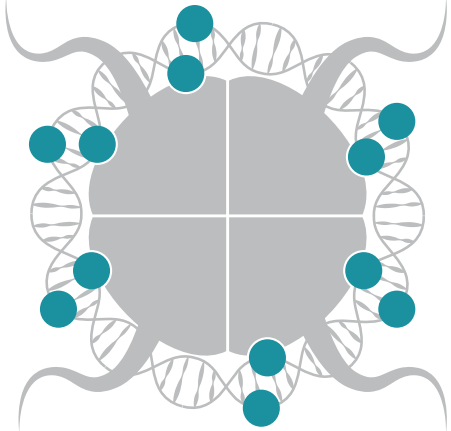
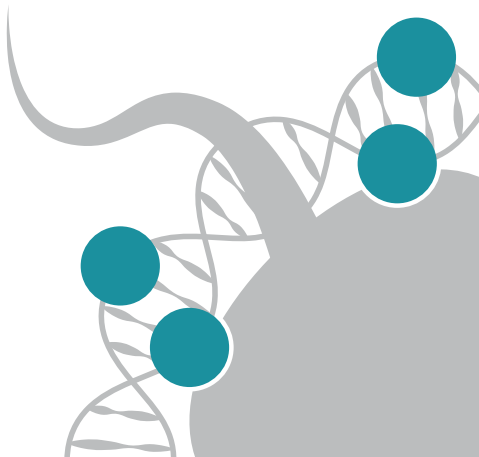


DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning

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

Genome Biology, 18(1), 67. doi:10.1186/s13059-017-1189-z

Presented by Fabian Müller

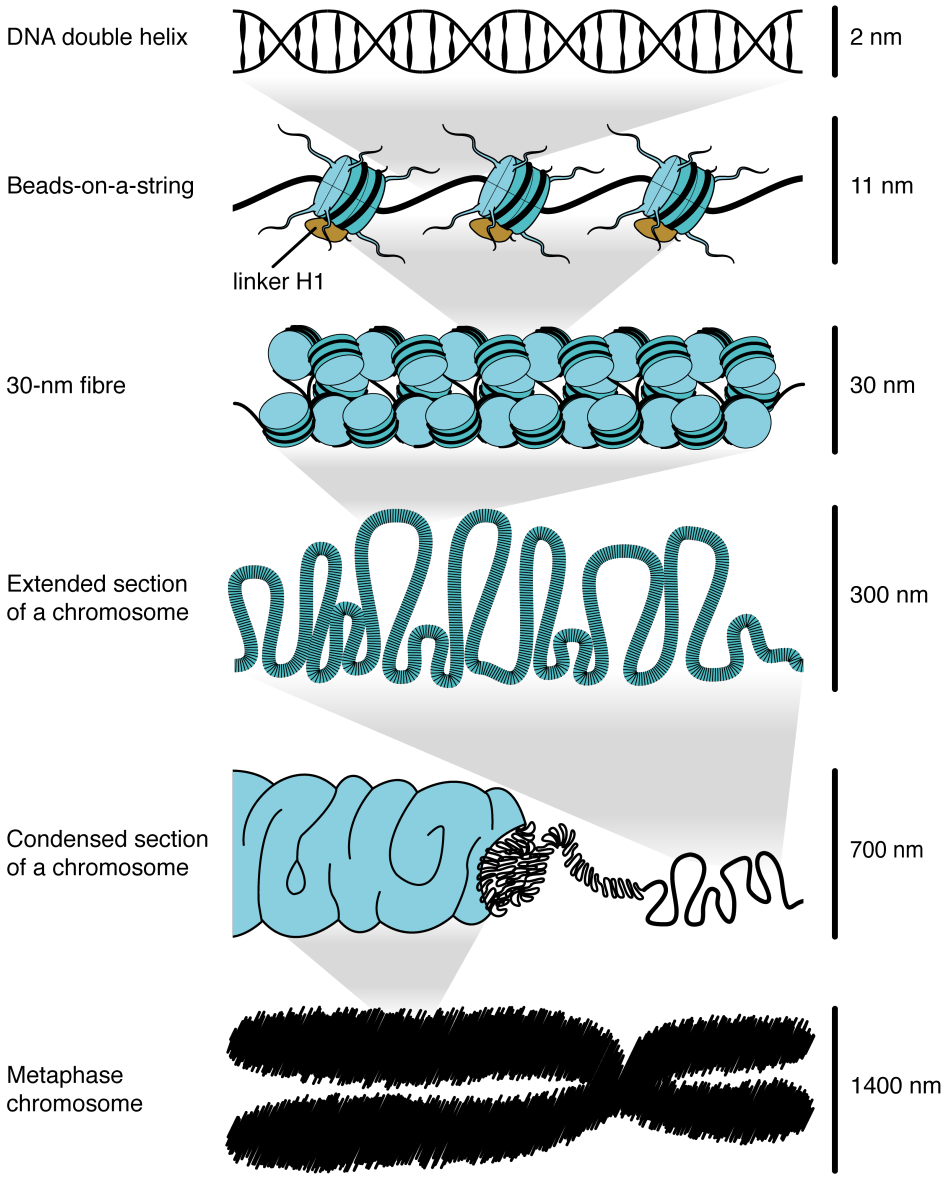


max planck institut informatik

Outline

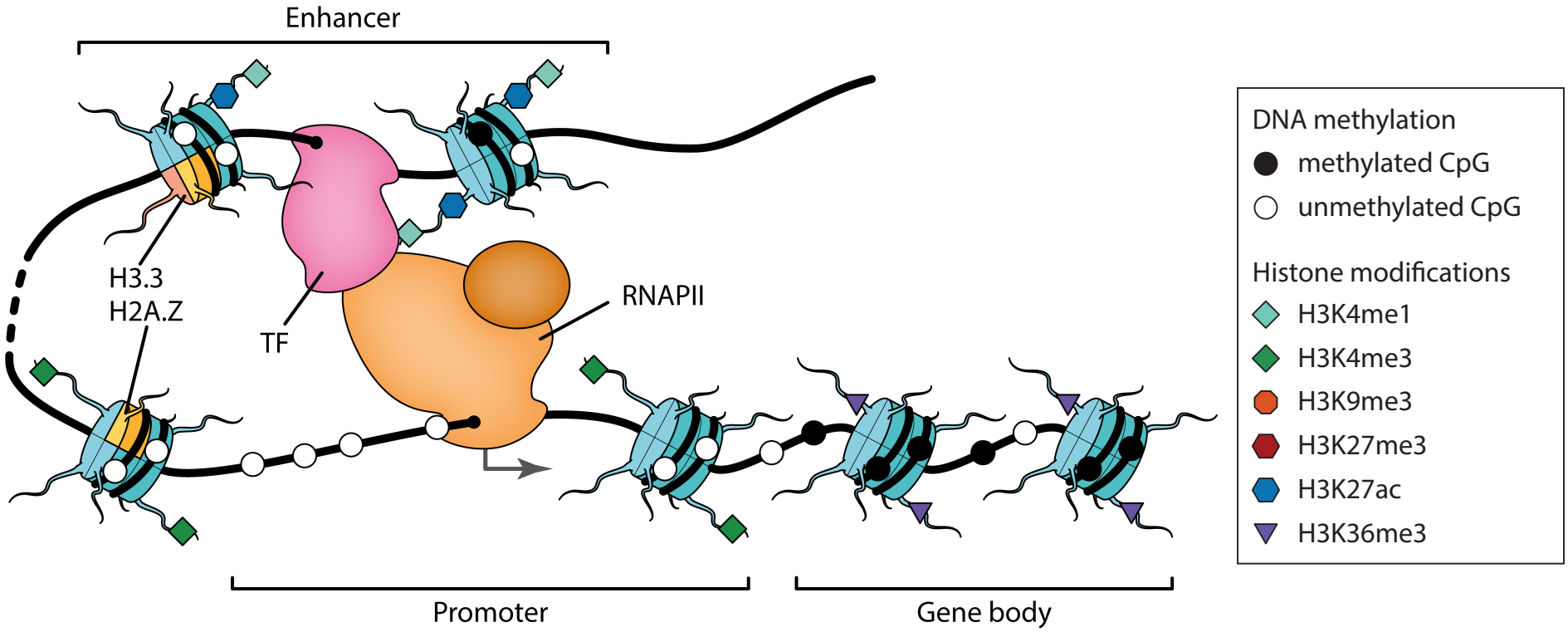
- Biological background
- DeepCpG model
  - Architecture
  - Validation
  - Interpretation
- Summary

Chromatin Organization



- Hierarchical organization
- Protein scaffold
  - Nucleosome: DNA wrapped around histone proteins
- Accessibility regulates gene expression
- Tightly regulated by a plethora of molecular factors

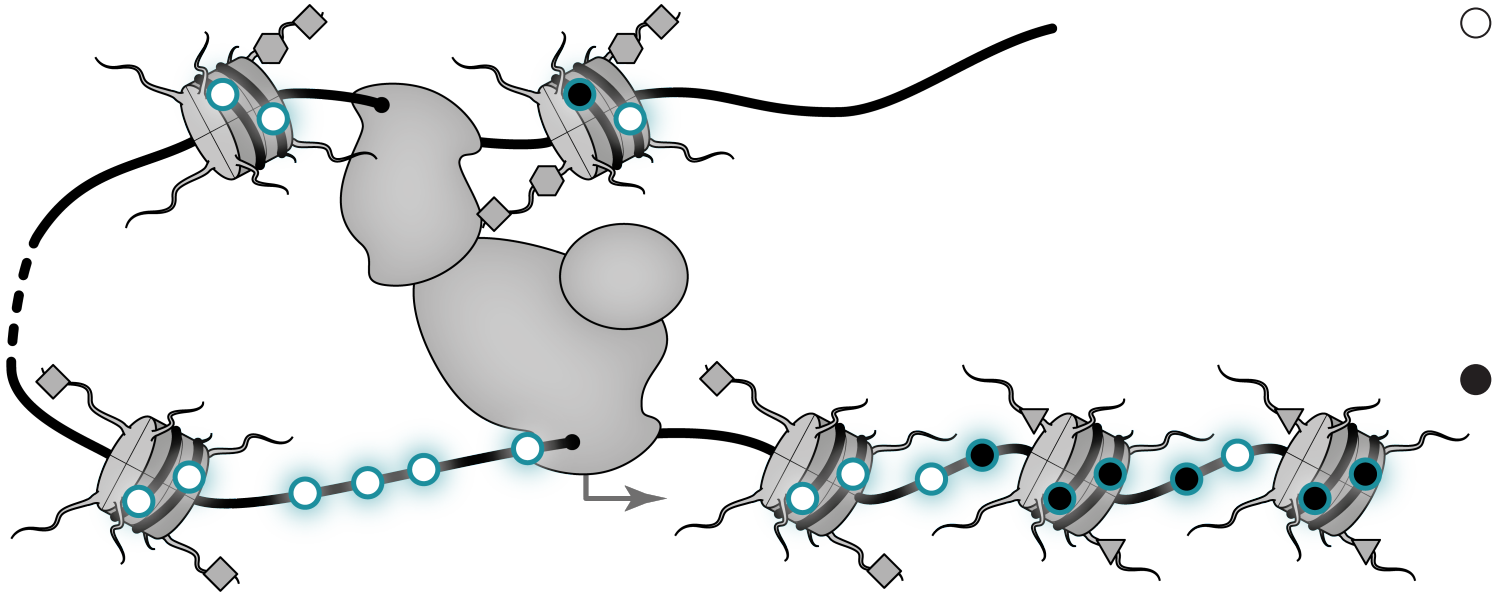
Epigenetic Regulation

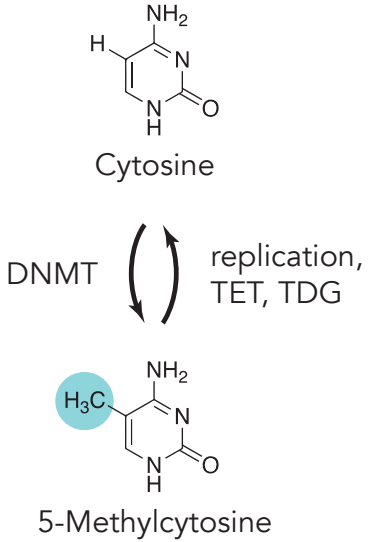


- Accessibility regulates gene expression
- Regulation beyond promoter regions
  - Enhancers, gene bodies, …
- 80.4% of the genome is attributed with regulatory roles [1]

[1] ENCODE Project Consortium (2012). *Nature*, 489(7414), 57–74

DNA Methylation

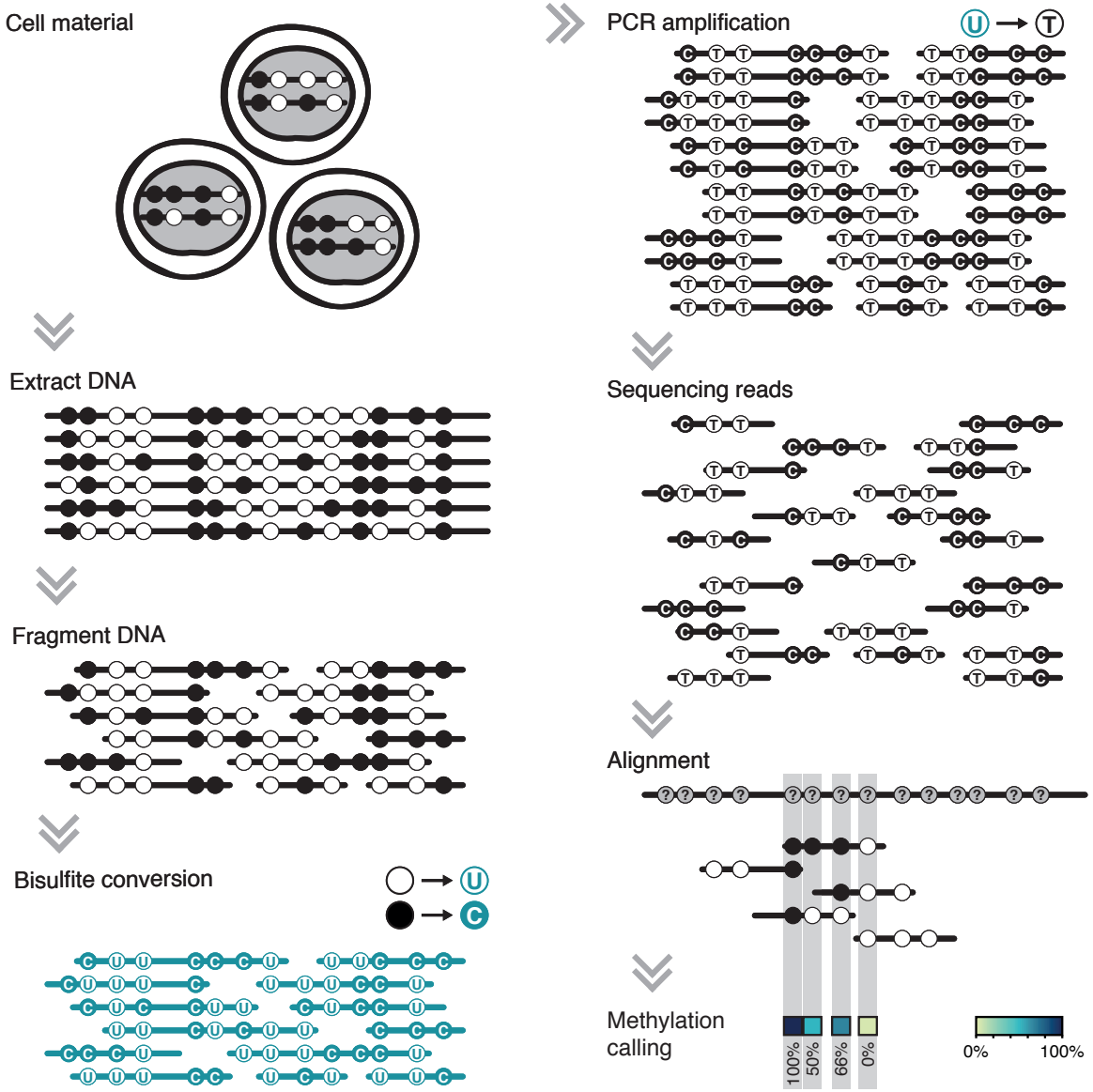




- Occurs predominantly in CpG dinucleotides in mammalian genomes
- Associated with regulation of gene expression

- Deregulated in disease
- Genome-wide quantification:
  - Bisulfite sequencing
  - Methylation arrays
  - Enrichment-based sequencing
  - …

Bisulfite Sequencing

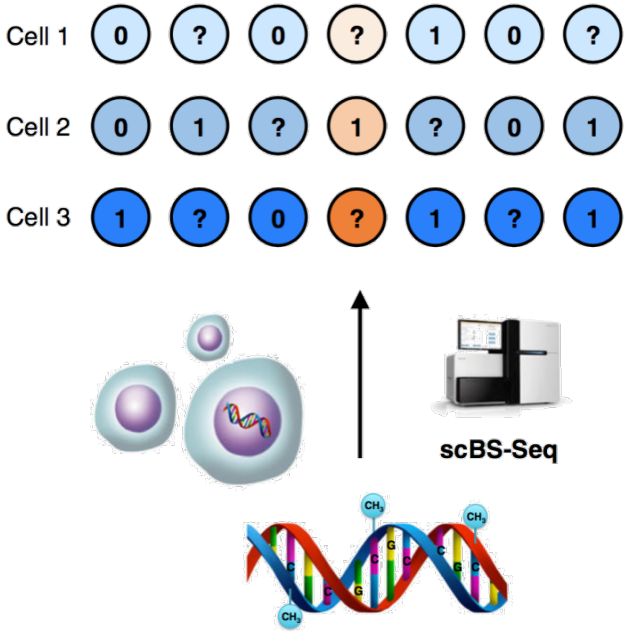


Prediction Task

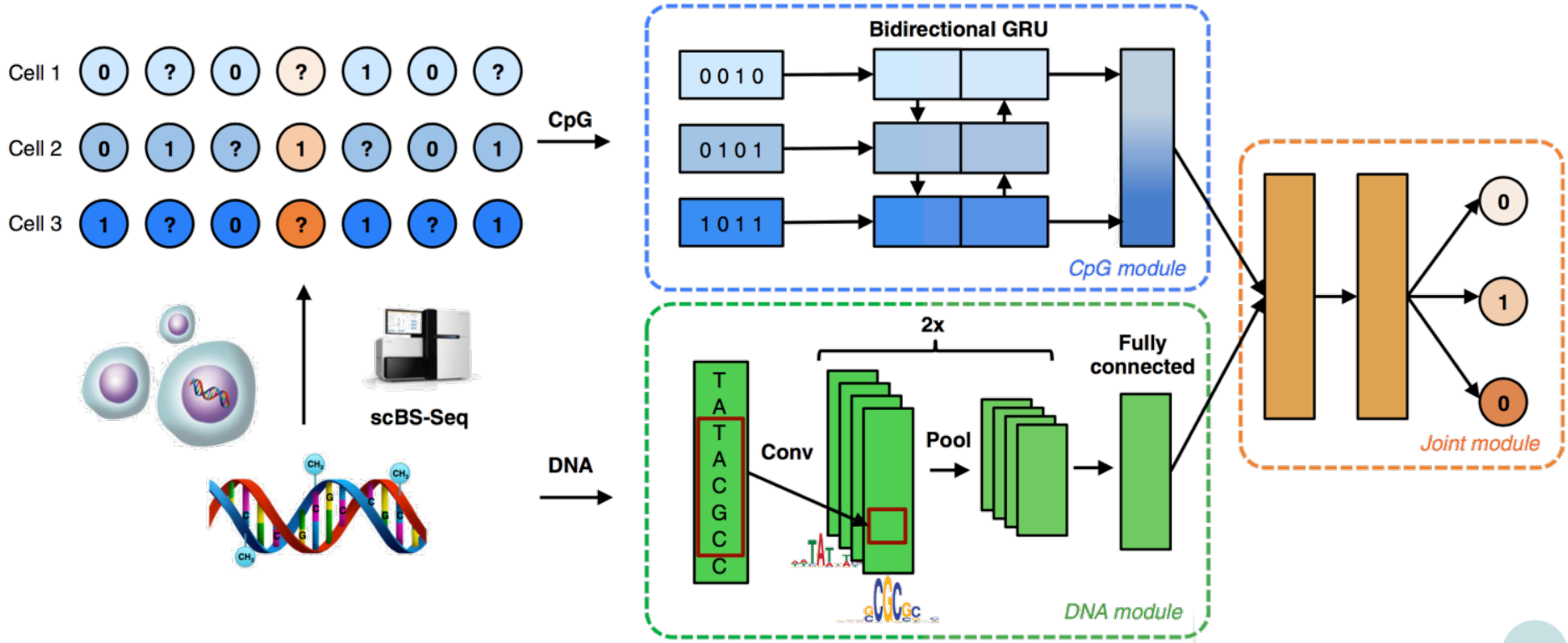
- Problem: single-cell bisulfite sequencing data is sparse

⇒ **Predict missing CpG methylation states in single cells**

- Output: methylation state (binary) for each CpG and cell
- Features:
  - Local DNA sequence
  - Observed CpG methylation



DeepCpG Model Architecture



- Combining CpG methylation and sequence features using different modules

DNA module

- Convolutional network using position weight matrices
- Input: 1001bp DNA sequence windows, centered at each target CpG (*n*): :

$$s_n \in \{0,1\}^{1001\times 4}$$

$$\begin{aligned} A &= [1,\ 0,\ 0,\ 0] \\ T &= [0,\ 1,\ 0,\ 0] \\ G &= [0,\ 0,\ 1,\ 0] \\ C &= [0,\ 0,\ 0,\ 1] \end{aligned}$$

- 1st convolution:

filter length

$$a_{nfi} = \mathrm{ReLU}\left(\sum_{l=1}^{L}\sum_{d=1}^{D} w_{fld}\, s_{n,i+l,d}\right)$$

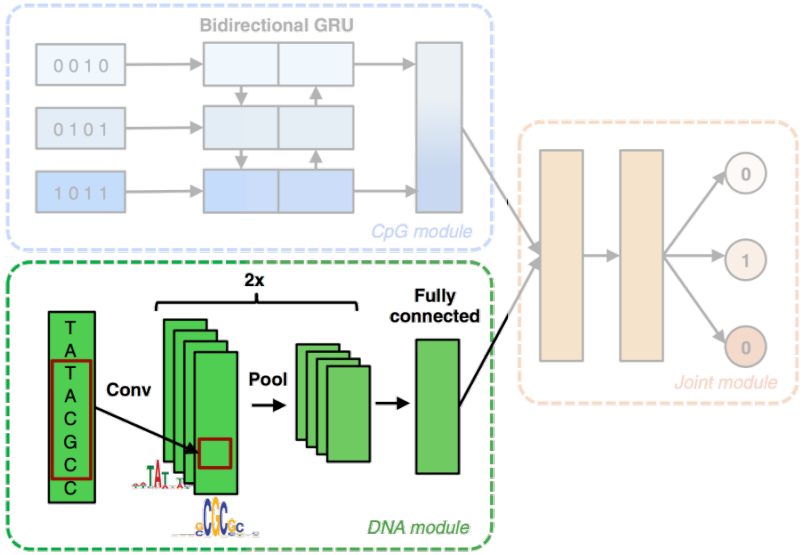
multiple filters *f*

with $\mathrm{ReLU}(x) = \max(0,\ x)$

- Max-pooling layer:

$$p_{nfi} = \max_{|k|<P/2}\left(a_{nf,i+k}\right)$$

step size for non-overlapping pooling



- Multiple convolution-pooling layers capture sequence-motif interactions
  - Number selected on validation set
- One final fully connected layer (e.g. 128 nodes, ReLU activation)

CpG Module

- Bidirectional gated recurrent network (GRU)
- Input: methylation states and relative distances for the 25 left and right neighbor CpGs of each target CpG in each of T cells:

$$x_1, \ldots, x_T \in (\{0,1\}^{50}, [0,1]^{50}) \subset [0,1]^{100}$$

- Embedding layer combines input and captures interactions for each cell:

$$\bar{x}_t = \mathrm{ReLU}(W_{\bar{x}} \cdot x_t + b_{\bar{x}}) \cdot$$
$$\in \mathbb{R}^{256}$$

- GRU layers (forward+backward) capture cell interactions :
  - Reset gate:

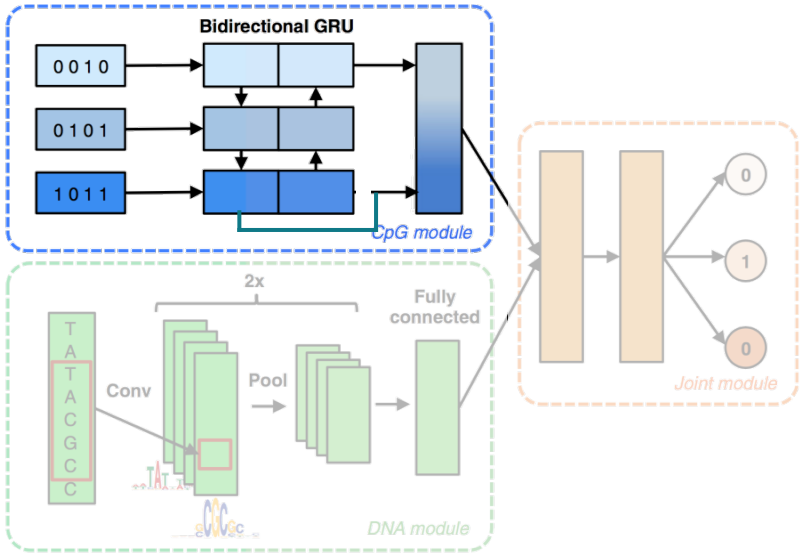
$$r_t = \mathrm{sigmoid}(W_{r\bar{x}} \cdot \bar{x}_t + W_{rh} \cdot h_{t-1} + b_r)$$

  - Update gate:

$$u_t = \mathrm{sigmoid}(W_{u\bar{x}} \cdot \bar{x}_t + W_{uh} \cdot h_{t-1} + b_u)$$

  - Hidden state:

$$h_t = (1-u_t) \odot h_{t-1} + u_t \odot \tilde{h}_t \cdot \longleftarrow \tilde{h}_t = \tanh\left(W_{\tilde{h}\bar{x}} \cdot \bar{x}_t + W_{\tilde{h}h} \cdot (r_t \odot h_{t-1}) + b_{\tilde{h}}\right)$$



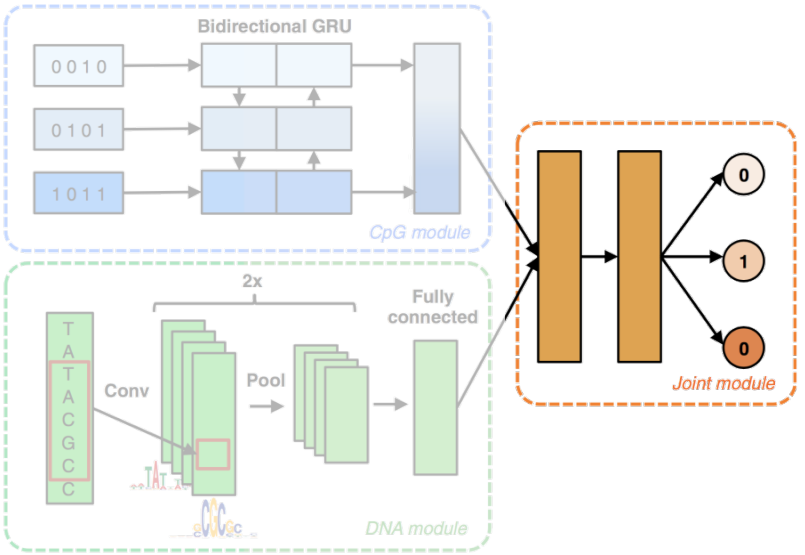
- Ouput: concatenation of forward and backward GRUs' last hidden state (512-dimensional)

Joint Module

- Combine outputs of DNA and CpG modules
- Two fully-connected layers (512 nodes, ReLU activation)
- Output: methylation state for each CpG *n* and cell *t*:

$$\hat{y}_{nt}(x) \in [0,1]$$

$$\hat{y}_{nt}(\mathrm{x}) = \mathrm{sigmoid}(\mathrm{x}) = \left(\frac{1}{1+\mathrm{e}^{-\mathrm{x}}}\right)$$



Model Training

- Minimize loss:

weight decay

$$L(w) = \mathrm{NLL}_w(\hat{y}, y) + \lambda_2 \|w\|_2$$

meth. State observed?

$$\mathrm{NLL}_w(\hat{y}, y) = -\sum_{n=1}^{N}\sum_{t=1}^{T} o_{nt}[y_{nt}\log(\hat{y}_{nt}) + (1-y_{nt})\log(1-\hat{y}_{nt})].$$

- Dropout with different rates for DNA, CpG and Joint module
- Optimization using mini-batch stochastic gradient descent
- Learning rate decay
- Early stopping: terminate if validation loss did not improve

- Hyper-parameter optimization: random sampling on validation set
- Training time:
  - DNA module: 24h
  - CpG module: 12h
  - Joint module: 4h
- Implemented in Python using `Theano` and `KERAS` [1]

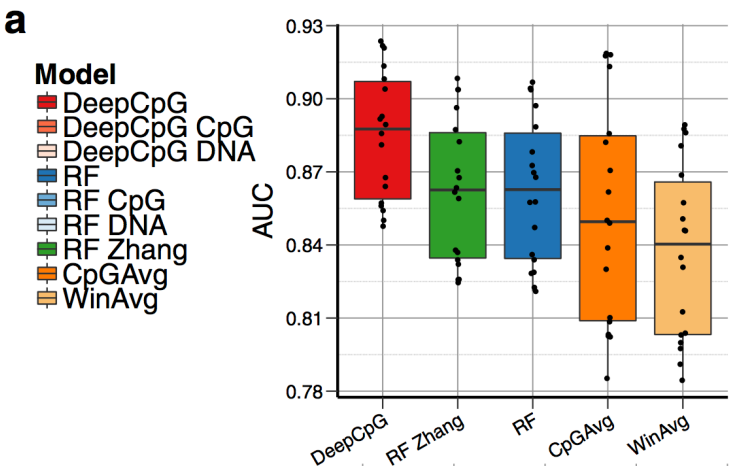
[1] https://github.com/PMBio/deepcpg

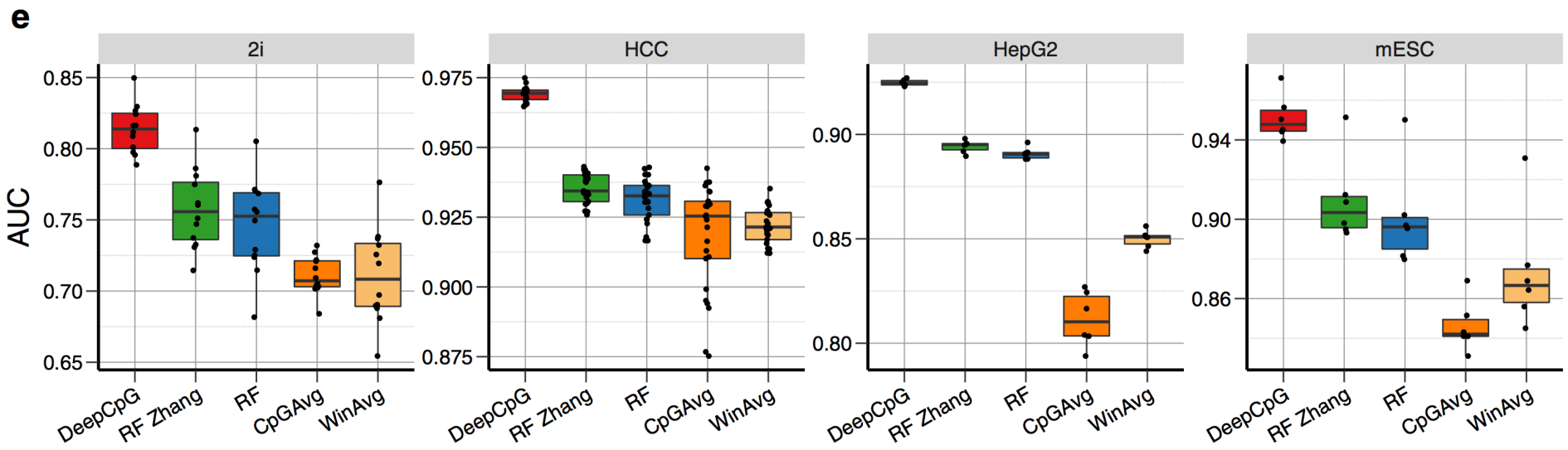
Performance Evaluation

- Comparison with
  - Genomic window averaging (*WinAvg*)
  - Averaging across cells (*CpGAvg*)
  - Random forest classifier using sequence and neighboring CpG states (*RF*)
  - Random forest classifier using comprehensive annotations (*RF Zhang*)
- Holdout validation
  - Different chromosomes in training, test and validation set

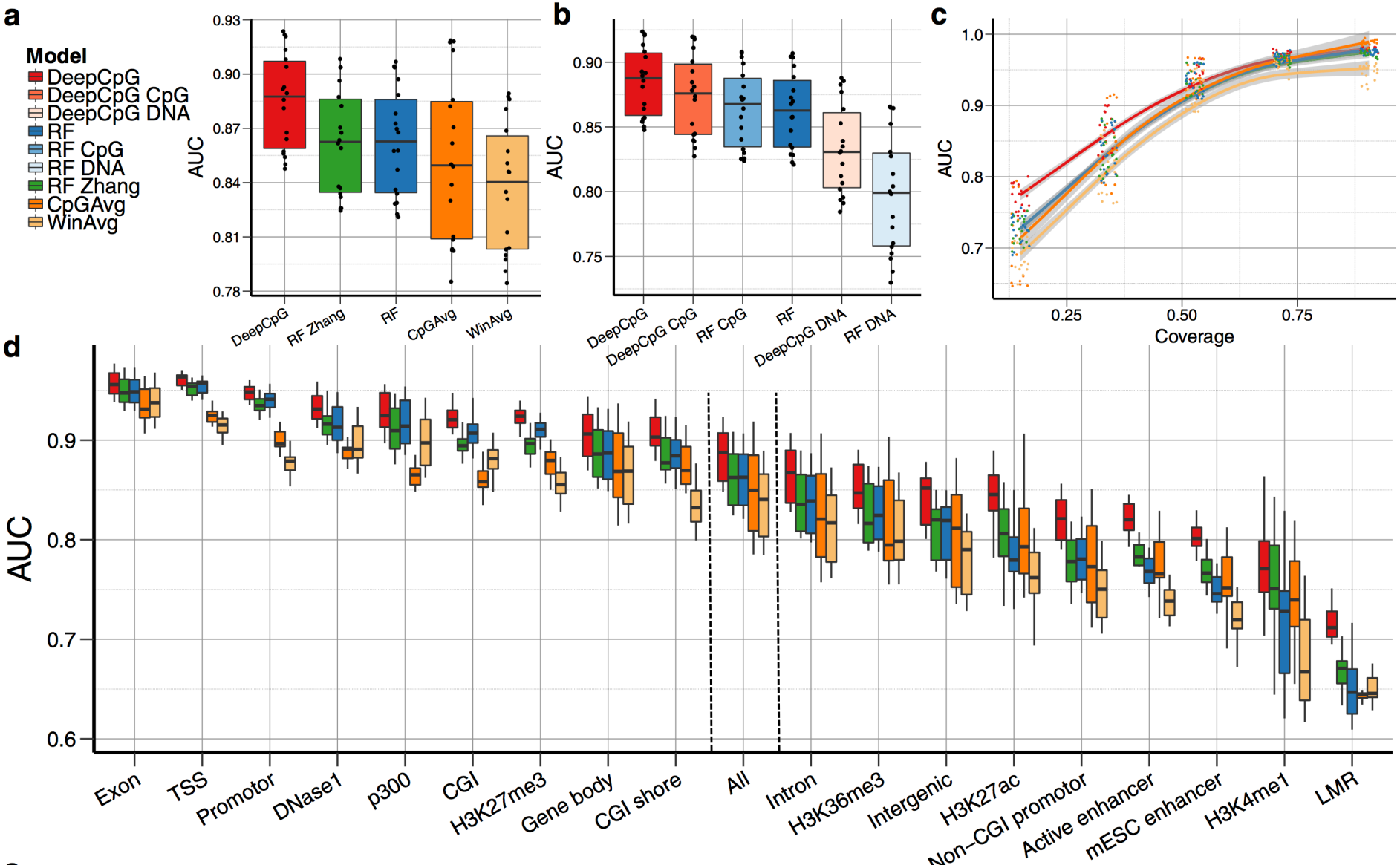
- Datasets:
  - Mouse Embryonic Stem Cells (mESCs; 18 serum, 12 2i medium; scWGBS)
  - 6 mESCs (scRRBS)
  - 25 human hepatocellular carcinoma (HCC; scRRBS)
  - 6 HepG2 (scRRBS)

Performance Evaluation





Performance Evaluation



Model Interpretability

- Motif analysis of DNA module
- For each filter: select sequence windows with

$$a_{nfi} > 0.5 \ \max_{ni}(a_{nfi})$$

filter length

$$a_{nfi} = \mathrm{ReLU}\left(\sum_{l=1}^{L}\sum_{d=1}^{D} w_{fld} s_{n,i+l,d}\right)$$

multiple filters *f*

sequence indicator matrix

- Filter activity: average weighted mean sequence activity $\bar{a}_{nf}$

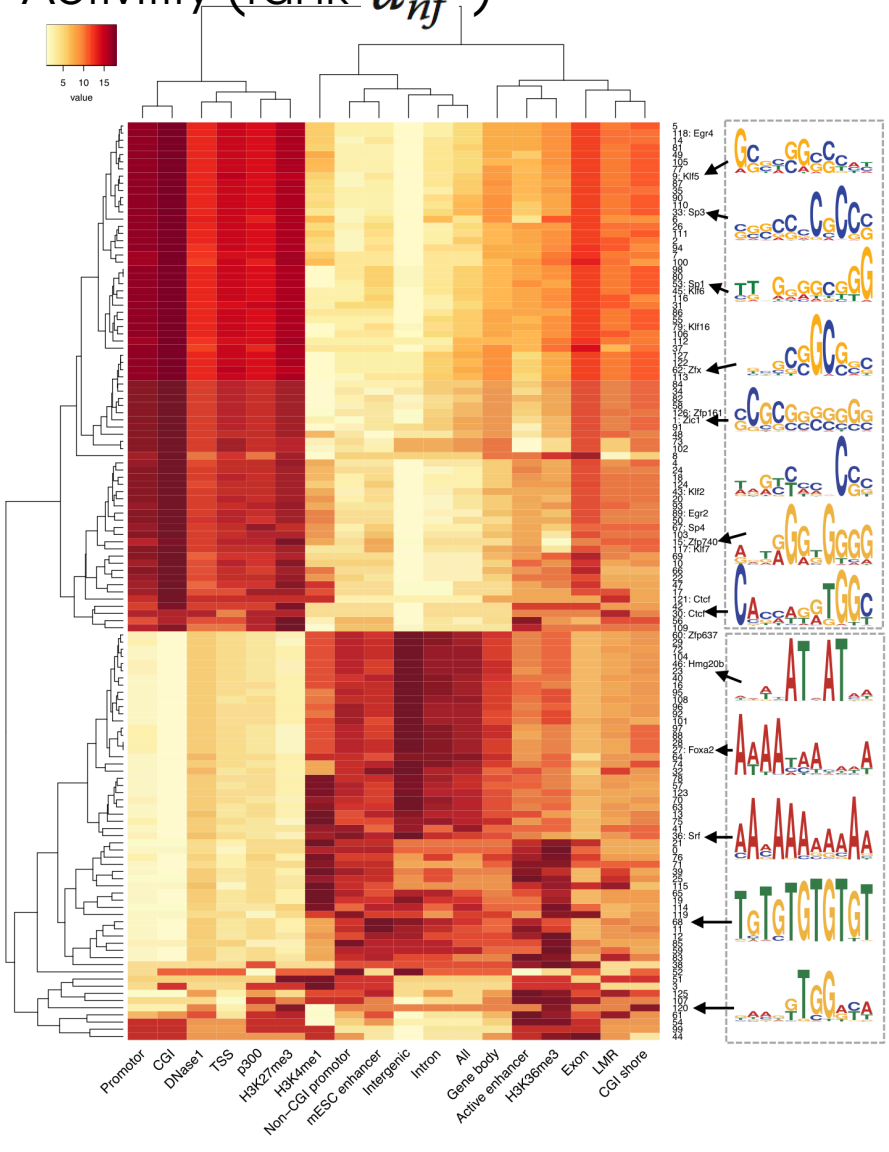
- Allows for the quantification of
  - Association with DNA methylation state:

$$r_{ft} = \mathrm{cor}_n(\bar{a}_{nf}, \hat{y}_{nt})$$

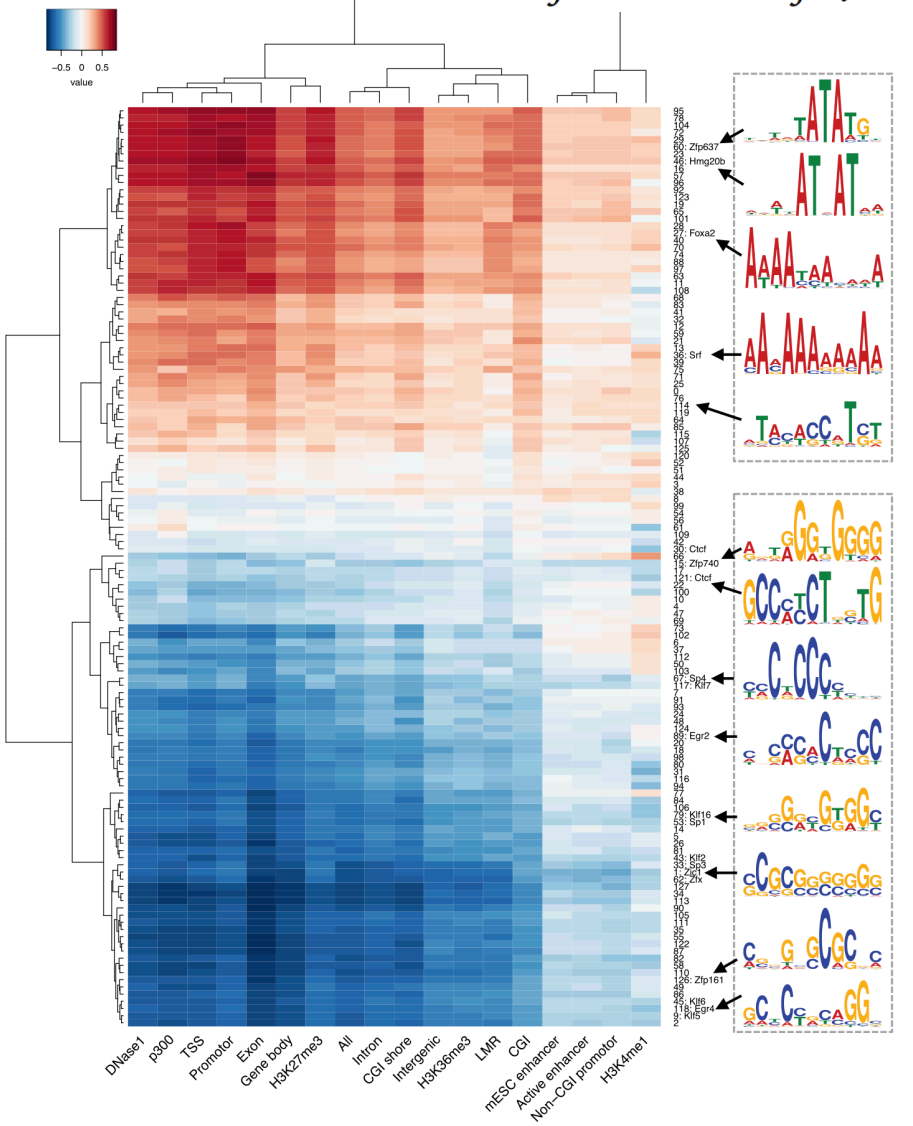
  - Sequence motifs

Filter Analysis

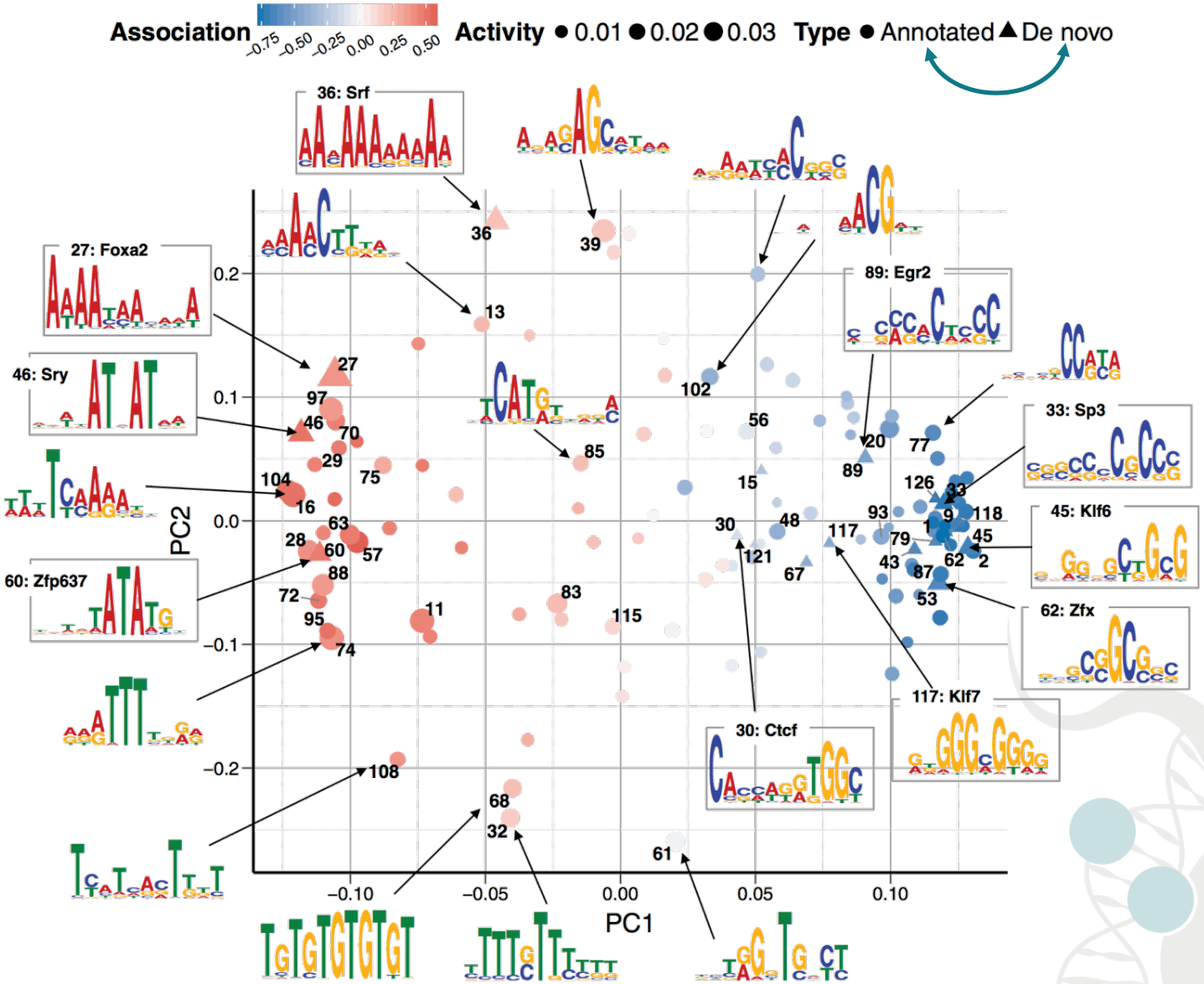
Activitity (rank $\bar{a}_{nf}$)



Effect on methylation $r_{ft} = \mathrm{cor}_n(\bar{a}_{nf}, \hat{y}_{nt})$



Filter Analysis



Effect of Sequence Mutations

- Estimate the effect of SNPs on the predicted methylation:

mean predicted methylation across cells

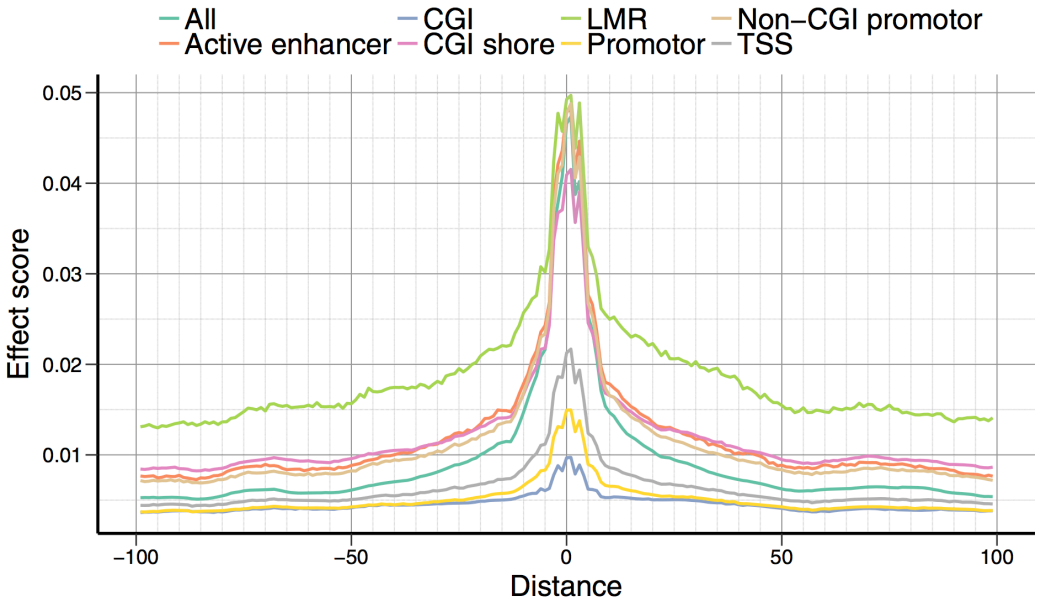
$$e^s_{nid} = \frac{\Delta\, \hat{y}_n(s_n)}{\Delta s_{nid}} * (1 - s_{nid})$$

gradient w.r.t. sequence

set effect of WT to 0

$$e^s_{ni} = \max_d \left| e^s_{nid} \right|$$

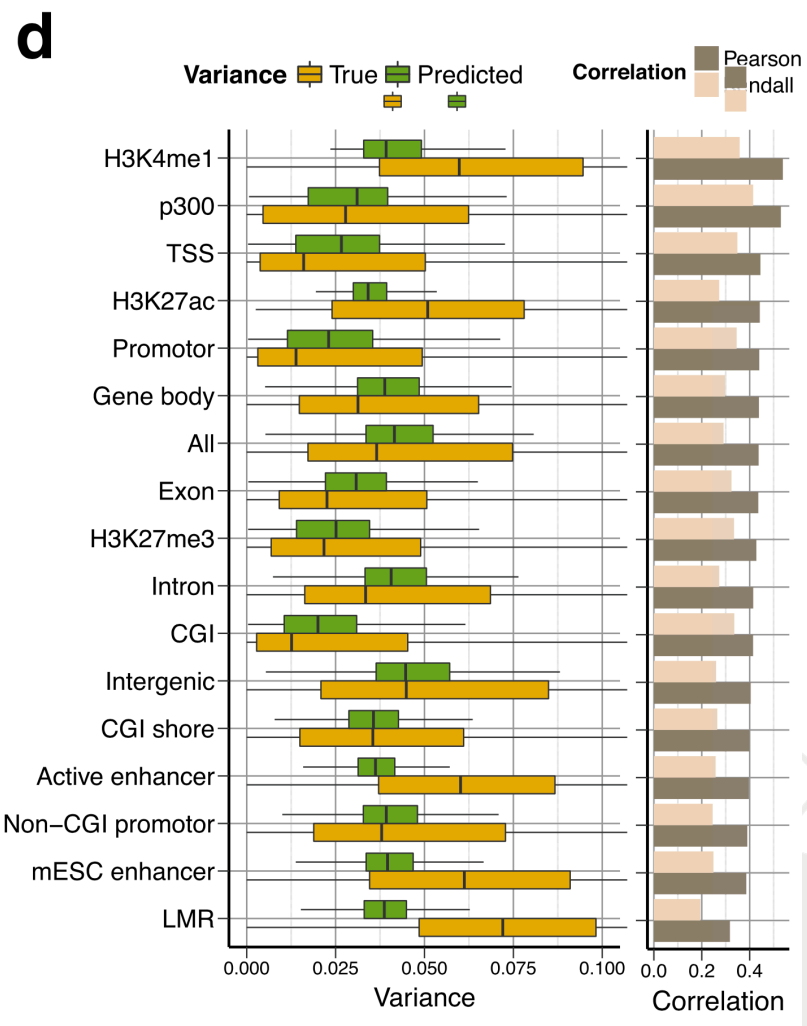
$$e^s_i = \text{mean}_n \left(e^s_{ni} \right)$$



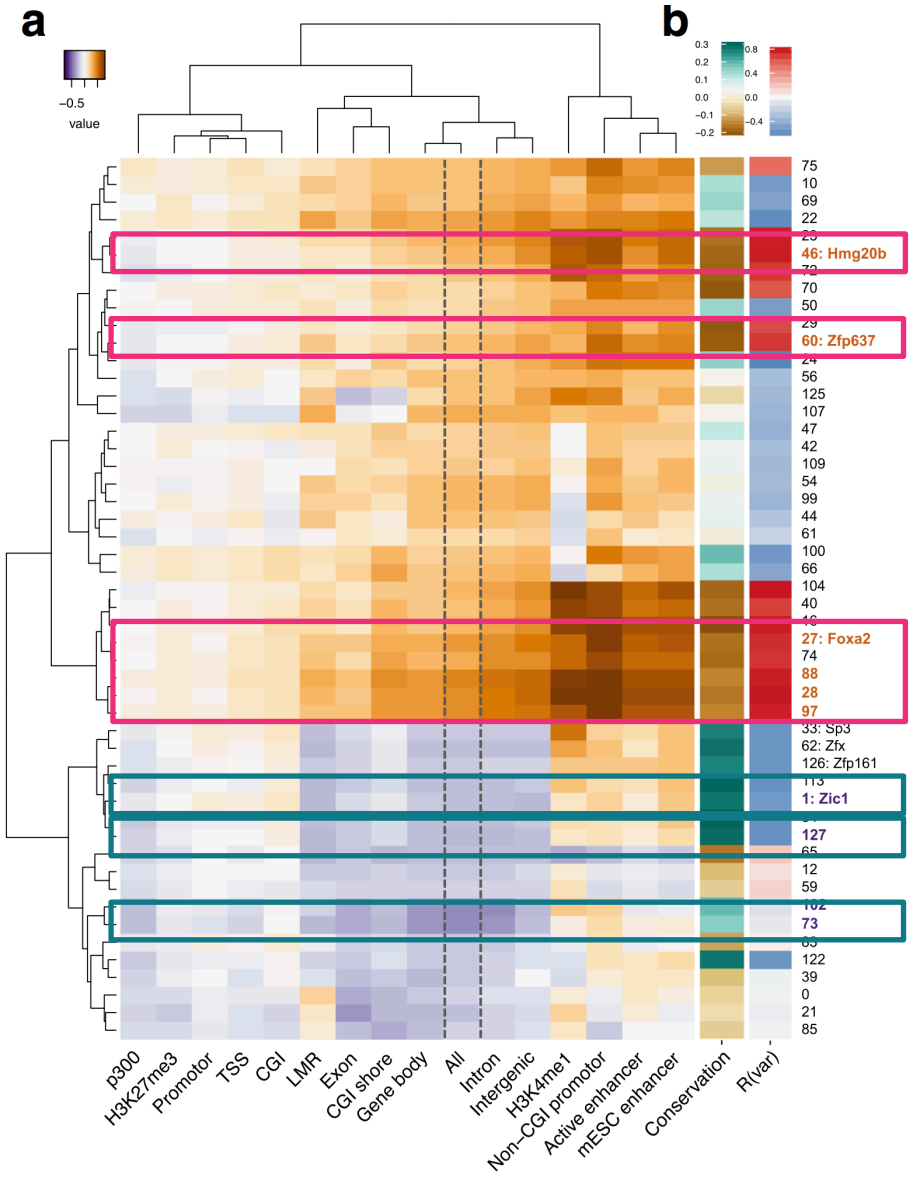
- Known mQTL mutations have significantly larger effect than random

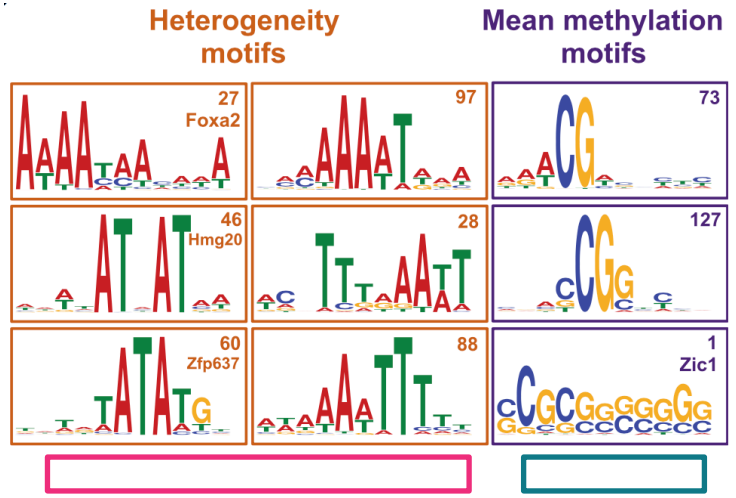
Methylation Variability

- Train another neural net
  - same architecture as the DNA module
  - Output layer:
    - Predict methylation variance and mean methylation across cells, averaged across windows
    - Sigmoid activation
- Compute association (correlation) of motif activity with mean methylation and variance



Methylation Variability





- Variance-associated motifs
  - AT-rich
  - Enhancers
  - Associate weakly with changes in gene expression

Summary

- DeepCpG employs deep neural networks to accurately predict missing methylation states based on DNA sequence and observed methylation
- Biologically relevant DNA sequence motifs are learned by the model
- Models can also be used to learn methylation variability

Open Issues

- Can the models be interpreted beyond motifs?
- Sequence-dependent DNA methylation?
- Other epigenetic features?
- Scaling to thousands/millions of single cells?