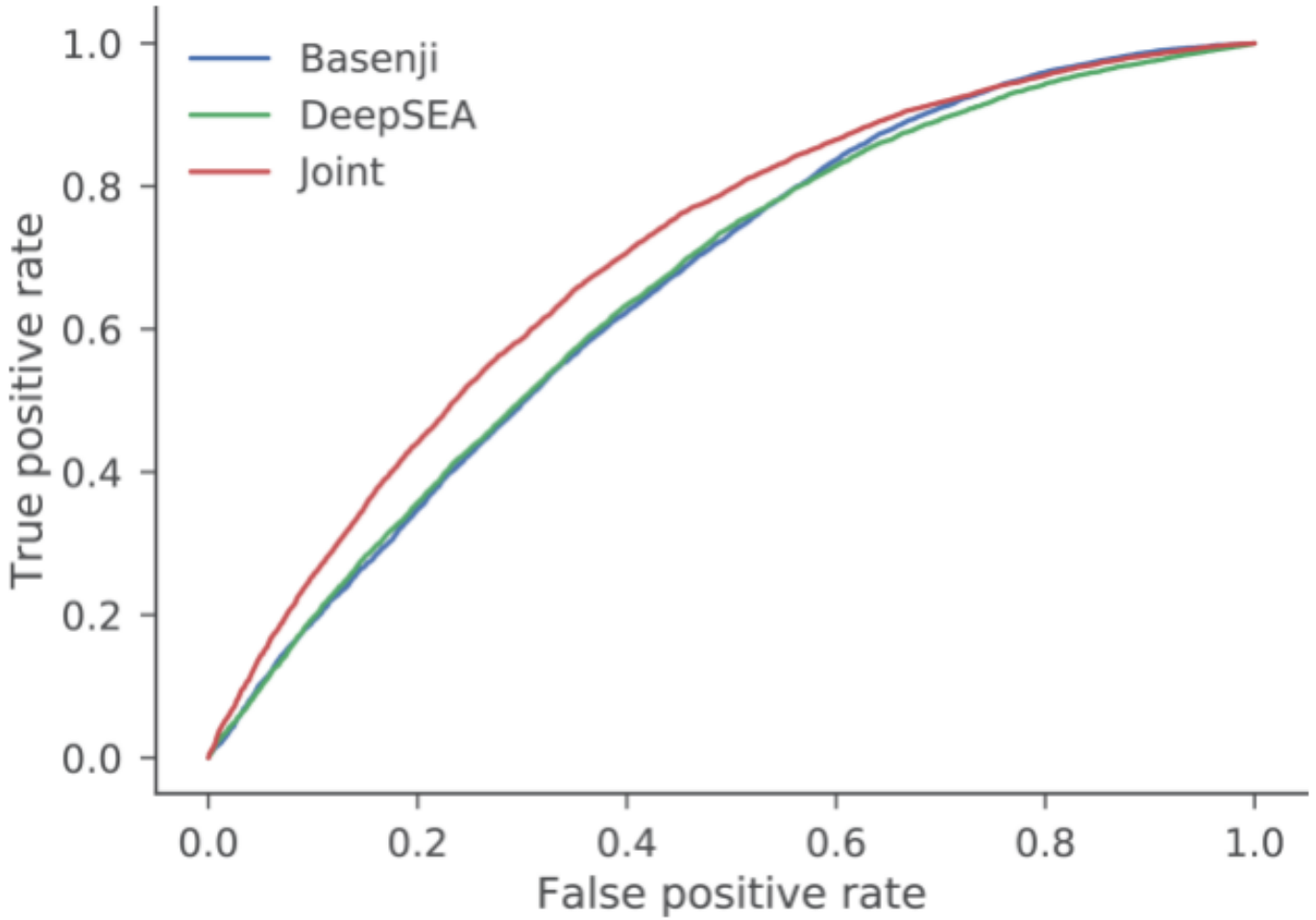
Predicting effects of distal genetic variants on expression (eQTLs)



Predicting effects of distal genetic variants on expression (eQTLs)



Predicting disease-associated variants



Supplementary Figure 8 – Basenji predictions exceed previous methods for GWAS classification.

Flawed set up of gold standard positive and negative set and flawed evaluation metric (auROC)

We computed SNP expression difference scores for a dataset containing 12,296 bi-allelic SNPs taken from the NIH GWAS Catalog database 24 and a negative set with matched minor allele frequency. We computed log2 fold changes between the predictions for the two alleles at each position in the surrounding region. We let the score for each SNP be the maximum of the absolute value of that fold change across the sequence. Finally, we reduced the dimensionality of the feature set to 200 with PCA and trained a logistic regression classifier to predict presence in the GWAS catalog. The DeepSEA authors previously computed predictions for this data using their method and conservation statistics in a more sophisticated model. Basenji-based scores match DeepSEA, and a joint model using both exceeds either one

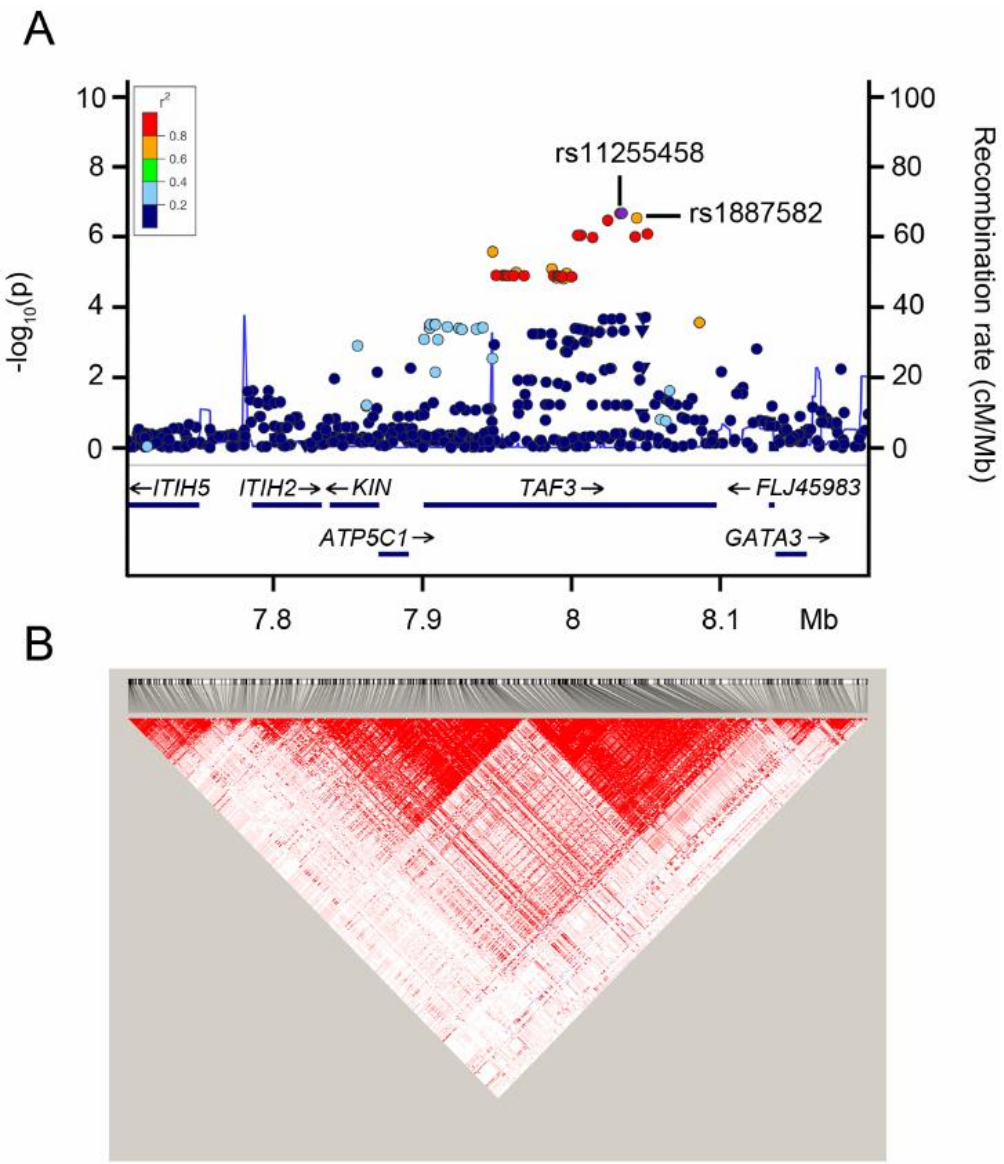
GWAS and the problem of linkage disequilibrium

<u>Positive set</u>
Lead or tag SNPs identified as statistically significant in a GWAS are more than often not the causal SNP. They more likely tag one or more causal SNPs in strong LD (correlated) with it.

<u>Negative set</u>
Balanced number of random SNPs only matched for minor allele frequency. No matching for distance, GC content and so many other confounders!

Also finding a causal SNP is a needle in a haystack problem. Highly unbalanced classification problem.

Using balanced set and auROC is flawed.



Predicting effect of disease-associated variant

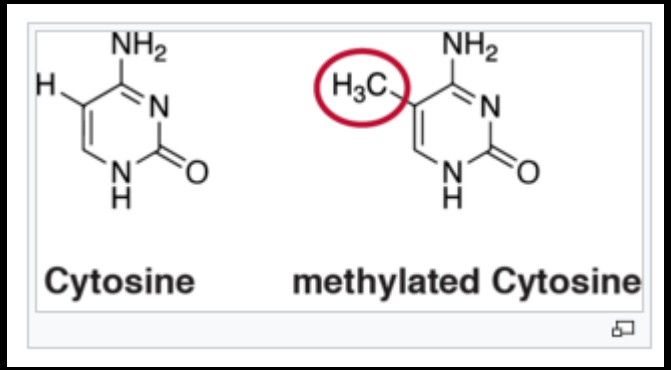
Figure 6 - Basenji gene-specific variant scores illuminate multiple sclerosis associated locus. Lead variant rs74796499 is associated with multiple sclerosis 53. Among the credible set of linked variants, Basenji predicts that rs78461372 would alter transcription of the nearby genes GPR65 and GALC. In immune cells, such as treated CD14+ cells depicted here, both genes are transcribed and the C>G introduces an ETS factor motif that enhances transcription. In contrast, in other cell types, e.g. insular cortex, where GPR65 is far less transcribed, Basenji predicts the same motifs play a role in repressing the gene.

Commentary

- Dilated convolutions are on the right track i.e. definitely want to model larger sequence contexts and connections between elements
- But dilated convolutions are a waste for long sequences with very sparse informative regions. Most the genome is not doing anything. So the dilated conv. are looking for a needle in the haystack.
- Combining the idea of dilated convolutions with some prior knowledge of important regions (masking regions that are definitely not regulatory) and their connections (leveraging 3D genome connectivity data)
- Alternative formulation is modeling the genome as a graph/network of connections and applying convolutions on graphs. You could cover much larger receptive field without wasting computation on non-regulatory regions in the genome

More on dilated convolutions

- http://www.inference.vc/dilated-convolutions-and-kronecker-factorisation/

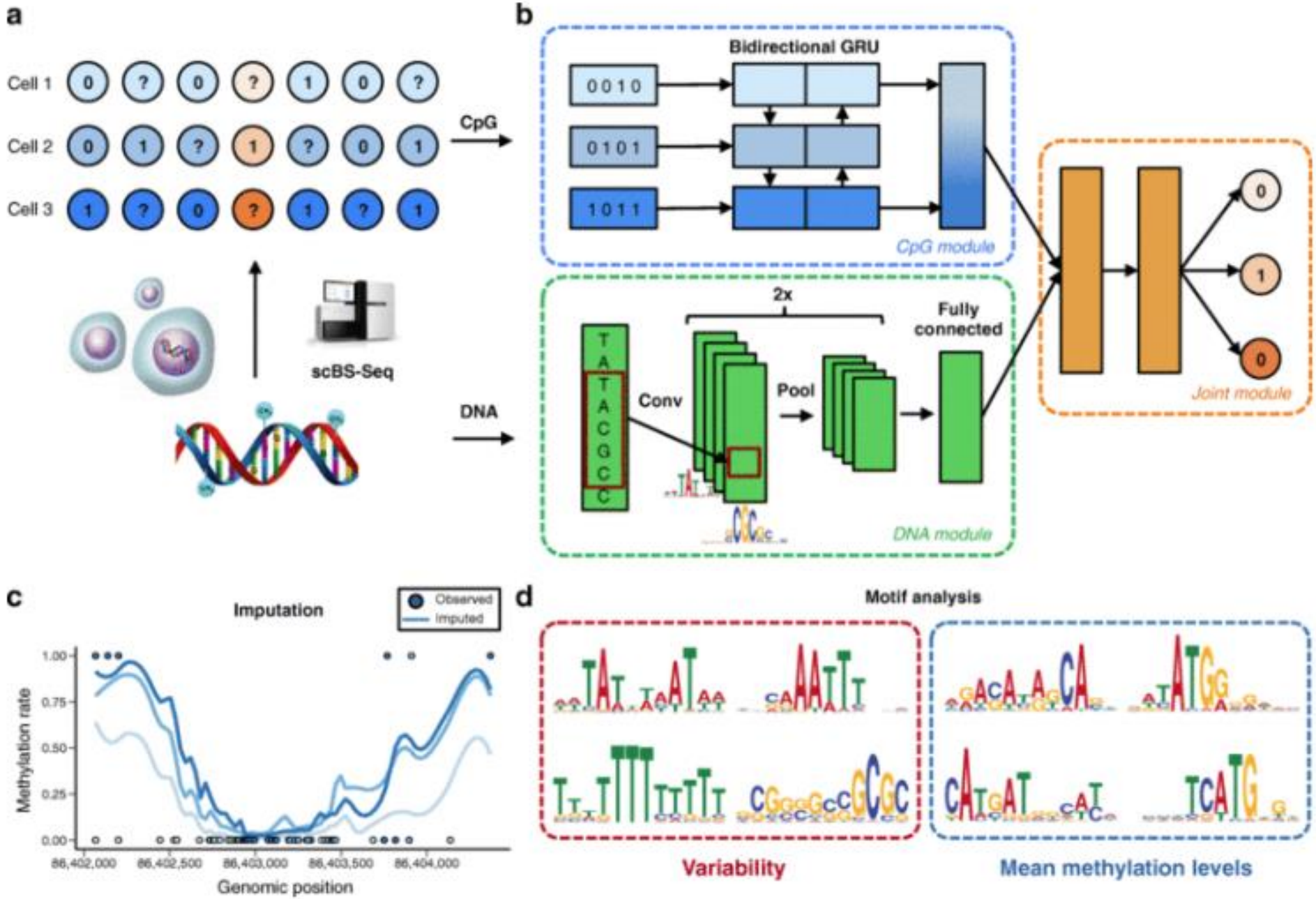


DeepCpG

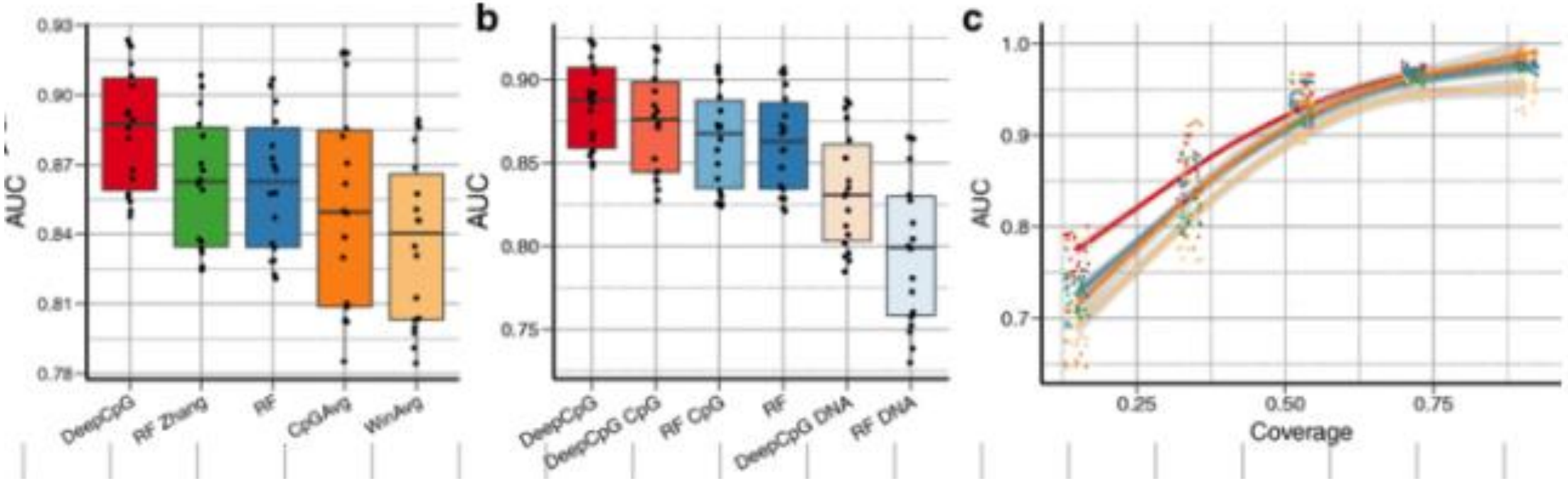
DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning

Christof Angermueller, Heather J. Lee, Wolf Reik and Oliver Stegle

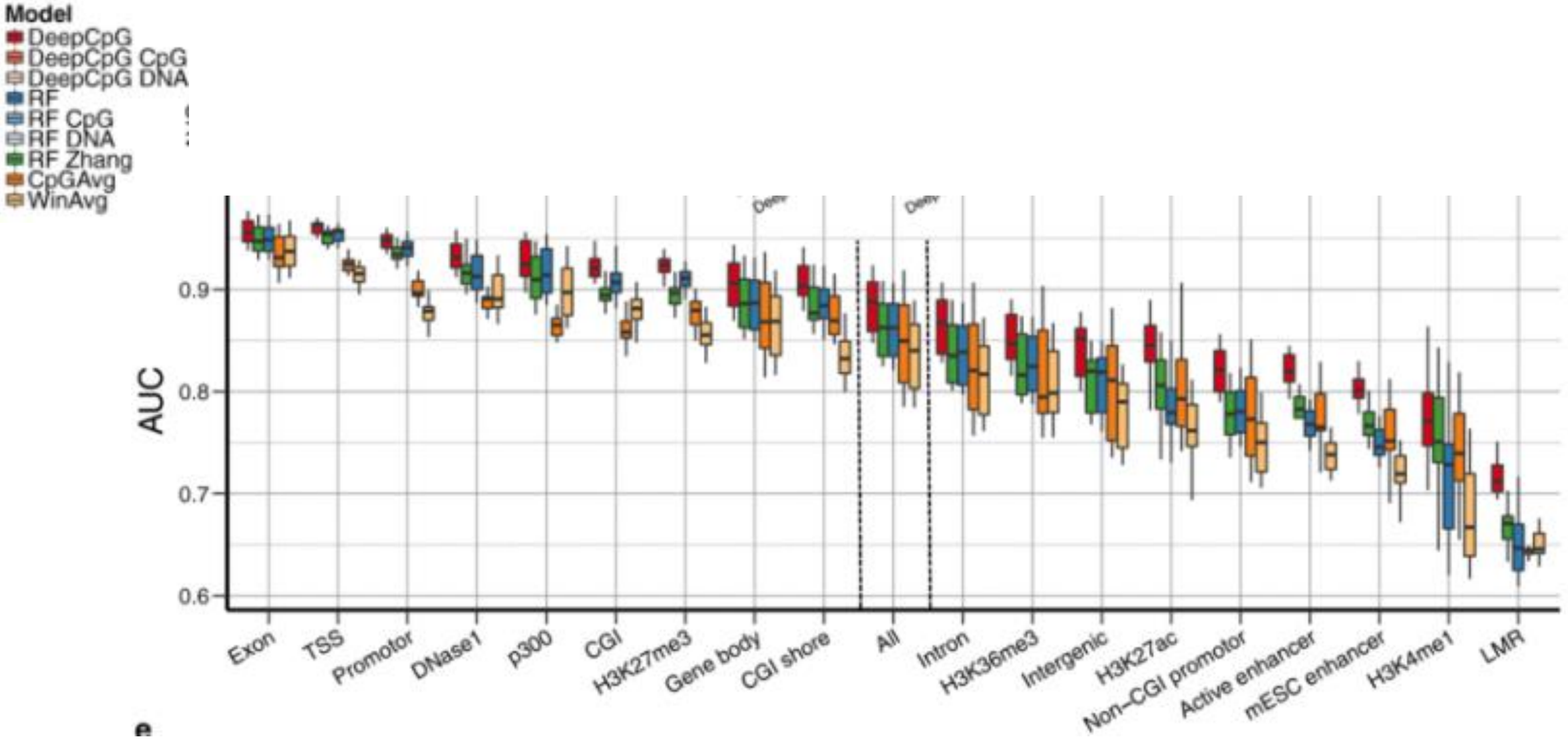
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1189-z


a
Cell 1
Cell 2
Cell 3
CpG
DNA
scBS-Seq
b
Bidirectional GRU
0010
0101
1011
CpG module
2x
Conv
Pool
Fully connected
DNA module
0
1
0
Joint module
c
Imputation
Observed
Imputed
Methylation rate
1.00
0.75
0.50
0.25
0.00
86,402,000
86,402,500
86,403,000
86,403,500
86,404,000
Genomic position
d
Motif analysis
Variability
Mean methylation levels

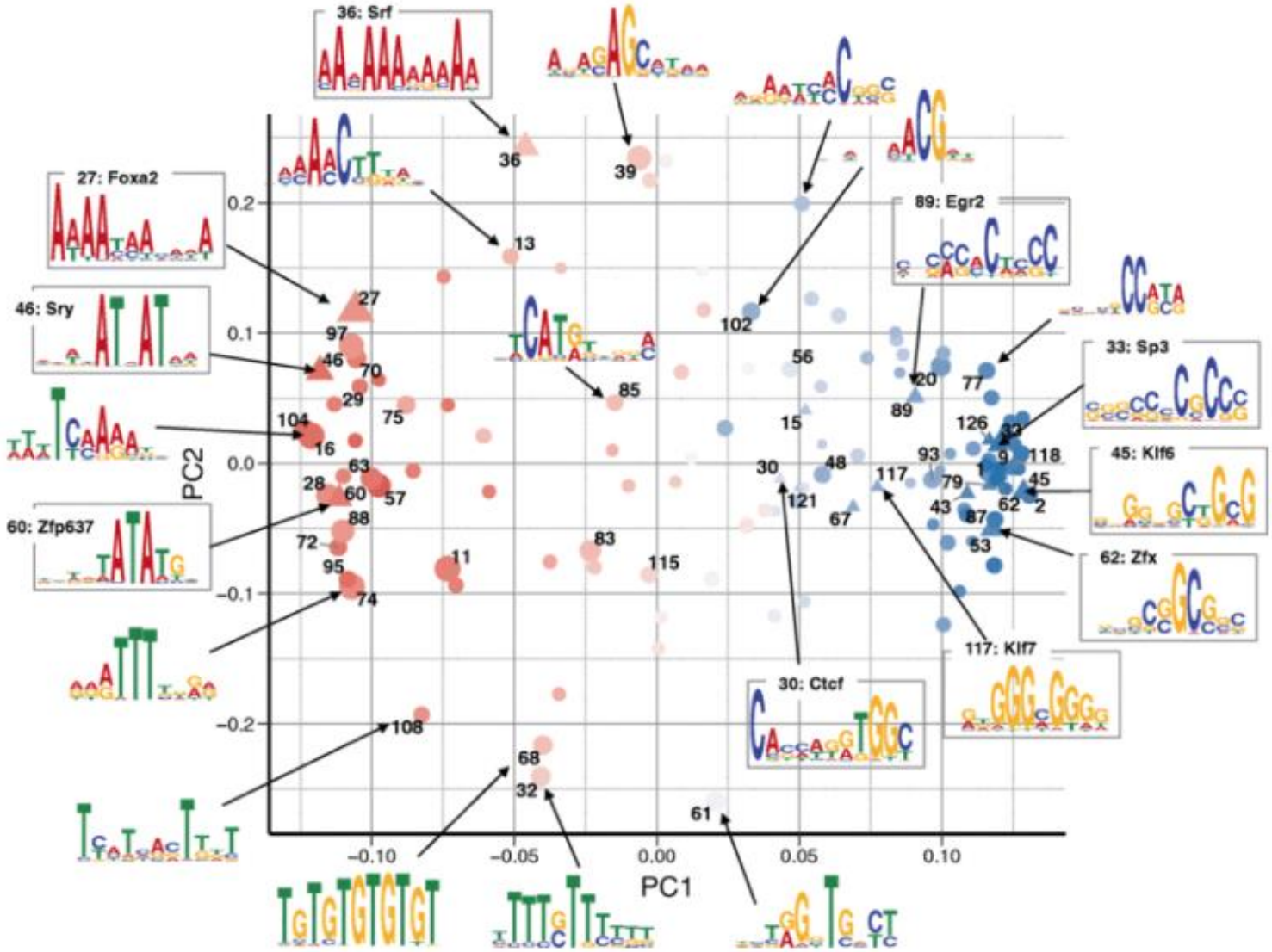
Prediction performance



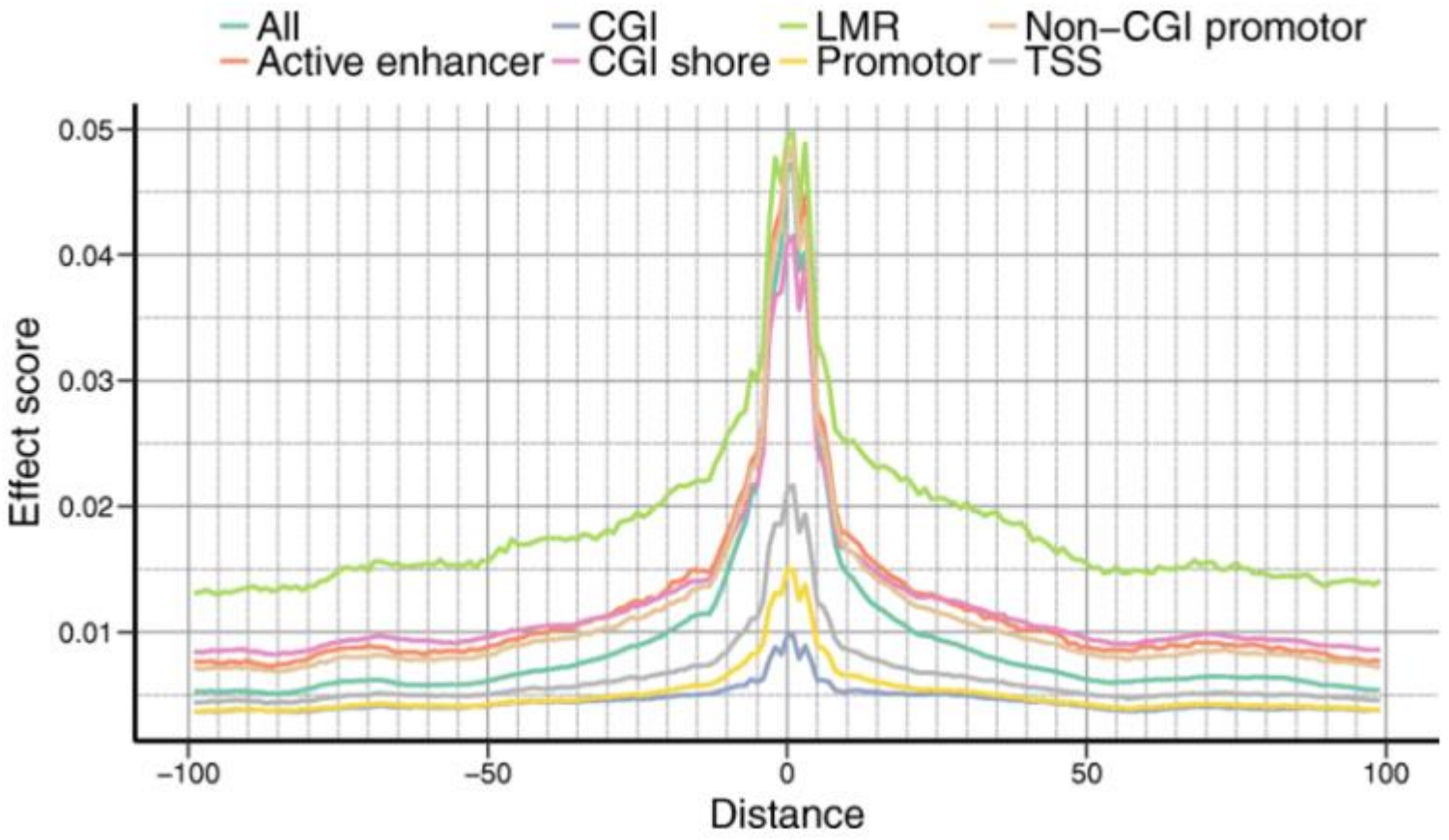
Prediction performance by annotation



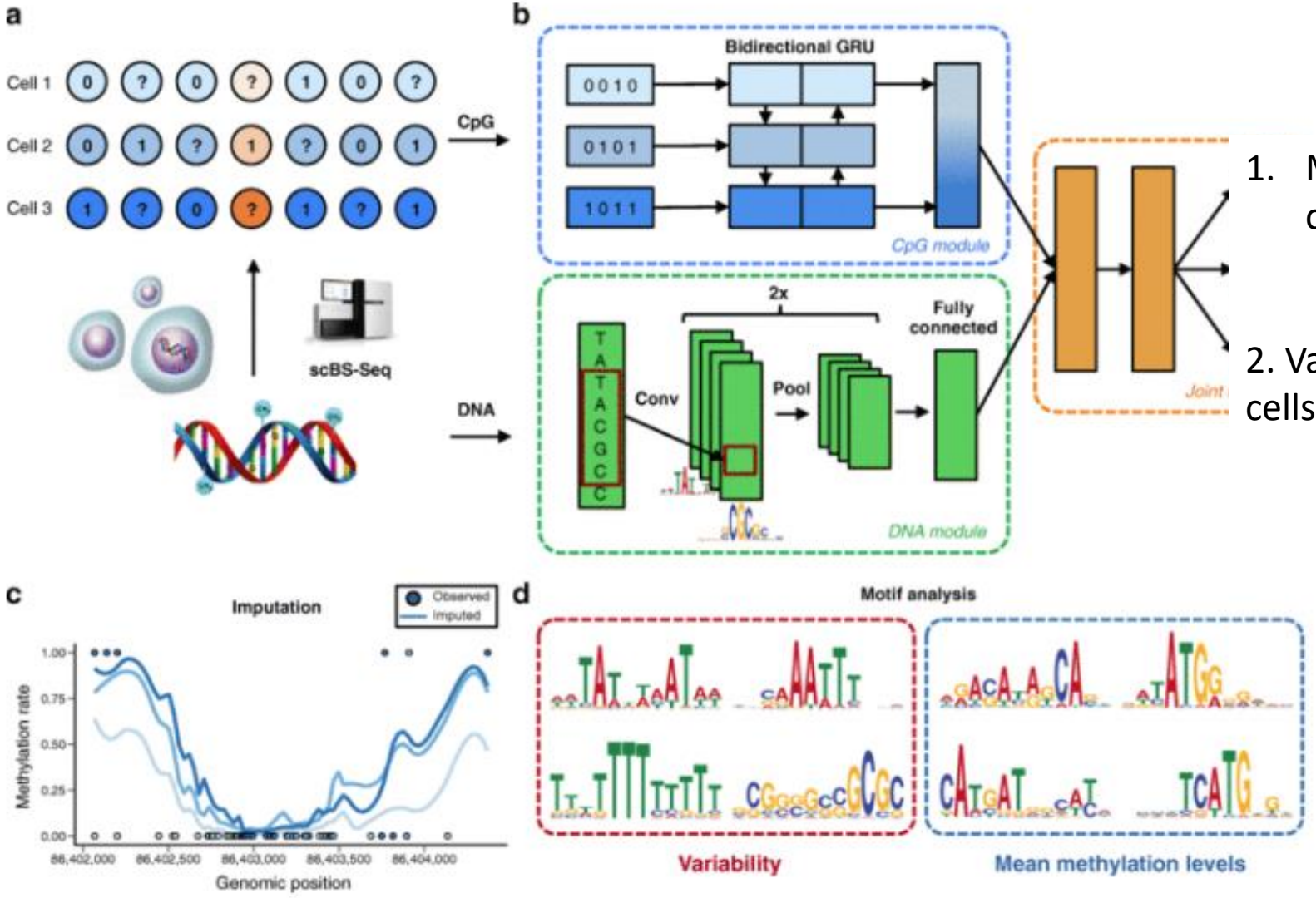
Interpreting learned sequence conv. filters based on occurrence/activation profile across sequences



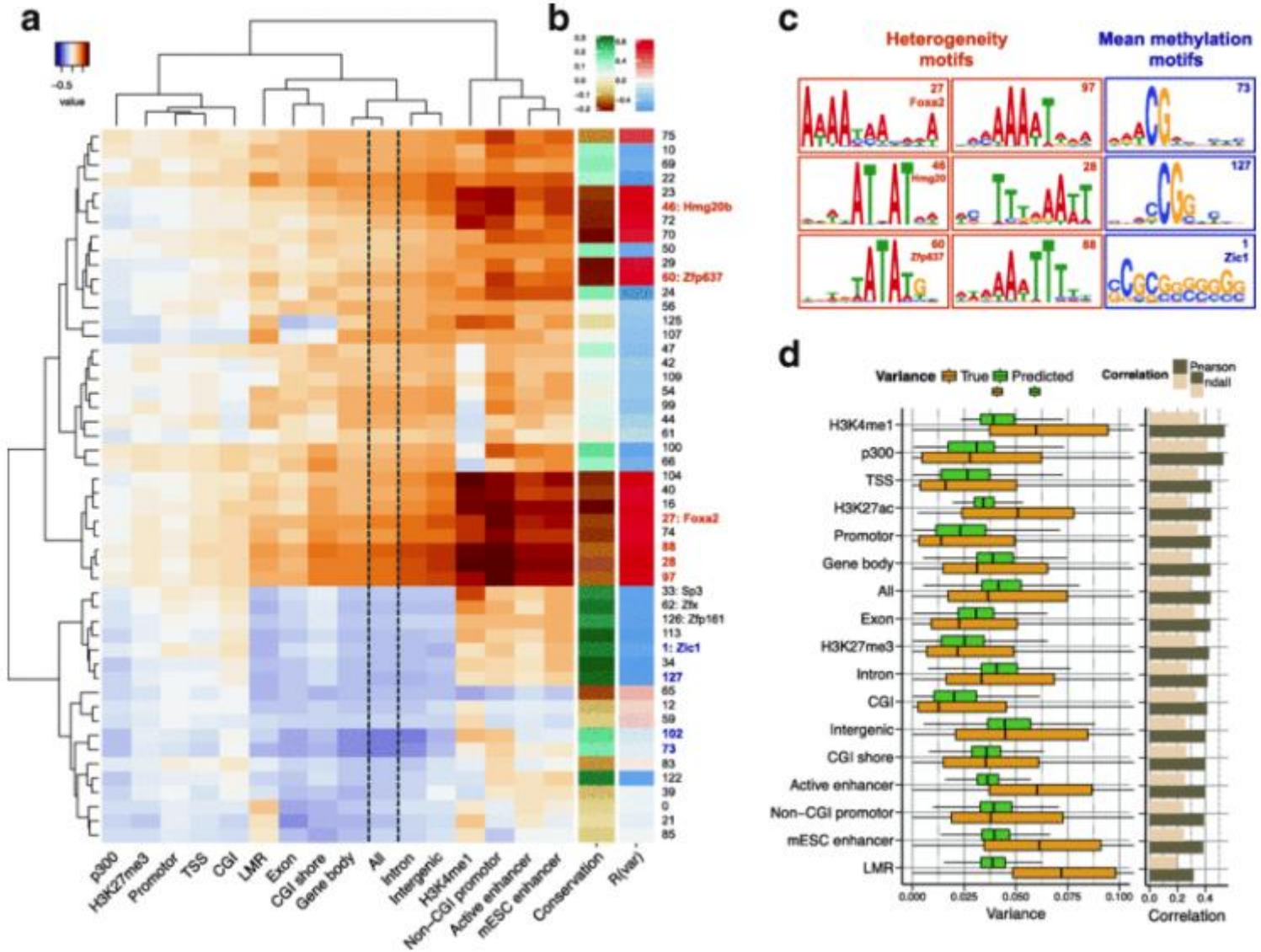
Saliency maps (gradients) to identify important sequence nucleotides affecting methylation



Predicting methylation variation across cells



1. Mean across cells
2. Variance across cells


a
b
c
d
Heterogeneity motifs
Mean methylation motifs
Variance True Predicted
Correlation Pearson Kendall
H3K4me1
p300
TSS
H3K27ac
Promotor
Gene body
All
Exon
H3K27me3
Intron
CGI
Intergenic
CGI shore
Active enhancer
Non−CGI promotor
mESC enhancer
LMR
Variance
Correlation
p300
H3K27me3
Promotor
TSS
CGI
LMR
Exon
CGI shore
Gene body
All
Intron
Intergenic
H3K4me1
Non−CGI promotor
Active enhancer
mESC enhancer
Conservation
R(var)

Evaluation of different strategies for training multi-task CNNs for genomics

Daniel Kim, Anshul Kundaje

(Unpublished)

Multi-task CNN

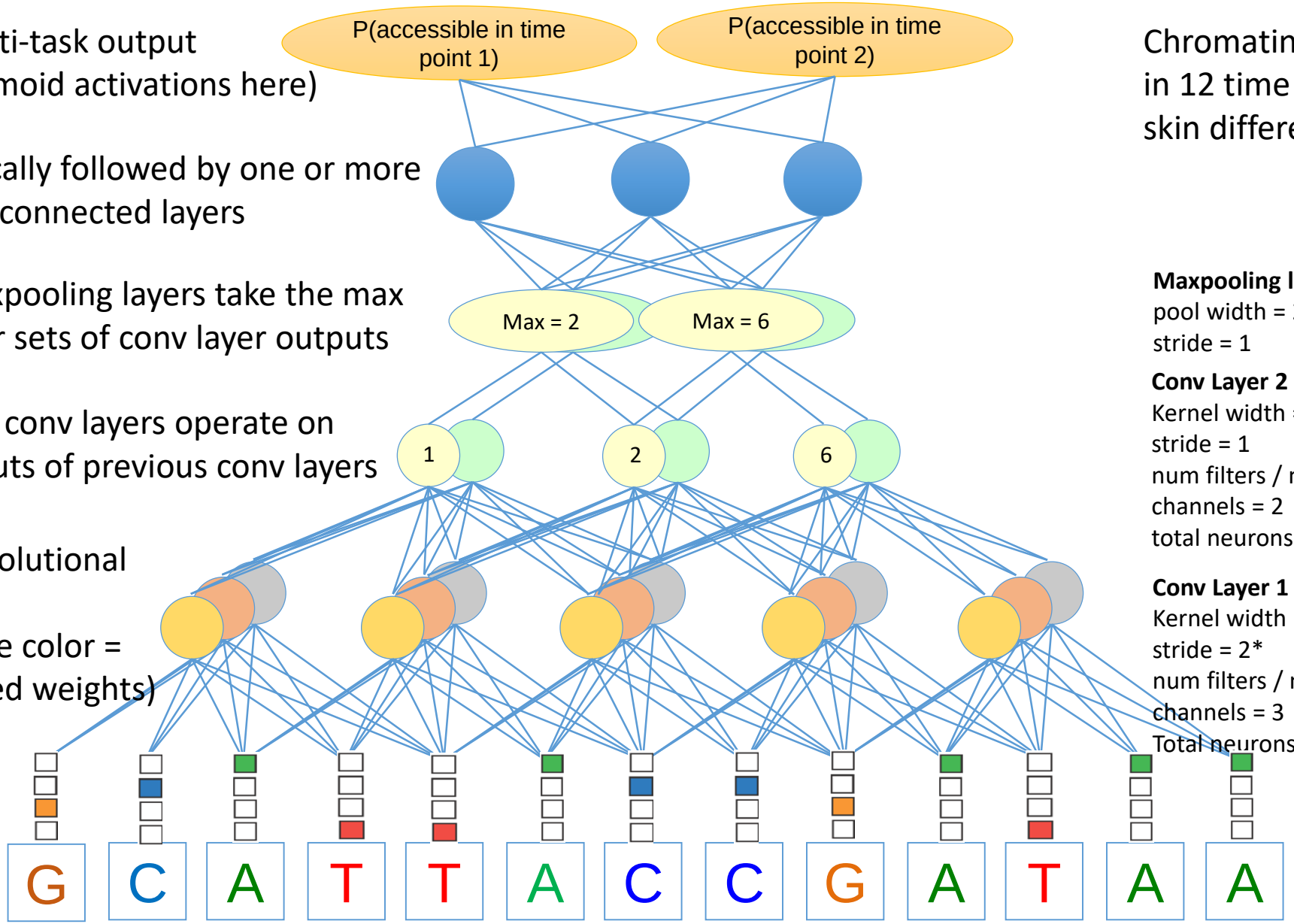
Multi-task output
(sigmoid activations here)

Typically followed by one or more fully connected layers

Maxpooling layers take the max over sets of conv layer outputs

Later conv layers operate on outputs of previous conv layers

Convolutional layer
(same color = shared weights)



Chromatin accessibility in 12 time points of skin differentiation

Maxpooling layer
pool width = 2
stride = 1

Conv Layer 2
Kernel width = 3
stride = 1
num filters / num channels = 2
total neurons = 6

Conv Layer 1
Kernel width = 4
stride = 2*
num filters / num channels = 3
Total neurons = 15

Choice of negative set

Table 1. Evaluation (AUPRC) of models trained with various negatives on datasets with different negative sets. Utilizing all negatives in training performs best across all datasets.

	Dataset: flank+DHS negatives only	**Dataset: random negatives only**	**Dataset: flank+DHS+ random negatives**
Training: flank+DHS negatives only	0.689321941465	0.803997392913	0.729479546005
Training: random negatives only	0.758205073644	0.829909381892	0.738491462625
Training: flank+DHS+random negatives	0.796838207242	0.83426273962	0.763176061265

Training chromosome-wide i.e. greatest diversity of negatives is most beneficial

Due to massive class imbalance you have to be careful how u create minibatches and weight examples

Augmenting with reverse complements

Table 2. Evaluation of models trained with or without reverse complement examples.

	AUROC	AUPRC	Recall at 5% FDR	Stopping Epoch
Without reverse complement	0.825404682283	0.748647678451	0.0795828995245	6
With reverse complement	0.840122696586	0.759827109164	0.078223558554	1

Minor improvements in accuracy but much faster training i.e. fewer epochs

Multi-task vs. single task

Table 3. Evaluation of models trained as single task or multi task models.

	AUROC	AUPRC	Recall at 5% FDR
Single task RF	0.714453756809	0.604013800621	0.000
Multi task RF	0.625453233719	0.431953579187	0.000
Single task CNN	0.825404682283	0.748647678451	0.0795828995245
Multi task CNN	0.819421813005	0.695408869799	0.0370161131853
Multi task CNN + differential	0.823156578034	0.681588440151	0.000

Naïve hard multi-task architecture where tasks split on final layer actually does WORSE than single task models

Need better multi-task architectures and training regimens that allow all gradients to optimize all tasks

Different types of architectures

Table 4. Evaluation of various model architectures.

	AUROC	AUPRC	Recall at 5% FDR	Model parameter total
CNN	0.825404682283	0.748647678451	0.0795828995245	5,969,201
CNN-RNN	0.843663621744	0.757844232883	0.000625715294193	3,942,961
Resnet	0.781753843879	0.689007260145	0.000	23,501,825

CNN-RNNs do about the same as CNNs. CNN-LSTMs have been claimed to do better.
Default ResNet architecture inherited from computer vision fails. Can't just transfer bonafide architectures across domains.

Transfer learning from massive reference data

Table 5. Evaluation of pre-training on ENCODE-Roadmap datasets and transfer learning.

	AUROC	AUPRC	Recall at 5% FDR
Single task CNN	0.825404682283	0.748647678451	0.0795828995245
Single task CNN pretrained	0.881574801742	0.821648594593	0.220124341705
Multi task CNN	0.819421813005	0.695408869799	0.0370161131853
Multi task CNN pretrained	0.869301980505	0.76648201909	0.118550785444
Multi task + differential CNN pretrained	0.869734772637	0.75949538282	0.109308876846

- Train a model on over 1800 tasks from reference datasets (like BASSET and DeepSEA papers)
- Initialize single or multi-task models on the skin differentiation tasks with these models
- Compare to de-novo learned single and multi-task skin models

- Multi-task pretrained models do MUCH better than denovo multi-task models
- Single-task pretrained models do the best