LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Journal Club:

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

by *C.Angermuelle, H.Lee, W.Reik and O. Stegle in Genome Biology, 2017*

Petr Nazarov

**2017-09-12**

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Paper

Angermueller *et al. Genome Biology* (2017) 18:67
DOI 10.1186/s13059-017-1189-z

Genome Biology

METHOD Open Access

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

Christof Angermueller[1*], Heather J. Lee[2,3], Wolf Reik[2,3] and Oliver Stegle[1*]

**Abstract**

Recent technological advances have enabled DNA methylation to be assayed at single-cell resolution. However, current protocols are limited by incomplete CpG coverage and hence methods to predict missing methylation states are critical to enable genome-wide analyses. We report DeepCpG, a computational approach based on deep neural networks to predict methylation states in single cells. We evaluate DeepCpG on single-cell methylation data from five cell types generated using alternative sequencing protocols. DeepCpG yields substantially more accurate predictions than previous methods. Additionally, we show that the model parameters can be interpreted, thereby providing insights into how sequence composition affects methylation variability.

**Keywords:** Deep learning, Artificial neural network, Machine learning, Single-cell genomics, DNA methylation, Epigenetics

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Outline

## Main Figure: the Idea

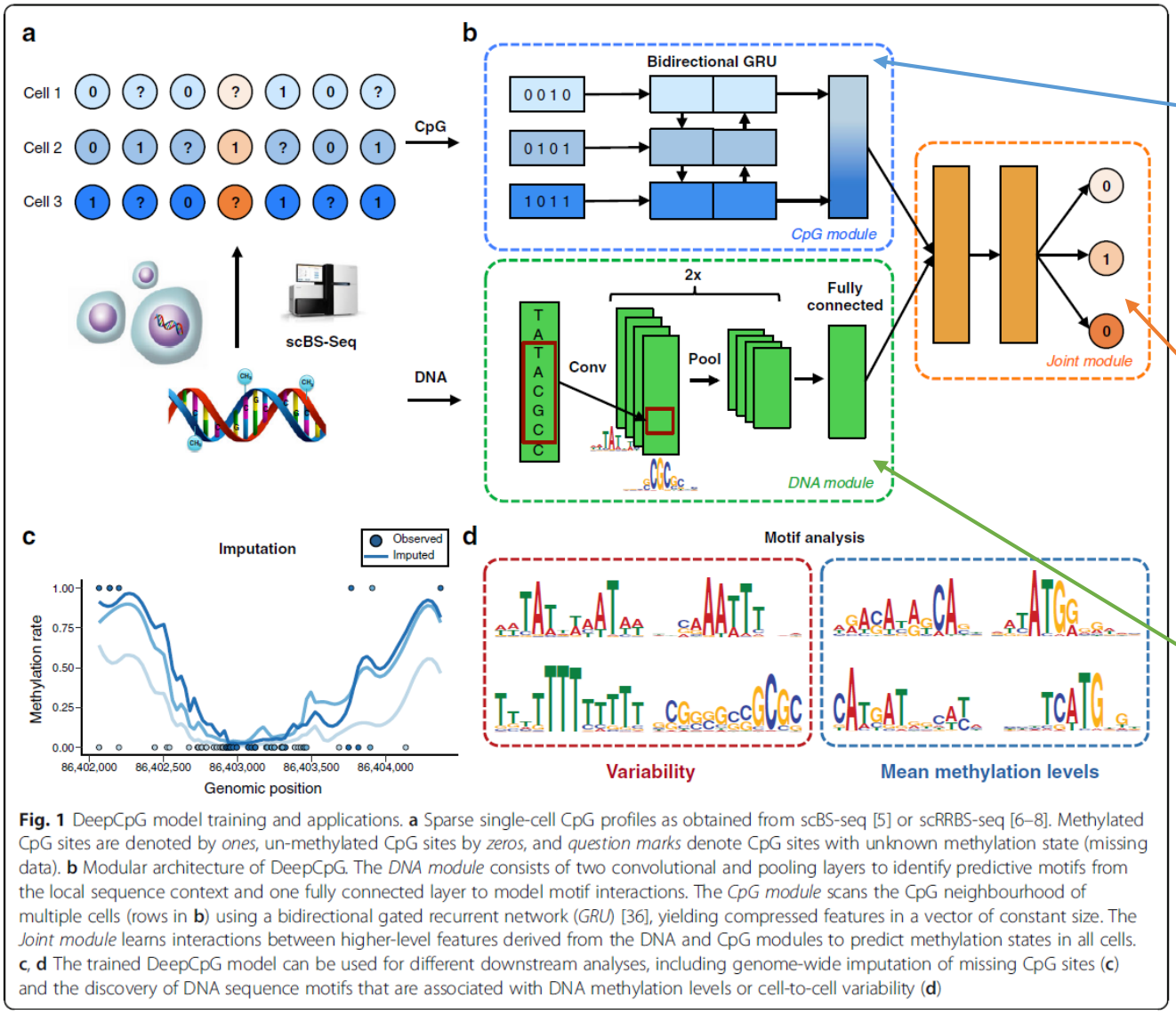



**Fig. 1** DeepCpG model training and applications. **a** Sparse single-cell CpG profiles as obtained from scBS-seq [5] or scRRBS-seq [6–8]. Methylated CpG sites are denoted by *ones*, un-methylated CpG sites by *zeros*, and *question marks* denote CpG sites with unknown methylation state (missing data). **b** Modular architecture of DeepCpG. The *DNA module* consists of two convolutional and pooling layers to identify predictive motifs from the local sequence context and one fully connected layer to model motif interactions. The *CpG module* scans the CpG neighbourhood of multiple cells (rows in **b**) using a bidirectional gated recurrent network (*GRU*) [36], yielding compressed features in a vector of constant size. The *Joint module* learns interactions between higher-level features derived from the DNA and CpG modules to predict methylation states in all cells. **c**, **d** The trained DeepCpG model can be used for different downstream analyses, including genome-wide imputation of missing CpG sites (**c**) and the discovery of DNA sequence motifs that are associated with DNA methylation levels or cell-to-cell variability (**d**)

Gated recursive neural network (GRN)

Fully-connected feed-forward network (FFN) or multi-layer perceptron (MLP)

Convolution neural network (CNN)

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Outline

There is no fun discussing the paper without knowing the instrument used for data processing and prediction.

Therefore:

- **Basics of artificial neural networks (ANN)**
  - *artificial neuron*
  - *feed forward network – FFN (a.k.a multi-layer perceptron MLP)*

- **Deep neural networks (DNN) and its application in the paper:**
  - *convolutional networks* (CNN)
  - *recurrent networks* (GRN)

- **Results & Discussion**

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

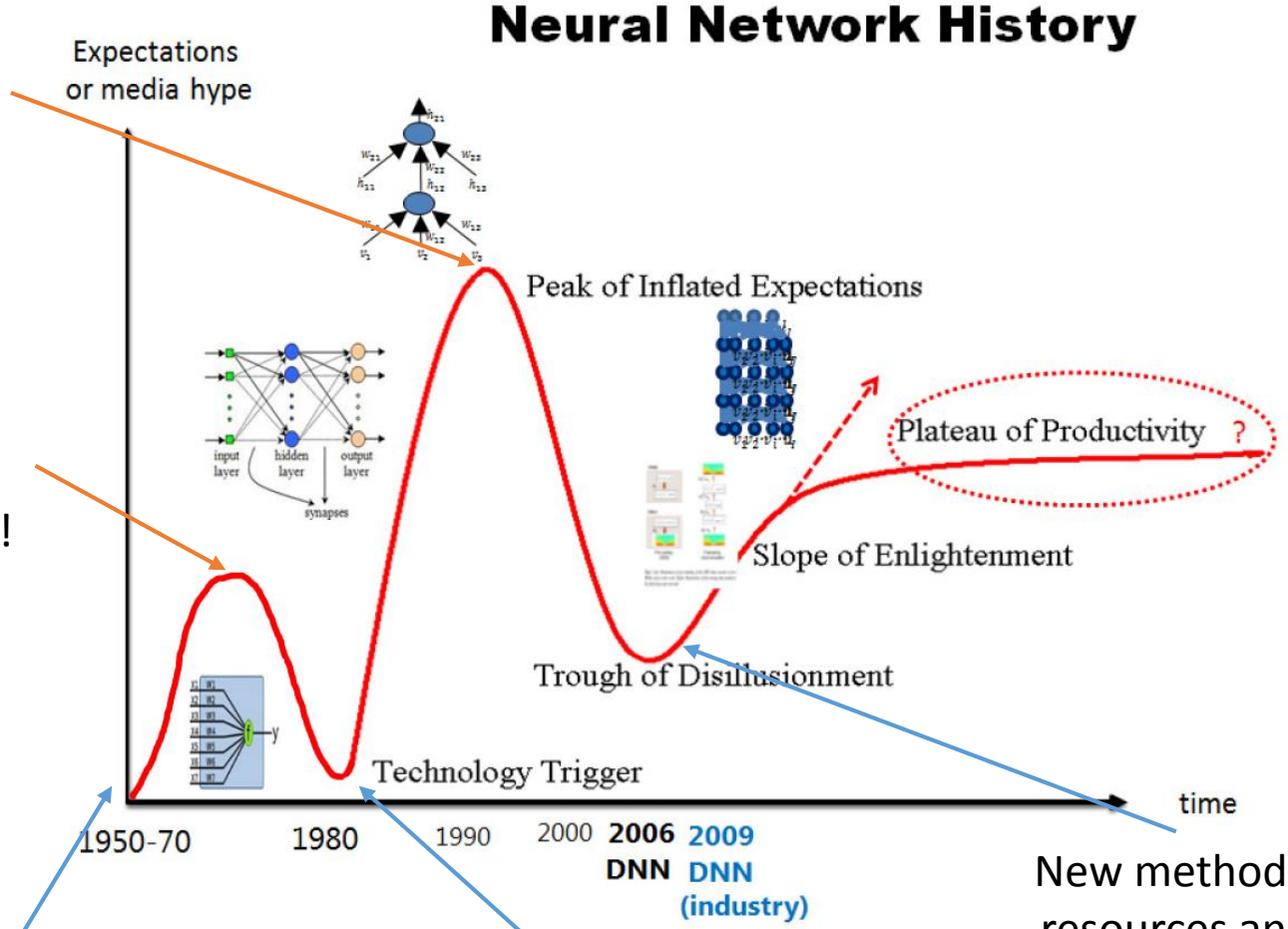
# Artificial Neural Networks

## History: Double "Gartner's Hype Cycle" ☺

1999: lack of computation power, lack of training data, gradient decay problem for deep nets.

**1970, Minsky, Papert**: Single-layer networks are limited!

1957 Frank **Rosenblatt** : Perceptron – a single neuron ☺ *in ferro*




1980s: The method to train multilayer networks (error backpropagation)

New methods, resources and BIG industrial players

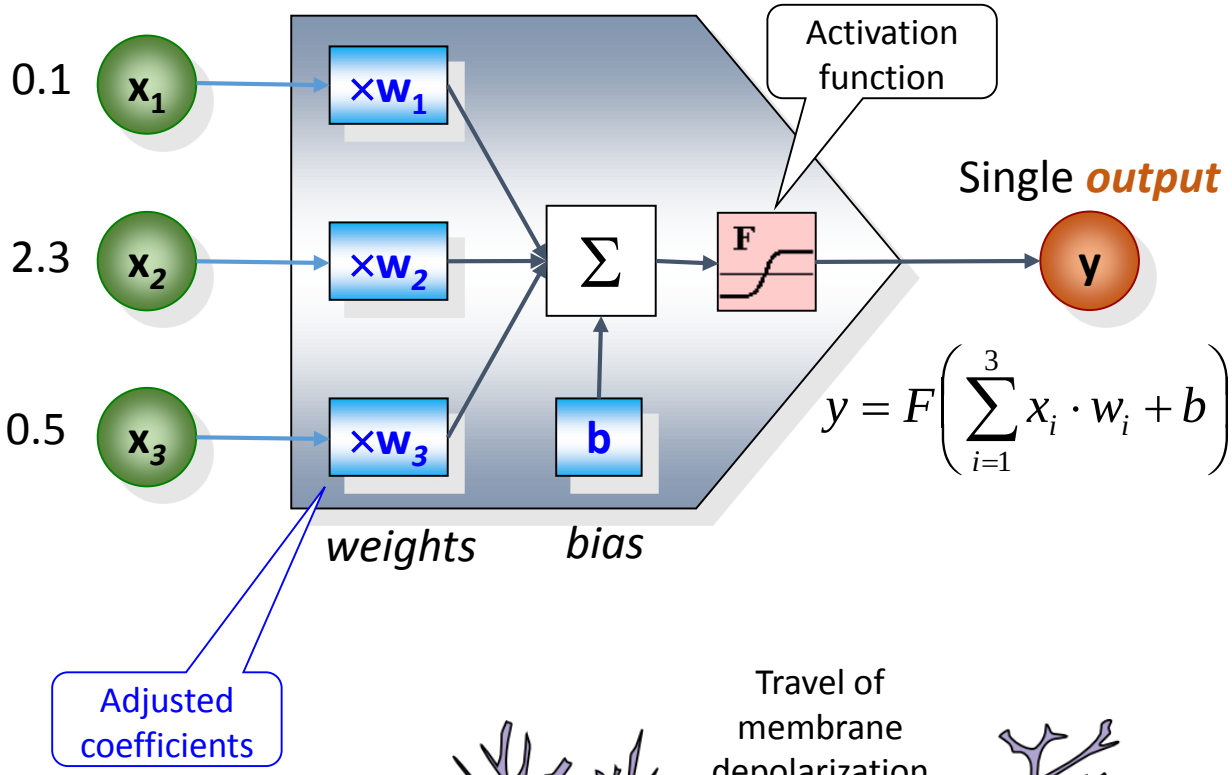
Modified from P. O. Glauner https://arxiv.org/abs/1504.06825

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

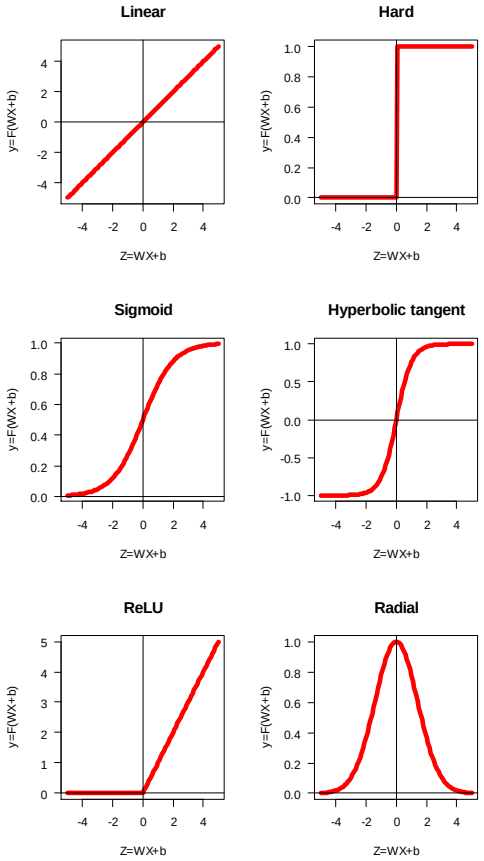
# Artificial Neural Networks

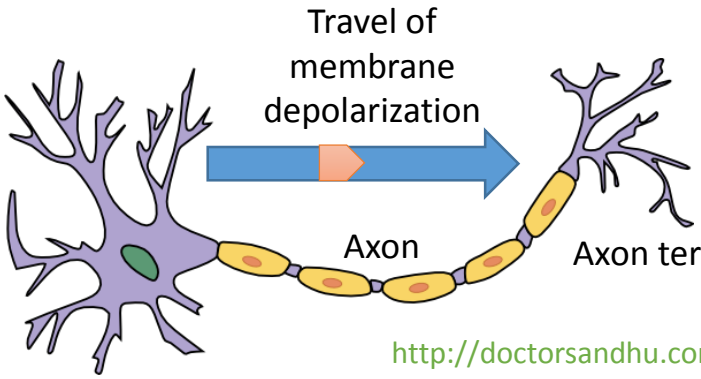
## Artificial Neuron – a Simple Processing Unit

Multiple *inputs*




$$y = F\left(\sum_{i=1}^{3} x_i \cdot w_i + b\right)$$




http://doctorsandhu.com/Neuron/Neuron.shtml

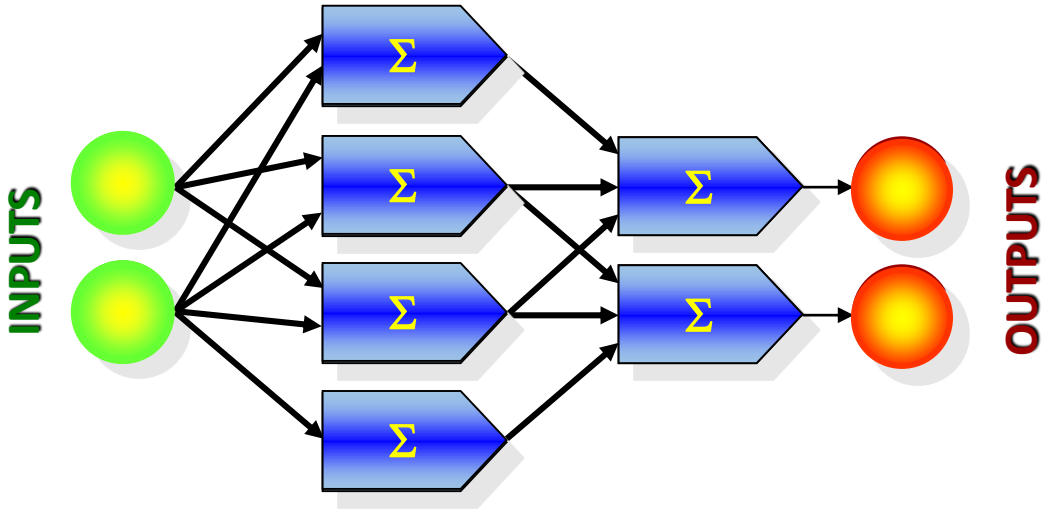
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Artificial Neural Networks

## Feed Forward Network (FFN), a.k.a. Multi-layer Perceptron (MLP)

**Forward propagation of information**

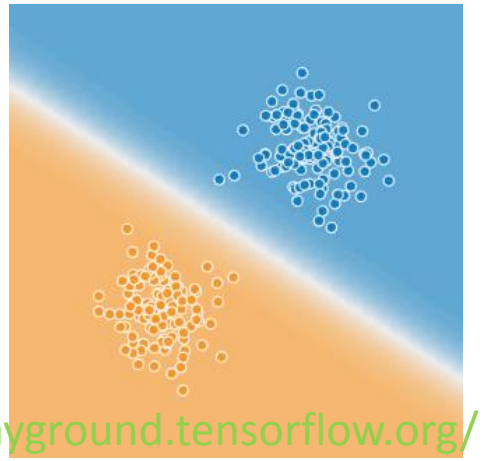
**Normalized data:** raw, features, variables etc.



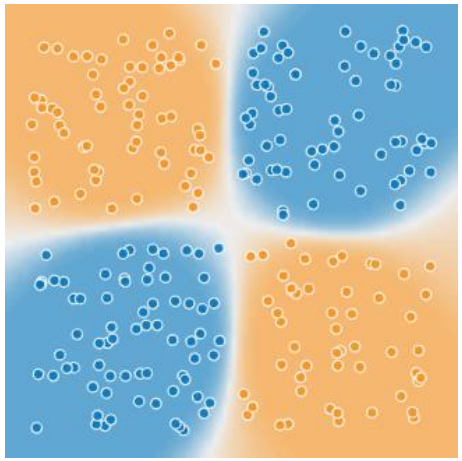

In classification the output is considered as probability of a class (with *softmax*)

$$p(y_i|X) = \frac{y_i}{\sum y_j}$$

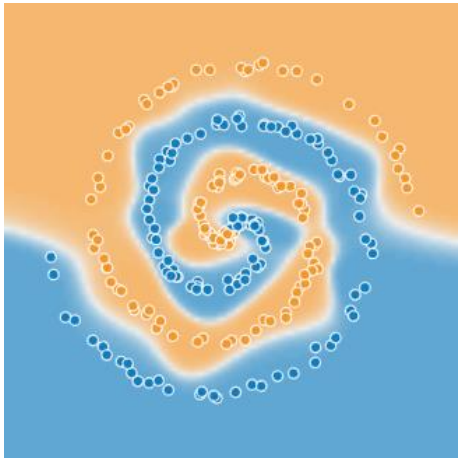
1 layer



2 layers



4 layers



http://playground.tensorflow.org/

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Deep Networks

## Convolutional Networks

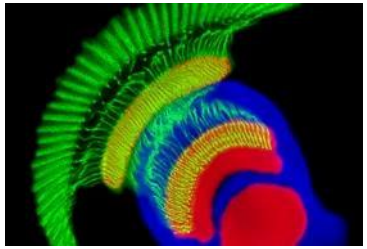
**How to distinguish cats from cancers with FFN ?**

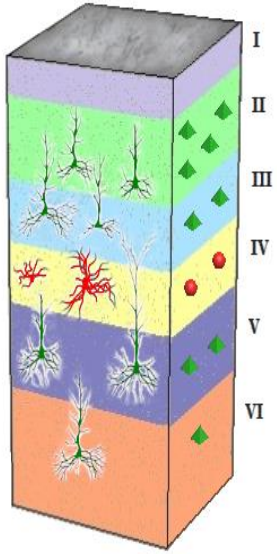

256 x 256 pixels, 3 color channels

**Inputs:**
256 x 256 x 3
≈ 200 k

**Layer 1:**
200 k x 100 ...
= **20 M** ← **Not feasible !**

**Why not to use what was already invented by evolution?**






**Visual cortex:**

- multi layer organization

- links only between spatially close neurons & receptors

- each layer reduces the size => extract features

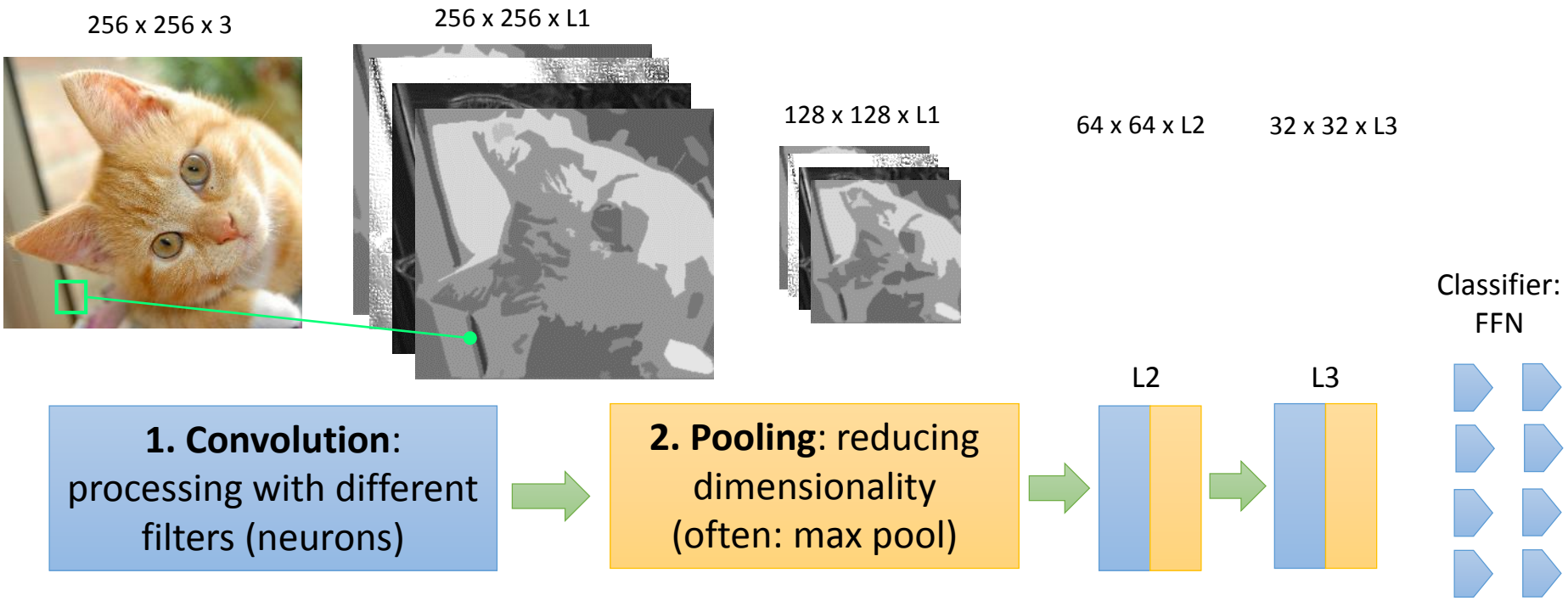
https://www.pinterest.com/pin/203647214370581827/

https://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/adaptation-and-neuronal-network-in-visual-cortex

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Deep Networks

## Convolutional Networks for Images



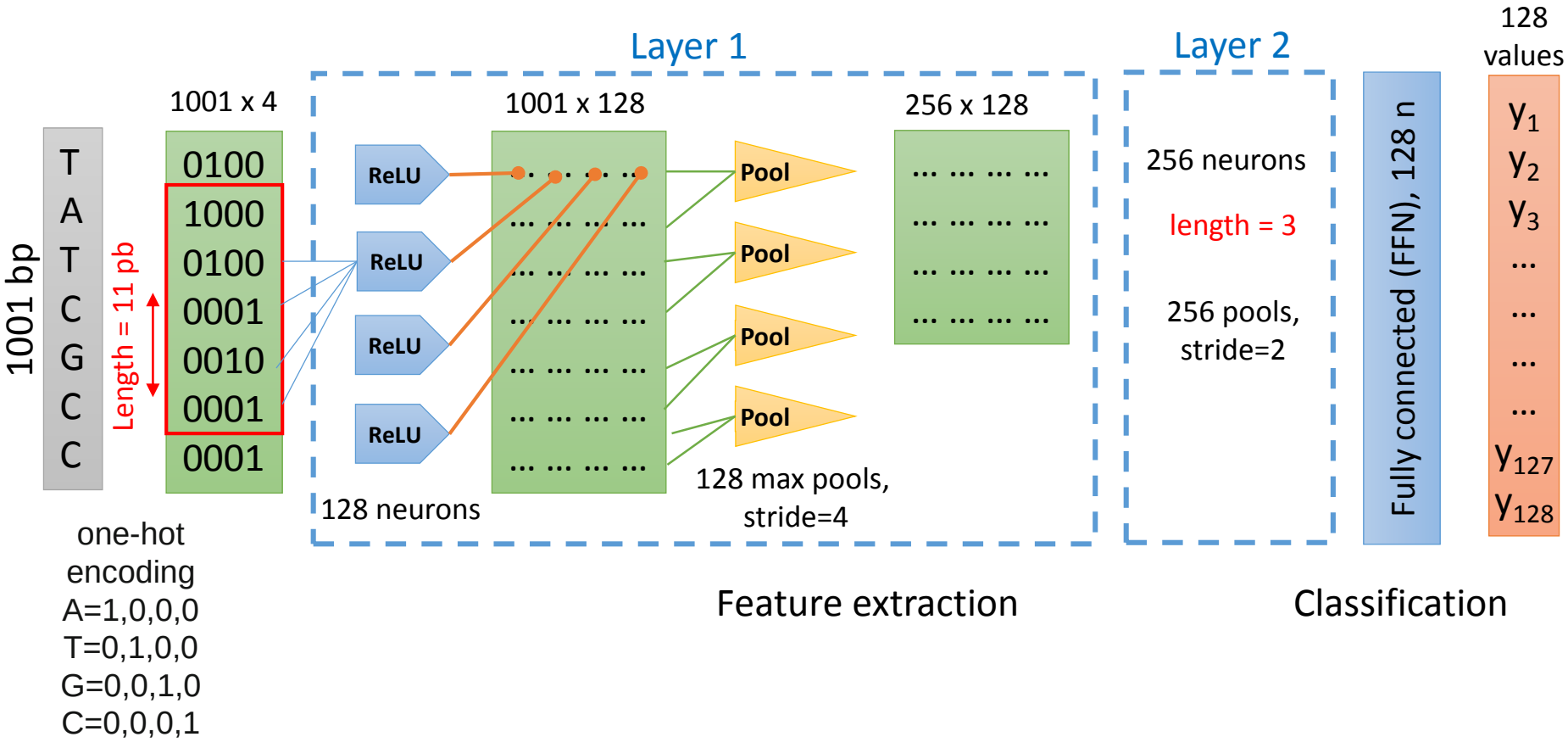
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Deep Networks

## Convolutional Networks for Sequences

Let's consider a simpler situation: genomic sequence instead of cat photos ☺

*Numbers are taken from the paper*

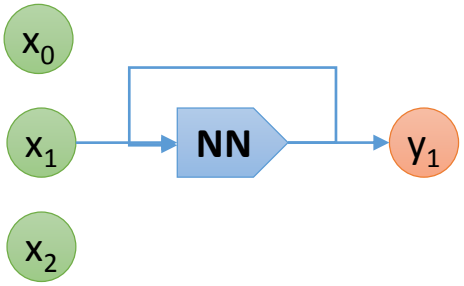


LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Deep Networks

## Recursive Networks: a Simplified Concept

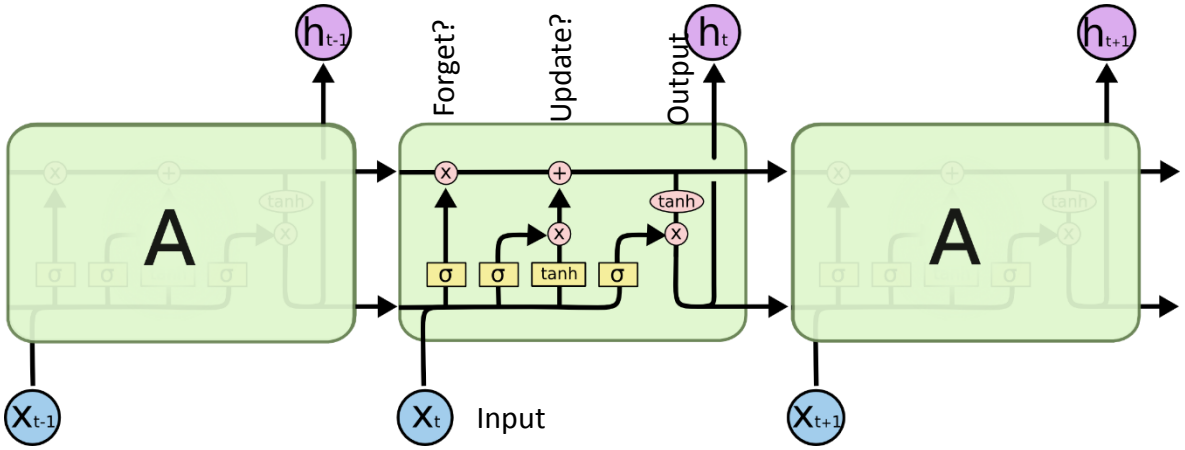
How about predicting or processing series of data? Speech, music, sequences, etc.




To combine current inputs and memory management:

**Long Short Term Memory - LSTM**

It is as "simple" as it sounds… ☺




**Features:**

- each neuron keeps in memory its state;
- can clear memory;
- can update memory;
- output is based on inputs and memory;
- robust in case of missing values

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

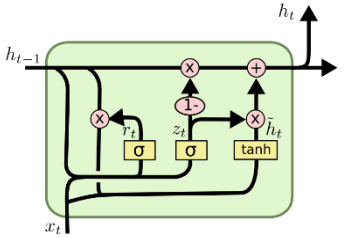
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Deep Networks

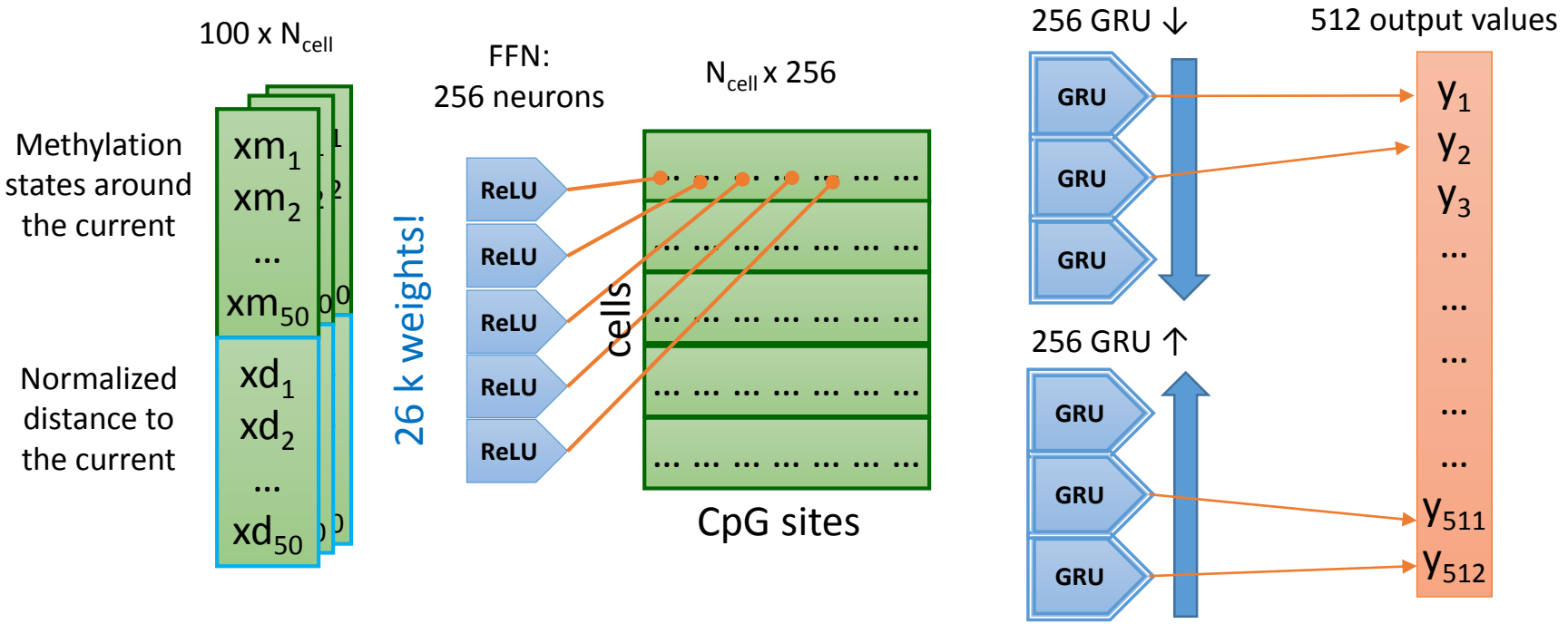
## Convolutional Networks for CpG Prediction

In paper they used FFN followed by simplified LSTM: Gated Recurrent Unit (GRU)

*This part was not very clear from the paper and needed investigation*




$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

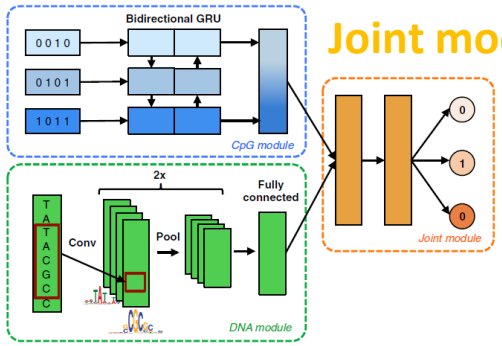


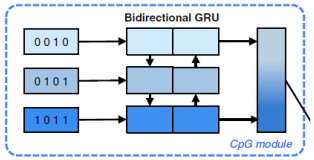

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Deep Networks

## Joint Module




Training time: 12h

CpG

$x_1$
$x_2$
$x_3$
...
...
...
...
$x_{511}$
$x_{512}$

328 k weights!

2 x 512 neurons

ReLU
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU

262 k weights!

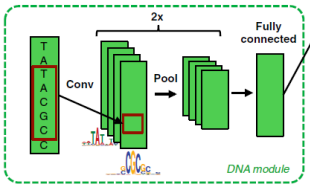
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU

5k-1M weights!

$N_{cell}$ neurons

$N_{cell}$ outputs

sigm → $y_1$
sigm → $y_1$
sigm
sigm → $y_N$




Training time: 24h

$x_1$
$x_2$
...
$x_{127}$
$x_{128}$

DNA

Training time: 4h

Estimated number of fitted parameters: 500k – 1.5 M !

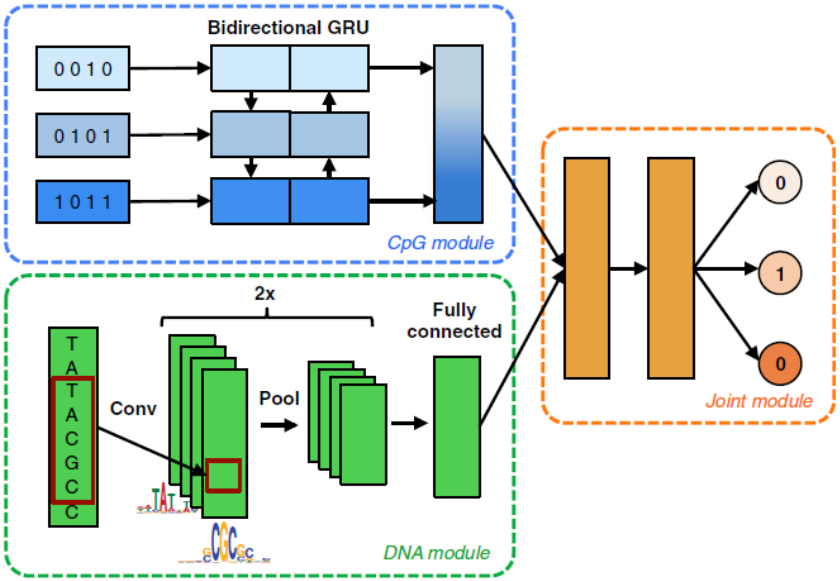
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## Architecture Overview

### CpG module

Learns correlations within and between cells.




**Training**: chr. 1, 3, 5, 7, 9, 11
**Validation** (aka "control"): chr. 13-19

**Testing**: chr. 2, 4, 6, 8, 10, 12

### Joint module

Combines observed neighboring methylations and sequence-based predictions into final prediction.

### DNA module

Accounts for sequence motifs that can predict the methylation state

**Outcomes:**

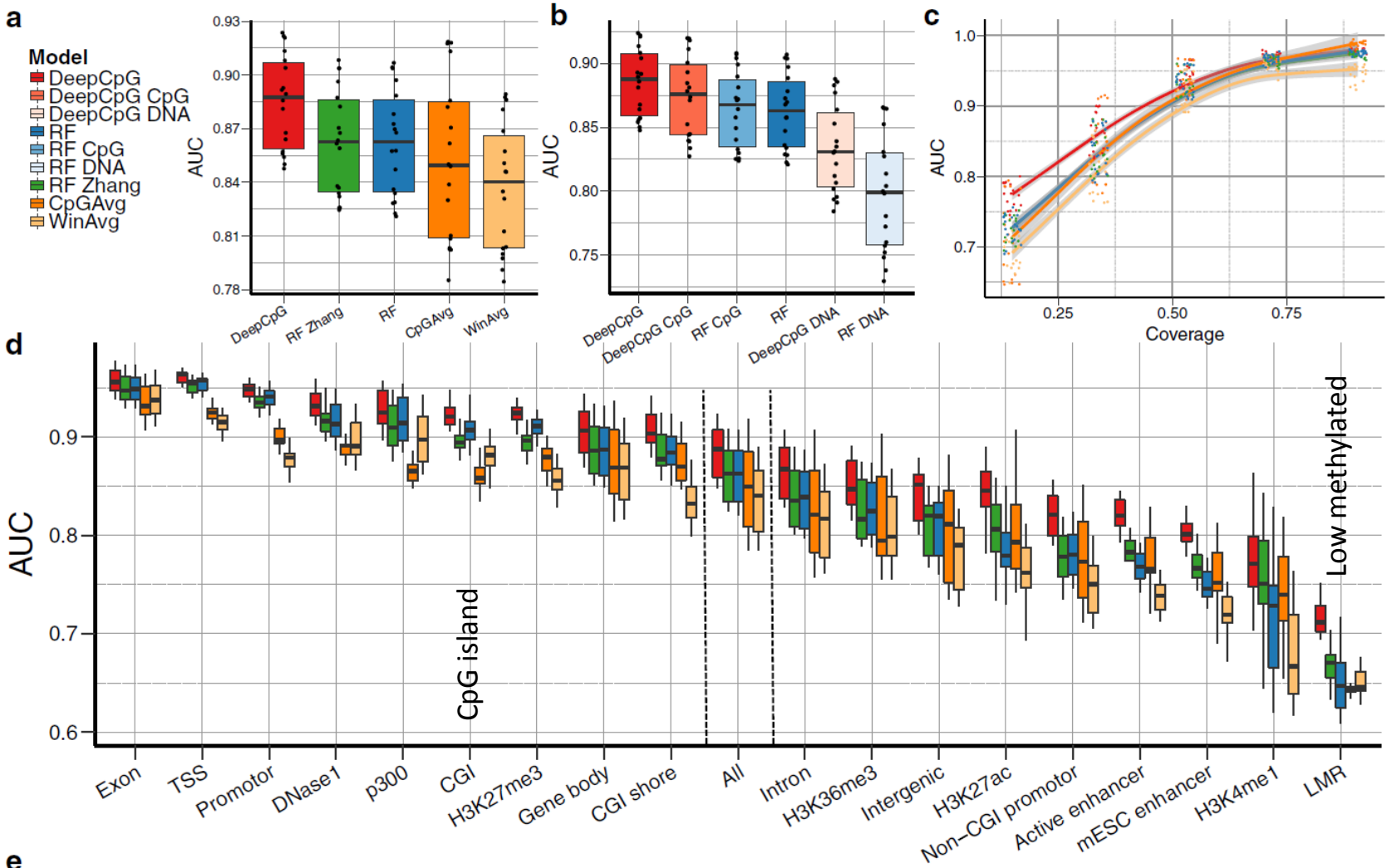
- imputation of missing methylation values
- discovering DNA motifs associated with methylation states

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## Imputation of Methylation States

18 serum-cultured mouse embryonic stem cells ("Serum"): CpG coverage = 17%, scBS-seq



LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## Imputation of Methylation States

Other cells:

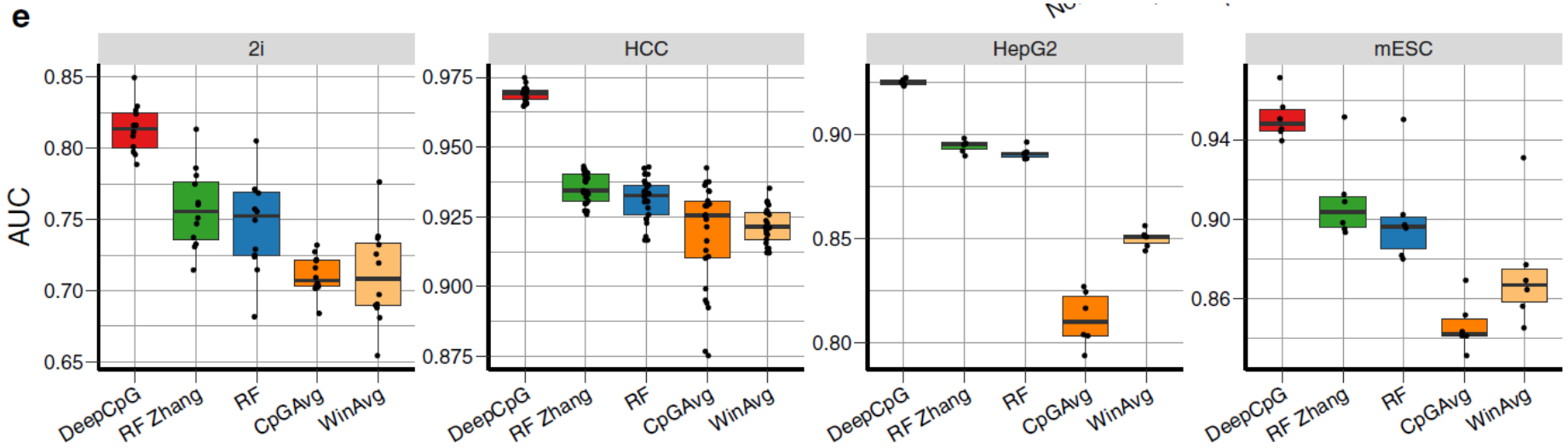



**Fig. 2** DeepCpG accurately predicts single-cell CpG methylation states. **a** Genome-wide prediction performance for imputing CpG sites in 18 serum-grown mouse embryonic stem cells (*mESCs*) profiled using scBS-seq [5]. Performance is measured by the area under the receiver-operating characteristic curve (*AUC*), using holdout validation. Considered were DeepCpG and random forest classifiers trained either using DNA sequence and CpG features (*RF*) or using additional annotations from corresponding cell types (*RF Zhang* [12]). Additionally, two baseline methods were considered, which estimate methylation states by averaging observed methylation states, either across consecutive 3-kbp regions within individual cells (*WinAvg* [5]) or across cells at a single CpG site (*CpGAvg*). **b** Performance breakdown of DeepCpG and RF, comparing the full models to models trained using either only methylation features (*DeepCpG CpG*, *RF CpG*) or only DNA features (*DeepCpG DNA*, *RF DNA*). **c** AUC of the methods as in (**a**) stratified by genomic contexts with increasing CpG coverage across cells. Trend lines were fit using local polynomial regression (*LOESS* [72]); *shaded areas* denote 95% confidence intervals. **d** AUC for alternative sequence contexts with *All* corresponding to genome-wide performance as in (**a**). **e** Genome-wide prediction performance on 12 2i-grown mESCs profiled using scBS-seq [5], as well as three cell types profiled using scRRBS-seq [8], including 25 human hepatocellular carcinoma cells (*HCC*), six HepG2 cells, and six additional mESCs. *CGI* CpG island, *LMR* low-methylated region, *TSS* transcription start site

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

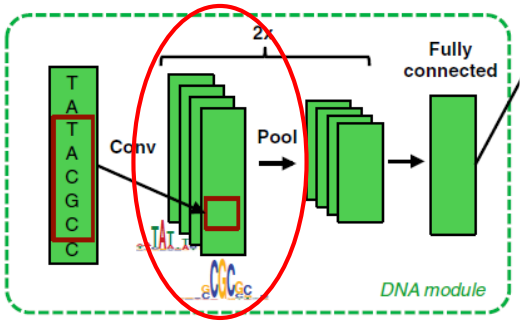
# Results

## Effects of DNA Motifs

**Q1. Discover methyl-associated motifs**

**Q2. Investigate the effect of SN mutation**

*1a. Motifs can be found from 1st layer of DNA module*




**First convolution layer:**
128 neurons or "filters", with weights that prioritize some motifs.
=> 128 motifs

*2. Linking changes in sequence to prediction*

$s_n$ – sequence, $d$ – nucleotide,
$i$ -position of change

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

**Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps**


Karen Simonyan Andrea Vedaldi Andrew Zisserman
Visual Geometry Group, University of Oxford
{karen,vedaldi,az}@robots.ox.ac.uk


*1b. Motif activity*

**Averaged** output of each neuron/filter over all scanned position is *activity* $a_{nf}$

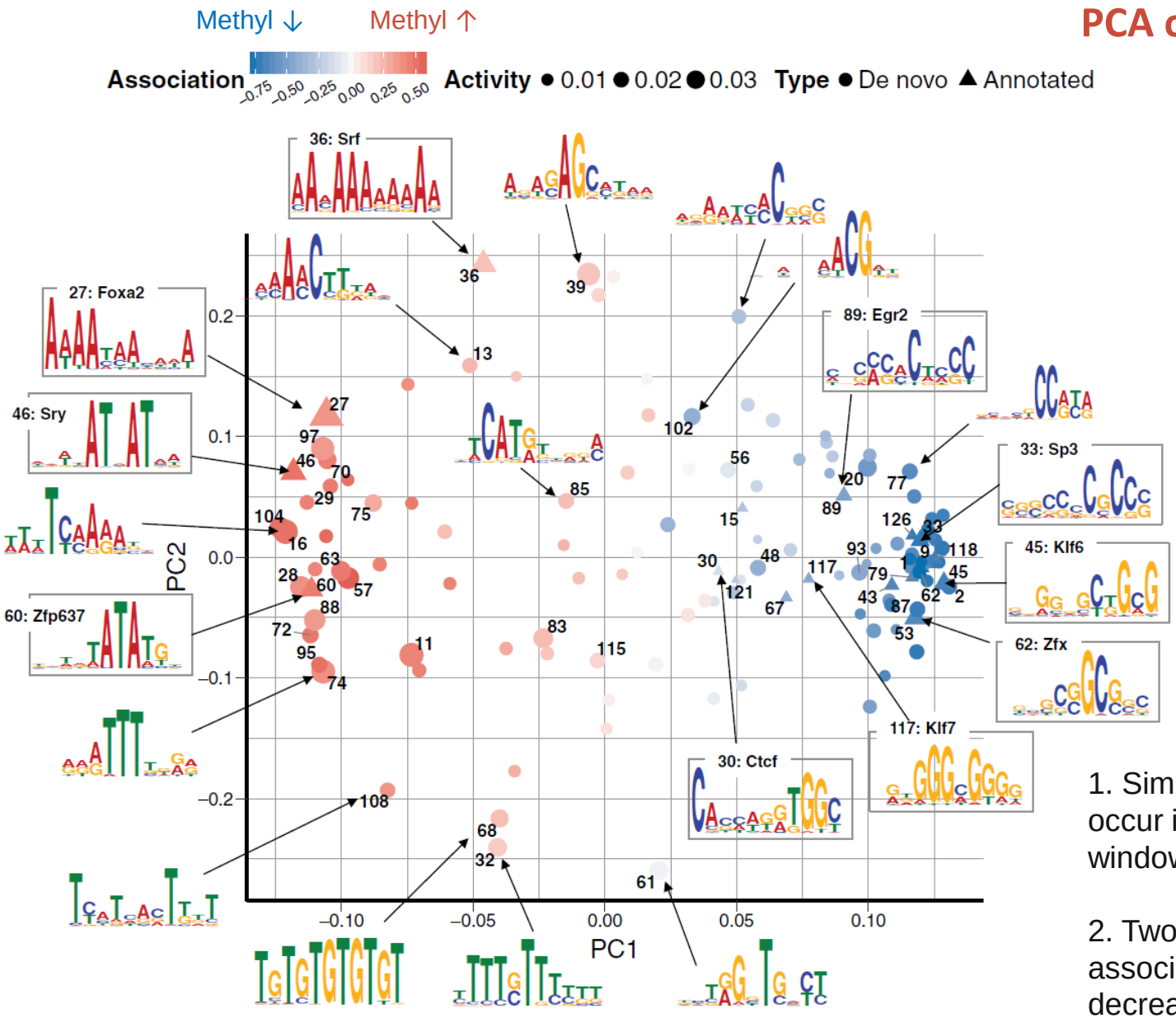
*1c. Motif influence on methylation*

Pearson correlation between *activity* and predicted methylation is a measure of *influence* $r_{ft} = cor(a_{nf}, y_{nt})$

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## PCA of Motifs Activity



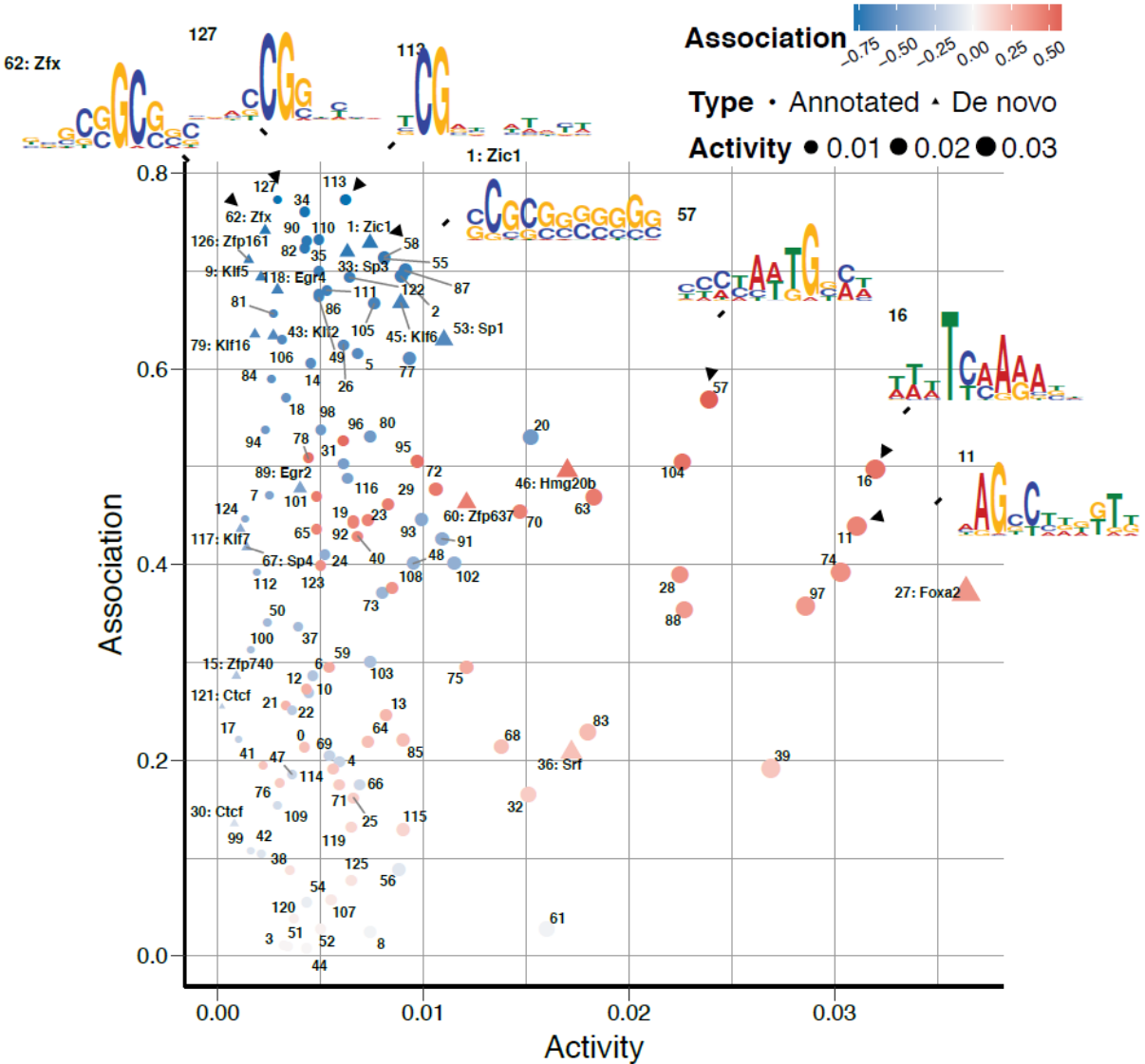

1. Similar motifs tend to co-occur in the same sequence windows;

2. Two major motif clusters are associated with increased or decreased methylation levels.

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## Activity and Association



LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## Effect of Mutations on Methylation

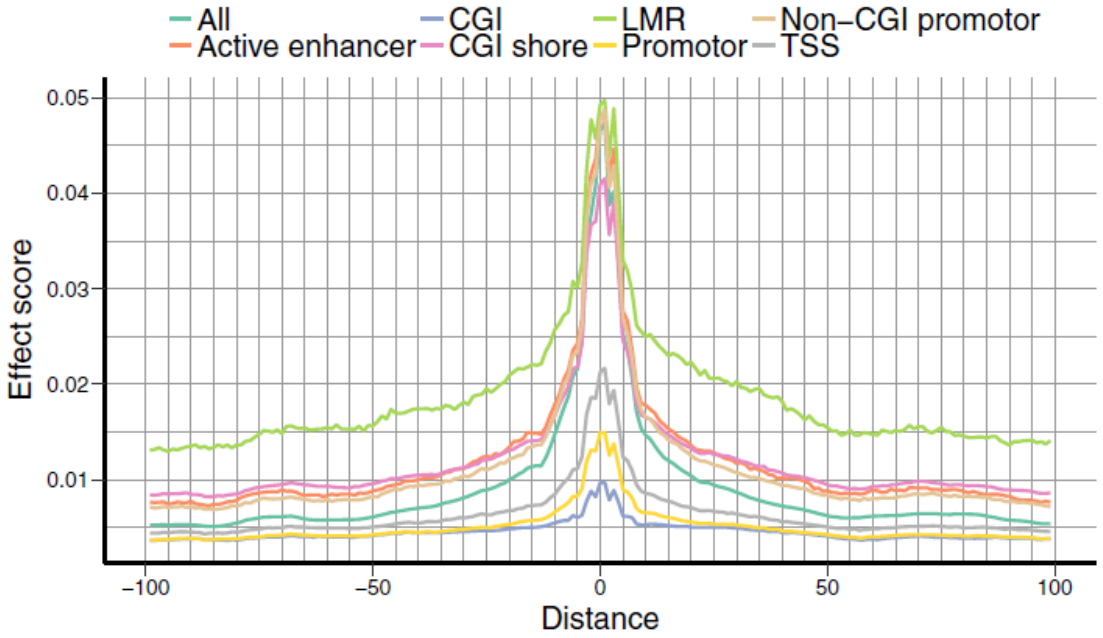



**Fig. 4** Effect of single-nucleotide mutations on DNA methylation. Average genome-wide effect of single-nucleotide mutations on DNA methylation estimated using DeepCpG, depending on the distance to the CpG site and genomic context. *CGI* CpG island, *LMR* low-methylated region, *TSS* transcription start site

1. Mutations in CG-dense regions tended to have smaller effects
2. Mutation in low-methylated region has the strongest effect. *But not discussed. Artefact?*

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Results

## Motifs and Variability between Cells

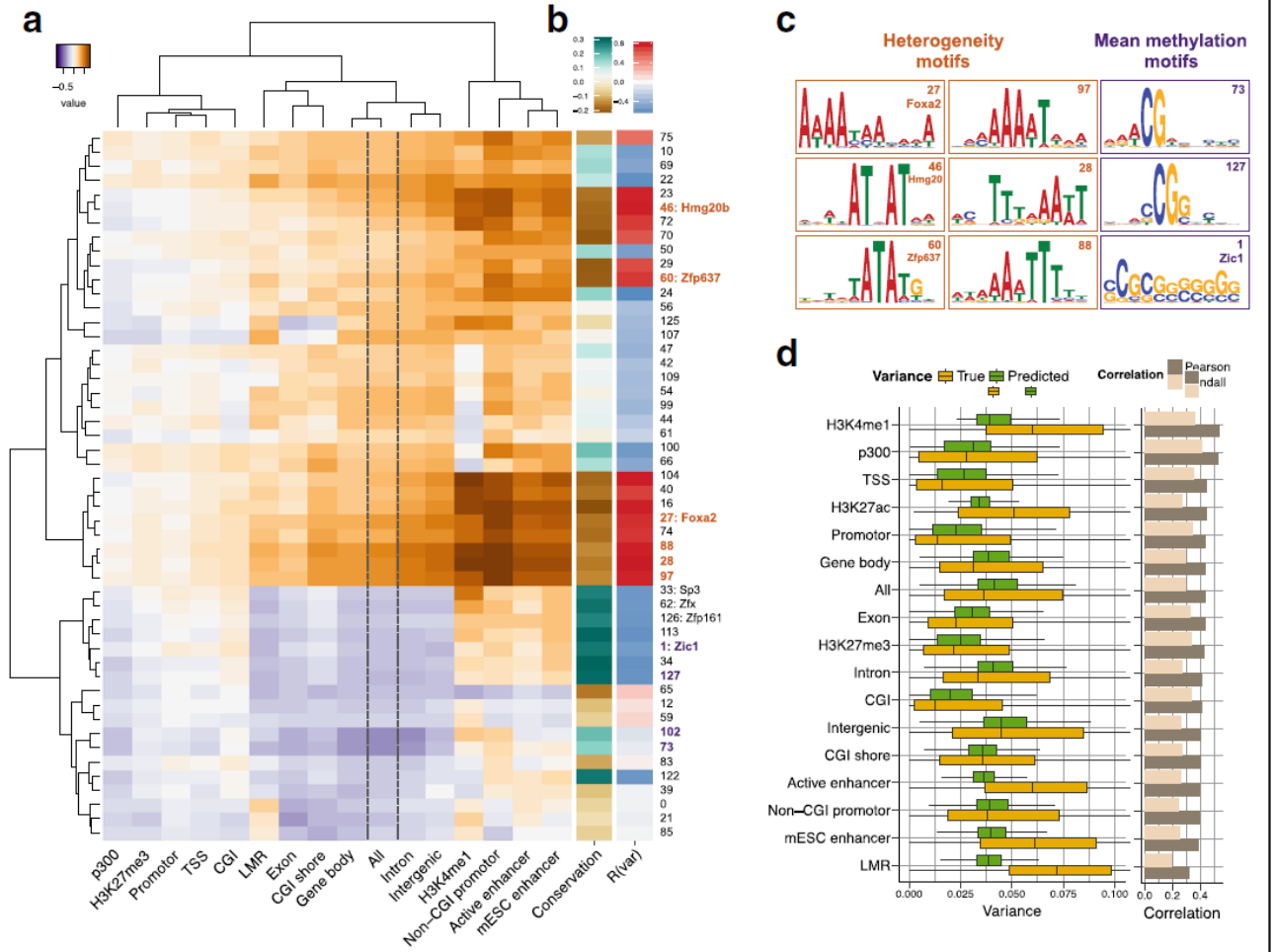
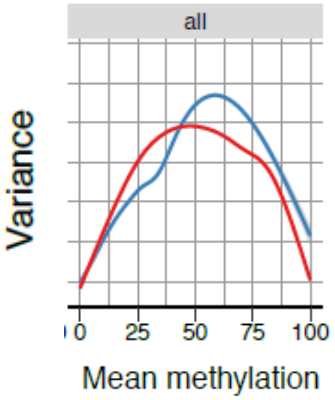



**Fig. 5** Prediction of methylation variability from local DNA sequence. **a** Difference of motif effect on cell-to-cell variability and methylation levels for different genomic contexts. Motifs associated with increased cell-to-cell variability are highlighted in *brown*; motifs that are primarily associated with changes in methylation level are shown in *purple*. **b** Genome-wide correlation coefficients between motif activity and DNA sequence conservation (*left*), as well as cell-to-cell variability (*right*). **c** Sequence logos for selected motifs identified in (**a**), which are highlighted with coloured text in (**b**). **d** Boxplots of the predicted and the observed cell-to-cell variability for different genomic contexts on held-out test chromosomes (*left*), alongside Pearson and Kendall correlation coefficients within contexts (*right*). *CGI* CpG island, *LMR* low-methylated region, *TSS* transcription start site

Here they used one more ANN to distinguish motifs that affect variability from those that affect methylation.

Let's skip…

Methylation is linked to variance




Hmm.. Isn't it pure statistics of Binomial distribution?

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Summary

## And Comments

1. DeepCpG accurately predicts missing values of methylation states

2. It detects sequence motifs that are associated with :
   - changes in methylation levels
   - cell-to-cell variability.

3. DeepCpG potentially can help studying methylation variability that is independent of DNA sequence effects (not in the paper, just stated)

4. Future: integrate multiple-omics data profiled in the same cells using parallel-profiling methods

*In general – I liked the paper. But it is quite methodologically complex. Some points need to be tried manually in order to be understood. Selection of some hyperparameters stay unclear for me… except the preference of authors to powers of 2 (64, 128,256,512) ☺*

*Only one discrepacy was found – Fig 1 (CpG module) does not fit to the M&M text.*

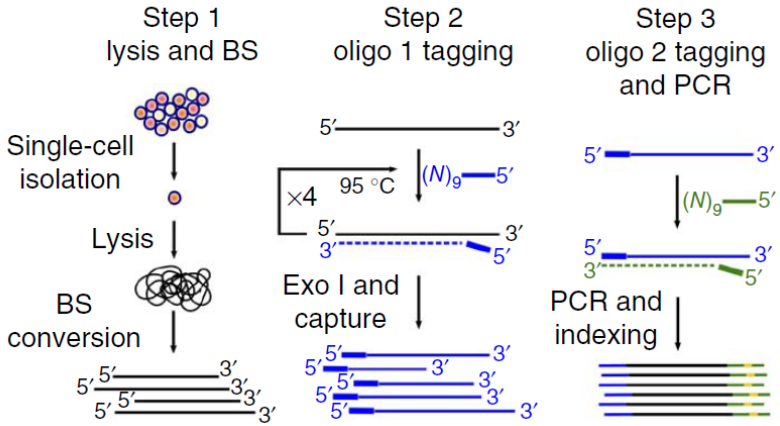
*To discuss: Can we use something similar for our SC RAN-seq data?*

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Appendix I

scBS-Seq

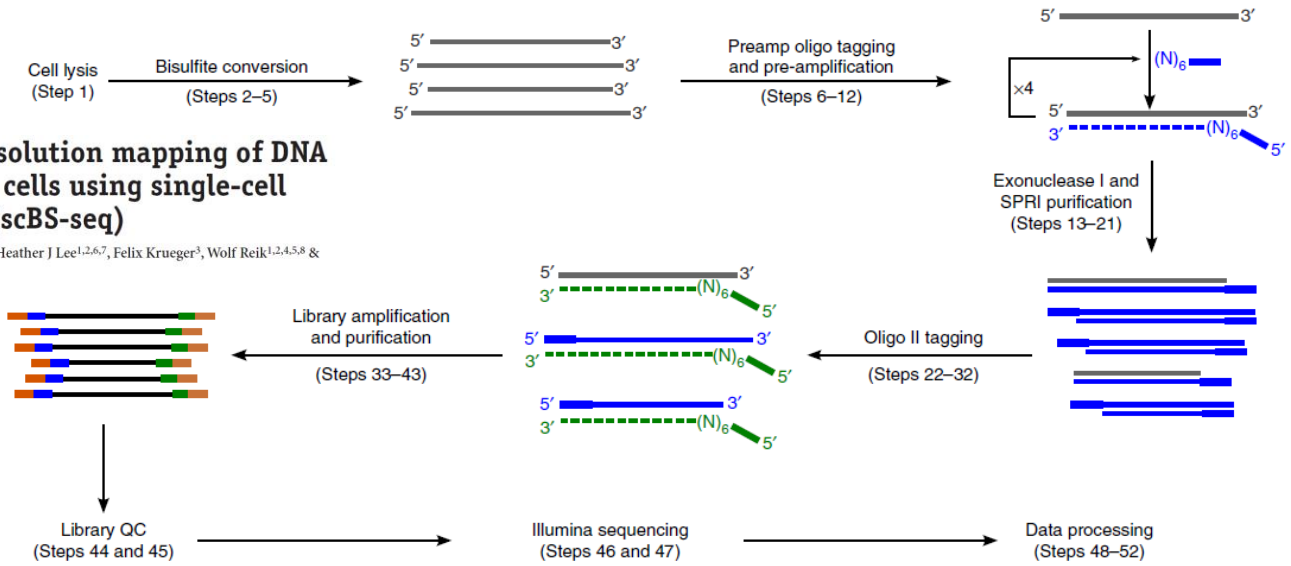
a




## Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity

Sébastien A Smallwood[1,6], Heather J Lee[1,2,6], Christof Angermueller[3], Felix Krueger[4], Heba Saadeh[1], Julian Peat[1], Simon R Andrews[4], Oliver Stegle[3], Wolf Reik[1,2,5,7] & Gavin Kelsey[1,5,7]

b




**Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq)**

Stephen J Clark[1,7], Sébastien A Smallwood[1,6,7], Heather J Lee[1,2,6,7], Felix Krueger[3], Wolf Reik[1,2,4,5,8] & Gavin Kelsey[1,4,8]

**Figure 1** | Overview of scBS-seq library preparation protocol. Single-cells are lysed and the DNA bisulfite is converted. Random priming and extension are used to preamplify and incorporate forward and, subsequently, reverse adaptor sequences. Finally, PCR is used to amplify and index the libraries before they are sequenced.

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

# Appendix II

PROTOCOL

scRRBS-Seq

## Profiling DNA methylome landscapes cells with single-cell reduced-repres bisulfite sequencing

Hongshan Guo[1,5], Ping Zhu[1,2,5], Fan Guo[1], Xianlong Li[1], Xinglong Wu[1,2], Xiaoy Fuchou Tang[1,3,4]

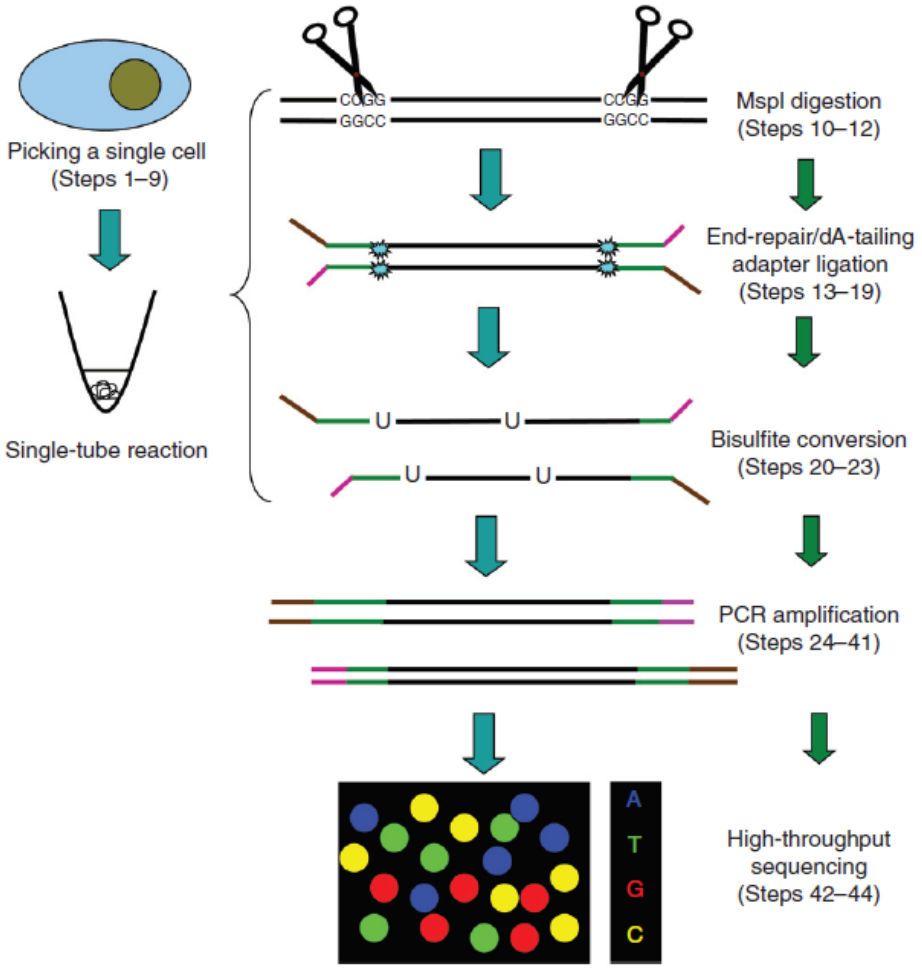



**Figure 1** | Flowchart of the experimental procedures of the scRRBS technique. Notably, we integrated cell lysis, MspI digestion, end repair/dA tailing, adapter ligation and bisulfite treatment into a single-tube reaction to avoid unnecessary DNA loss.