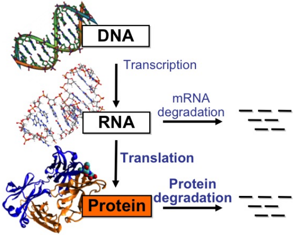
# Exploiting spatial patterns in the analysis of BS-Seq data.

Guido Sanguinetti

School of Informatics and SynthSys, University of Edinburgh

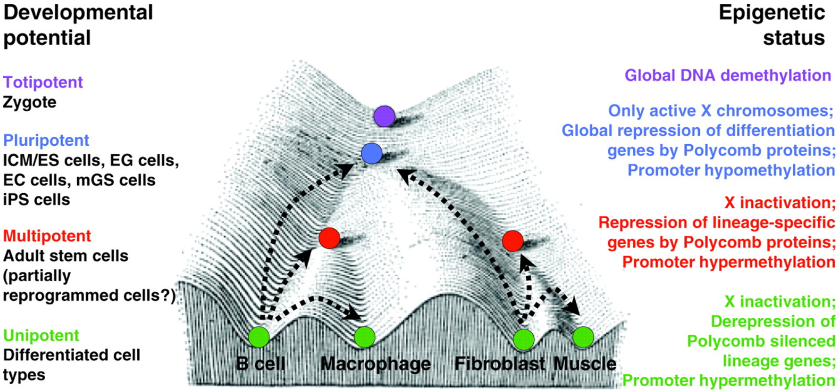
Karstenfest 10/2016

# The central dogma



Where does variability come into play? What can we measure?

# Epigenetics



**Developmental potential**

**Totipotent**
**Zygote**

**Pluripotent**
**ICM/ES cells, EG cells, EC cells, mGS cells iPS cells**

**Multipotent**
**Adult stem cells (partially reprogrammed cells?)**

**Unipotent**
**Differentiated cell types**

**Epigenetic status**

**Global DNA demethylation**

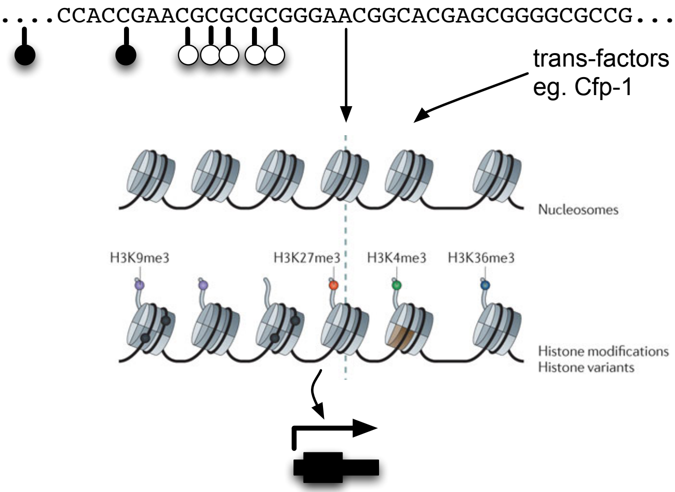
**Only active X chromosomes; Global repression of differentiation genes by Polycomb proteins; Promoter hypomethylation**

**X inactivation; Repression of lineage-specific genes by Polycomb proteins; Promoter hypermethylation**

**X inactivation; Derepression of Polycomb silenced lineage genes; Promoter hypermethylation**

A modeller's dream!

# A more accurate picture?



**Genome**
DNA sequence

....CCACCGAACGCGCGGAACGGCACGAGCGGGGCGCCG...

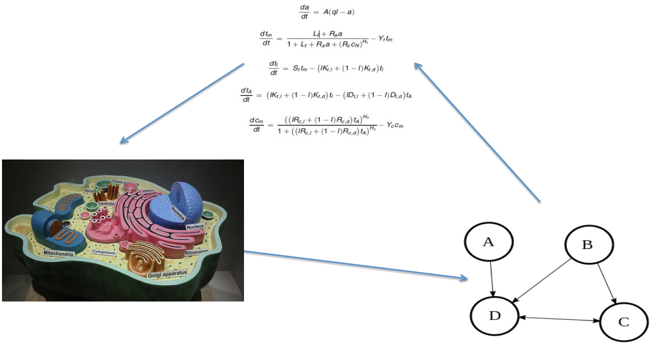
trans-factors
eg. Cfp-1

**Epigenome**

**Transcriptome**
RNA-Seq / Pol-II

Zhou *et al.*, Nat Rev Genet, 2011

# The modelling cycle



$$\frac{da}{dt} = A(ql - a)$$

$$\frac{dt_m}{dt} = \frac{L_t + R_a a}{1 + L_t + R_a a + (R_c c_N)^{H_t}} - Y_t t_m$$

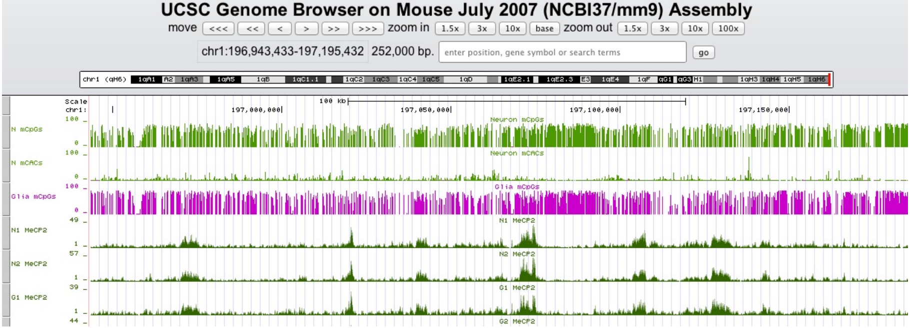
$$\frac{dt_l}{dt} = S_t t_m - (lK_{t,l} + (1-l)K_{t,d})t_l$$

$$\frac{dt_A}{dt} = (lK_{t,l} + (1-l)K_{t,d})t_l - (lD_{t,l} + (1-l)D_{t,d})t_A$$

$$\frac{dc_m}{dt} = \frac{((lR_{c,l} + (1-l)R_{c,d})t_A)^{H_c}}{1 + ((lR_{c,l} + (1-l)R_{c,d})t_A)^{H_c}} - Y_c c_m$$

Informatics will provide the synthesis!

# Epigenetics: what the data looks like



UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

chr1:196,943,433-197,195,432 252,000 bp.

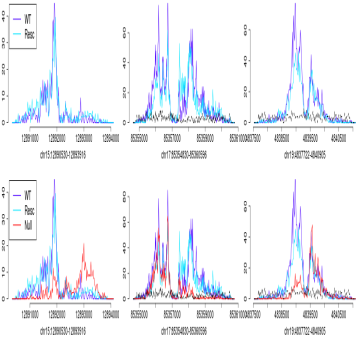
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
enter position, gene symbol or search terms
go
chr1 (qH6) 1qA1 A2 1qA3 1qA5 1qB 1qC1.1 1qC2 1qC3 1qC4 1qC5 1qD 1qE2.1 1qE2.3 E3 1qE4 1qF qG1 qG3 H1 1qH3 1qH4 1qH5 1qH6
Scale
chr1:
100 kb
197,000,000
197,050,000
197,100,000
197,150,000
N mCpGs
100
0
Neuron mCpGs
N mCACs
100
0
Neuron mCACs
Glia mCpGs
100
0
Glia mCpGs
N1 MeCP2
49
1
N1 MeCP2
N2 MeCP2
57
1
N2 MeCP2
G1 MeCP2
39
1
G1 MeCP2
44
G2 MeCP2

Each row is a tiny fraction of a next-generation sequencing experiment's data. Each row $\geq$1GB of data.

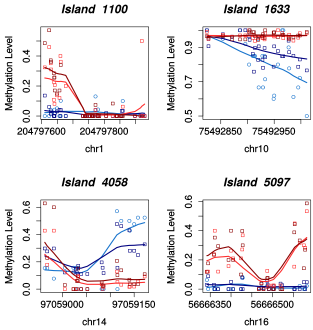
# What the data looks like

after QC, mapping, alignment,

Histone modification data



DNA Methylation data



# Obvious problems

- Small data, with each data point being very big

# Obvious problems

- Small data, with each data point being very big
- Even restricting to regions (e.g. genes), the data is high dimensional and non-trivial

# Obvious problems

- Small data, with each data point being very big
- Even restricting to regions (e.g. genes), the data is high dimensional and non-trivial
- How can we even determine statistical differences?

# Obvious problems

- Small data, with each data point being very big
- Even restricting to regions (e.g. genes), the data is high dimensional and non-trivial
- How can we even determine statistical differences?
- What is a suitable probability model for each of these high-dimensional, non-Gaussian items?

# Obvious problems

- Small data, with each data point being very big
- Even restricting to regions (e.g. genes), the data is high dimensional and non-trivial
- How can we even determine statistical differences?
- What is a suitable probability model for each of these high-dimensional, non-Gaussian items?
- Data associated with different genes may be of intrinsically different dimensionality. How can I do even basic things like clustering?

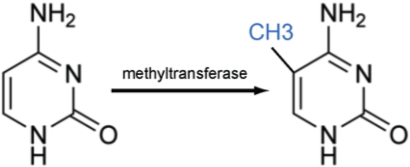
# Obvious problems

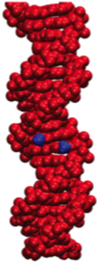
- Small data, with each data point being very big
- Even restricting to regions (e.g. genes), the data is high dimensional and non-trivial
- How can we even determine statistical differences?
- What is a suitable probability model for each of these high-dimensional, non-Gaussian items?
- Data associated with different genes may be of intrinsically different dimensionality. How can I do even basic things like clustering?
- How can we model in the presence of very strong redundancies (dimensionality reduction)?

# Talk outline

1. Background
2. Shape-based testing for methylation profiles (T. Mayo)
3. Spatial methylation and gene expression (Andreas Kapourani)

# DNA Methylation





- Addition of a methyl group to a cytosine
- Predominantly occurs in the CpG context
- Tightly controlled epigenetic phenomenon

# DNA Methylation - why study it?

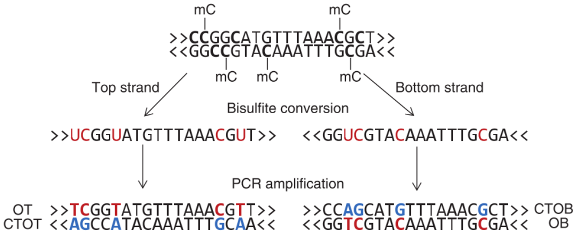
DNA methylation has been associated with

- Cellular processes: genomic imprinting, cell differentiation, retrotransposon silencing, gene regulation
- Diseases: Cancer, heart disease.
- Canonical view: methylation of promoters (CpG islands) silences gene

As such, epigenetic therapies are being developed which specifically target methylation

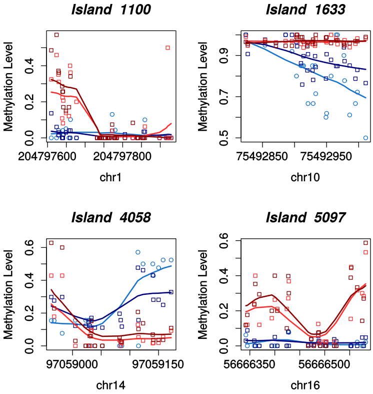
Epigenome-wide association studies (EWAS) incorporating methylation

# Methylation Data



- Bisulfite conversion: unmethylated Cytosine to Uracil
- NGS, conversion aware alignment
- RRBS: focus on CpG-rich regions

# A look at the data



Data exhibits strong spatial correlations conserved across replicates

# Existing methods

Typical approaches test individual cytosines and aggregate (not MAGI).

BSmooth

- Uses local likelihood smoothing to filter noise
- Replicates are aggregated to a single methylation profile

MethylSig & BiSeq

- Beta-binomial approach to model variability, at each cytosine
- Differ in approach to multiple comparison testing

MAGI

- Pre-selects regions and assigns global methylation state via thresholding
- Uses Fisher exact test on binary string

# Existing Methods: Problems

In general:

- Require high replication & coverage
- Loss of significance due to multiple comparisons
- Ignore spatial correlations in the data
- Hence, require uninterrupted, large methylation changes to occur at individual Cs.

Beta-Binomial methods:

- Require large number of replicates
- Require high coverage at each C in large number of samples
- Variability is modelled individually at each cytosine

# Formulate the test question

We wish to test whether the methylation profile in a region is different between two samples.

# Formulate the test question

We wish to test whether the methylation profile in a region is different between two samples.

**Idea**: treat data as outcome of a generative process where CpG sites are randomly assigned reads and methylation state on each read

- $n$ observations in data set $s$ (e.g. WT)

$$X^s = \{\mathbf{x}_1^s, ..., \mathbf{x}_n^s\}$$

- $m$ observations in data set $s'$ (e.g. Null),

$$X^{s'} = \{\mathbf{x}_1^{s'}, ..., \mathbf{x}_m^{s'}\}$$

where $\mathbf{x}^s$, $\mathbf{x}^{s'}$ random variables
drawn i.i.d. from probability distributions $p$ and $p'$.

# Formulate the test question

We wish to test whether the methylation profile in a region is different between two samples.

**Idea**: treat data as outcome of a generative process where CpG sites are randomly assigned reads and methylation state on each read

- $n$ observations in data set $s$ (e.g. WT)

$$X^s = \{\mathbf{x}_1^s, \ldots, \mathbf{x}_n^s\}$$

- $m$ observations in data set $s'$ (e.g. Null),

$$X^{s'} = \{\mathbf{x}_1^{s'}, \ldots, \mathbf{x}_m^{s'}\}$$

where $\mathbf{x}^s$, $\mathbf{x}^{s'}$ random variables
drawn i.i.d. from probability distributions $p$ and $p'$.

**Can we decide whether $p \neq p'$?**

# MMD: non-parametric testing for distributions

- MMD: Kernel-based non-parametric test
- recently developed by Gretton et al., 2008, 2012
- retains information of any order within the testing procedure.

# MMD: non-parametric testing for distributions

- MMD: Kernel-based non-parametric test
- recently developed by Gretton et al., 2008, 2012
- retains information of any order within the testing procedure.

## Maximum Mean Discrepancy (MMD)

Starting point:
Define feature map, which maps the distributions into a high dimensional reproducing Kernel Hilbert Space (RKHS).

In this space, two distributions are identical
if and only if their kernel mean is identical.

Distance between means is a good quantitative measure for difference between two distributions.

# MMD Test statistics

- Nonlinear kernel function $k(\mathbf{x}^s, \mathbf{x}^{s'})\rightarrow$ the *mean embedding* of a distribution $p$ (in the RKHS $\mathcal{F}$) contains the information of all higher-order moments.

# MMD Test statistics

- Nonlinear kernel function $k(\mathbf{x}^s, \mathbf{x}^{s'}) \rightarrow$ the *mean embedding* of a distribution $p$ (in the RKHS $\mathcal{F}$) contains the information of all higher-order moments.
- The *maximum mean discrepancy, (MMD)* is the distance between mean embeddings

$$MMD[\mathcal{F}, p, p'] = sup_{f \in \mathcal{F}}(\mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{x \sim p'}[f(x')])$$

# MMD Test statistics

- Nonlinear kernel function $k(\mathbf{x}^{s}, \mathbf{x}^{s'})\rightarrow$ the *mean embedding* of a distribution $p$ (in the RKHS $\mathcal{F}$) contains the information of all higher-order moments.
- The *maximum mean discrepancy, (MMD)* is the distance between mean embeddings

$$MMD[\mathcal{F}, p, p'] = sup_{f \in \mathcal{F}}(\mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{x \sim p'}[f(x')])$$

- Theorem: $MMD^{p,p'} = 0$ if and only if $p = p'$

# MMD Test statistics

- Nonlinear kernel function $k(\mathbf{x}^s, \mathbf{x}^{s'})\rightarrow$ the *mean embedding* of a distribution $p$ (in the RKHS $\mathcal{F}$) contains the information of all higher-order moments.
- The *maximum mean discrepancy, (MMD)* is the distance between mean embeddings

$$MMD[\mathcal{F}, p, p'] = sup_{f\in\mathcal{F}}(\mathbf{E}_{x\sim p}[f(x)] - \mathbf{E}_{x\sim p'}[f(x')])$$

- Theorem: $MMD^{p,p'} = 0$ if and only if $p = p'$
- Finite sample estimates of MMD will be different from zero, but their distribution can be estimated (by bootstrapping)
- MMD can be efficiently computed in terms of Kernel functions

$$MMD^{(s,s')} = \left[\frac{1}{(n)^2}k(\mathbf{x}^s, \mathbf{x}^s) - \frac{2}{n\cdot m}k(\mathbf{x}^s, \mathbf{x}^{s'}) + \frac{1}{m^2}k(\mathbf{x}^{s'}, \mathbf{x}^{s'})\right]^{\frac{1}{2}}$$

# Choice of Kernel

Each mapped cytosine is an individual data point: $x_j = (C_j, Meth_j)$

```
  ATGGCATTGCAA
   TGGCATTGCAATTTG
AGATGGTATTG
```

Composite kernel

- $k_{full}(x_i, x_j) = k_{RBF}(x_i, x_j) k_{STR}(x_i, x_j)$
- $k_{RBF}(x_i, x_j) = exp[-(C_i - C_j)^2 / 2\sigma^2]$
- $k_{STR}(x_i, x_j) = 1$ if $Meth_i = Meth_j$, 0 else

$\sigma$ is modelled from the data as $\sigma^2 = \bar{x}^2/2$ where $\bar{x}$ is the median observed distance in the region.

# Handling Coverage

- The MMD tests whether samples are drawn from the same distribution.
- The frequency that data is drawn - the coverage - is independent of the methylation profile.
- We adapt the method by subtracting an appropriate 'coverage only' metric.
- The MMD with an RBF kernel on genomic location only (no methylation considered)

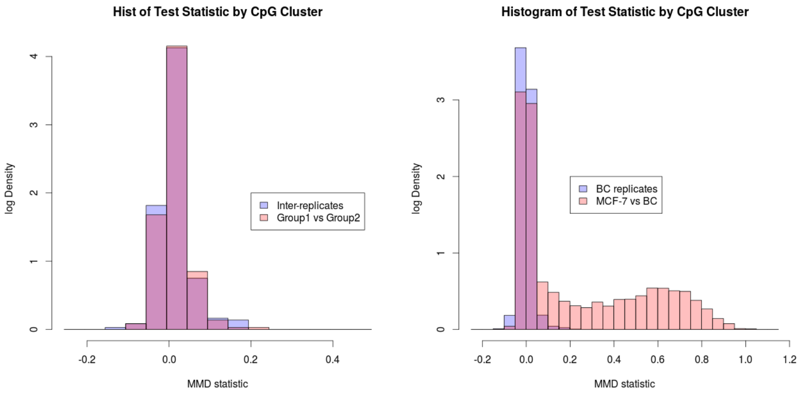
# Test-statistic

M$^3$D test-statistic

$$M^3D[X, Y] = MMD[X, Y, k_{full}] - MMD[X, Y, k_{RBF}]$$

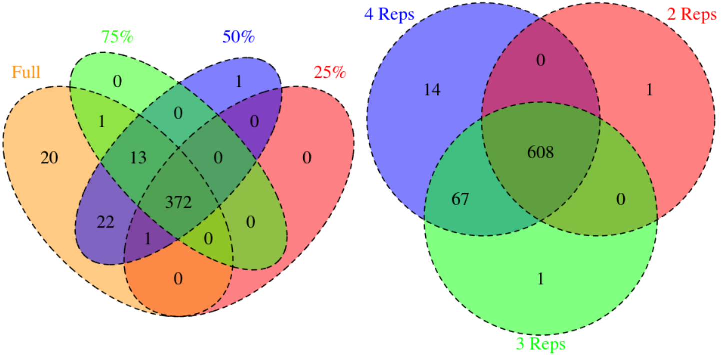
- The test statistic over all replicate pairs forms our testing distribution
- For a given region, the mean of the inter-group comparisons is tested against this distribution
- This gives the empirical probability of finding the cross-group difference in methylation profiles among the replicates

# $M^3D$ produces nice histograms



$M^3D$ statistic between replicates (left) and between different conditions (K562 vs H1 cells).

# $M^3D$ is robust to low replication/ coverage



$M^3D$ test results is robust to low coverage (left) and low replication (right).

# Talk outline

1 Background

2 Shape-based testing for methylation profiles (T. Mayo)

3 Spatial methylation and gene expression (Andreas Kapourani)

# Spatial methylation patterns

- Spatial methylation patterns appear to be strongly reproducible hence they yield a very powerful test
- Do they mean anything?

# Spatial methylation patterns

- Spatial methylation patterns appear to be strongly reproducible hence they yield a very powerful test
- Do they mean anything?
- To answer this question, we need to quantify precisely methylation patterns of regions
- M3D avoided the issue using the kernel trick
- Quantifying patterns is tricky as different regions have different numbers of CpGs

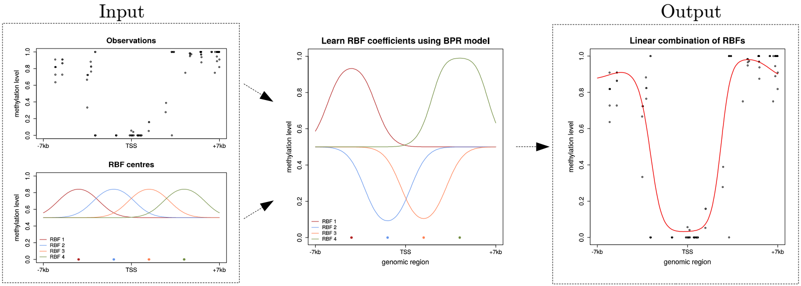
# The BPRM model

- We assume the methylation pattern of a region to be determined by an unobserved methylation function $f(x) = \Phi\big(g(x)\big)$, where $\Phi$ is the probit transform, defined on the whole region (not just CpGs)
- We represent the unconstrained function $g(x) = \mathbf{w}\xi(\mathbf{x})$ as a linear combination of fixed basis functions $\xi_j$ (e.g. RBF)
- The actual number of methylated reads at position $i$ is binomial distributed

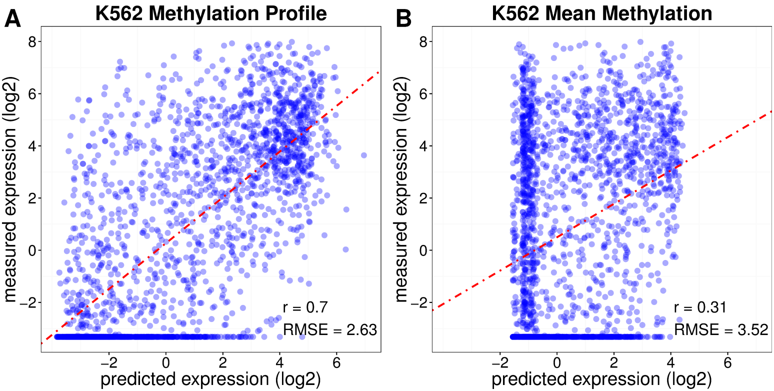
  $$n_i \sim \mathrm{Bin}\left(m_i, f(x_i)\right) \tag{1}$$

  with $m_i$ the coverage at position $i$.
- Optimising the likelihood given by (1) w.r.t. the weights $\mathbf{w}$ associates each region with *methylation profile features*

# The BPRM model - cartoon

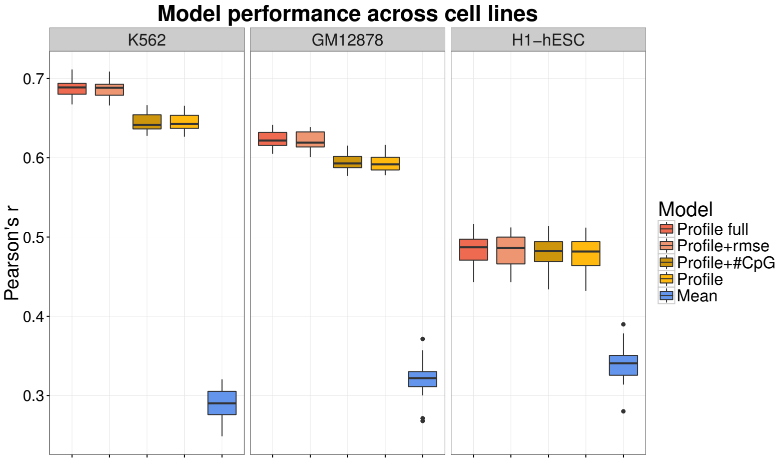


# Predicting gene expression



Predicting gene expression from methylation profiles (left) or mean methylation levels (right). Overall improvement in Pearson *r* from 0.31 to 0.72.

# Effect of different features



BPRM model predictions on different cell lines/ using different features.

# Conclusions

- MMD-based statistics enable more powerful tests than currently used approaches
- MMDiff is complementary to count-based methods: changes that only alter counts (keeping shape fixed) cannot be captured
- MMD is potentially of use in other scenarios where distributions arise naturally, e.g. methylation or metagenomics
- Machine learning can help extract patterns from high-throughput epigenomic data which may suggest biological functions/ clarify links between epigenetics and gene regulation

# Thanks

**School of Informatics**

- Gabriele Schweikert
- Tom Mayo
- Andreas Kapourani

**Wellcome Trust Centre for Cell Biology**

- Adrian Bird

**MRC-HGU/ IGMM**

- Duncan Sproul

Funding: EU FP7 Marie Curie Actions, ERC, EPSRC.

# References

- G. Schweikert et al, MMDiff: quantitative testing for shape changes in ChIP-Seq data sets, BMC Genomics 14:826, 2013
- MMDiff2 bioconductor package

  `http://www.bioconductor.org/packages/release/bioc/html/MMDiff2.html`
- T. Mayo et al, $M^3D$: a kernel-based test for spatially correlated changes in methylation profiles, Bioinformatics 31(6), 809-816, 2015
- M3D bioconductor package

  `http://www.bioconductor.org/packages/devel/bioc/html/M3D.html`
- A. Kapourani and G.S., Higher order methylation features for clustering and prediction in epigenomic studies, Bioinformatics **32**(17), i405-i412, 2016 (Proc of ECCB16)