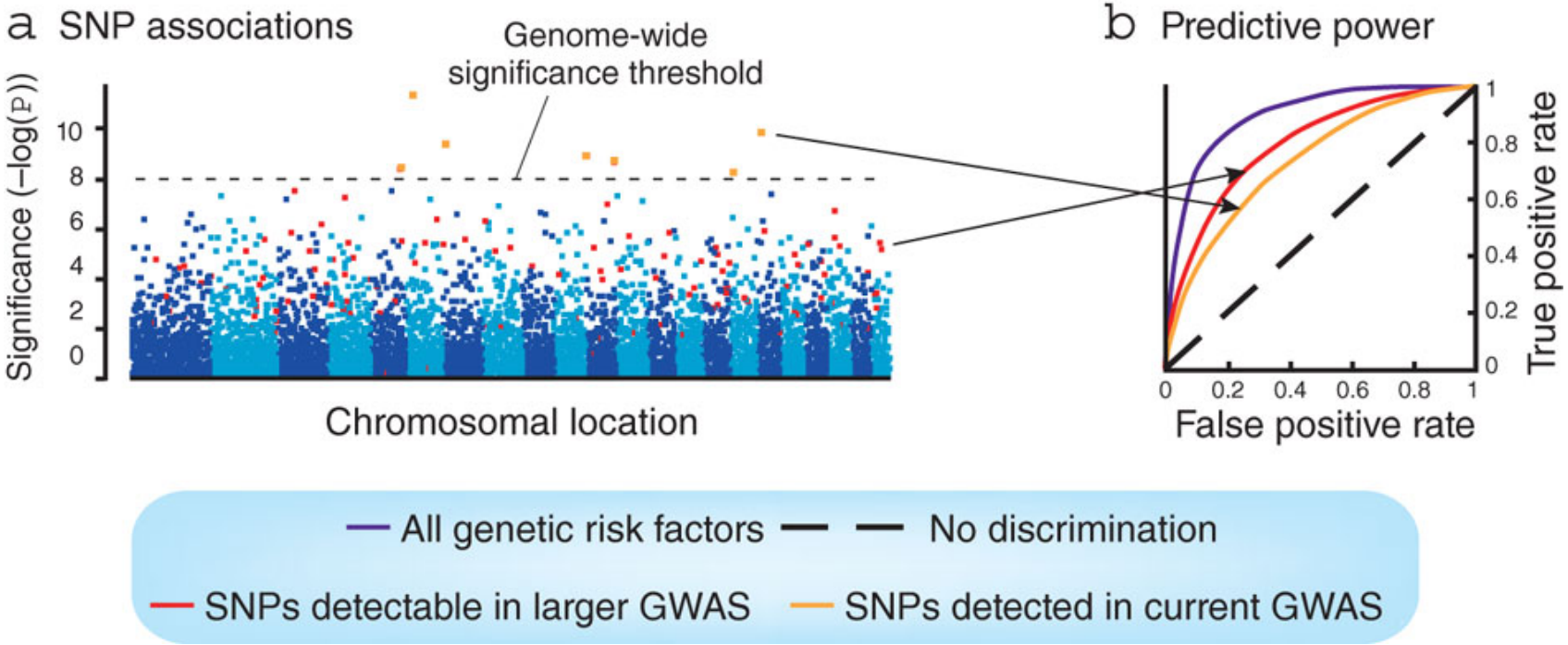
Machine learning methods for splicing quantification at extremely low coverage

Guido Sanguinetti

School of Informatics, University of Edinburgh

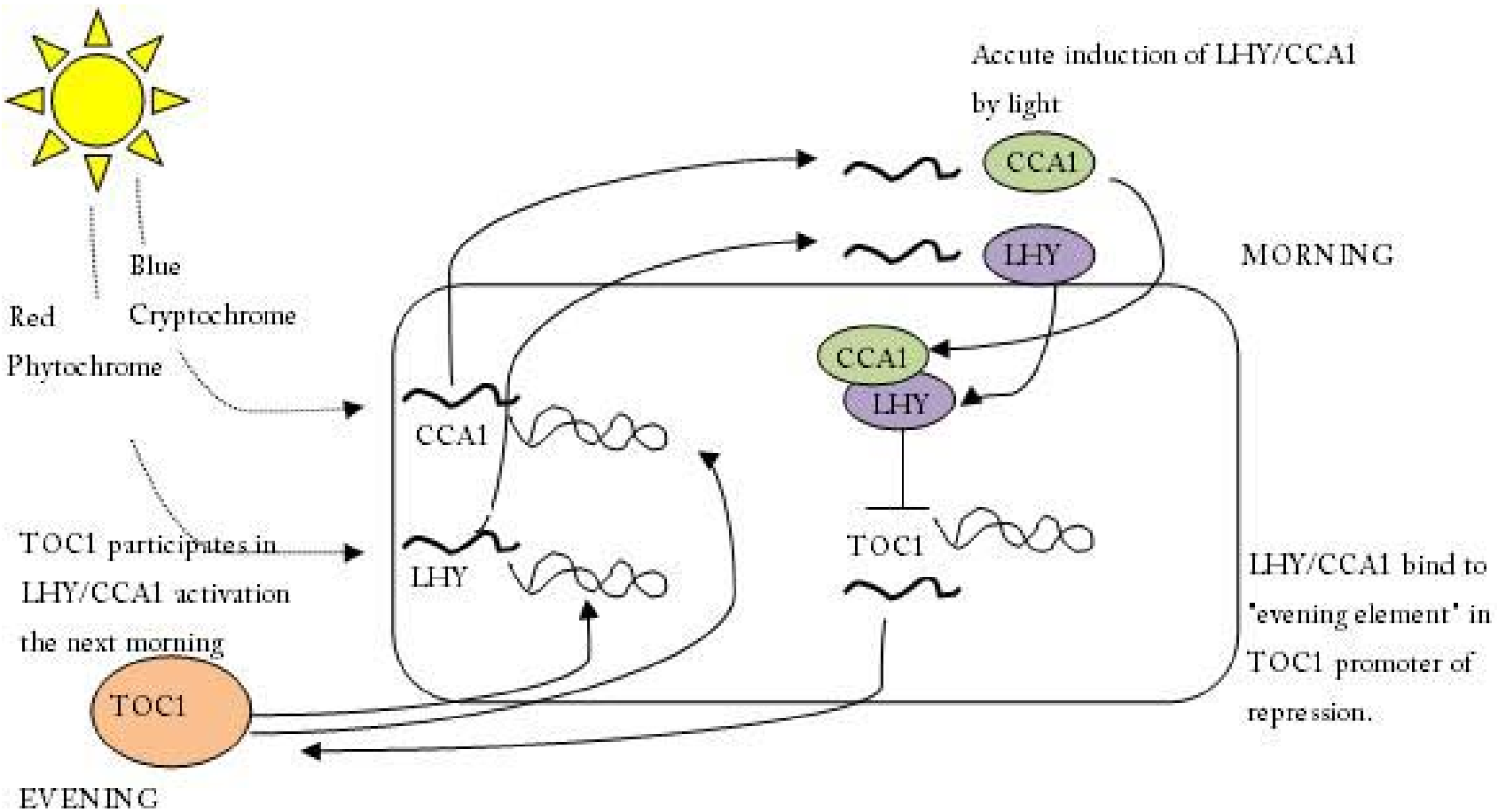
ICTP 2017

From genotype to phenotype



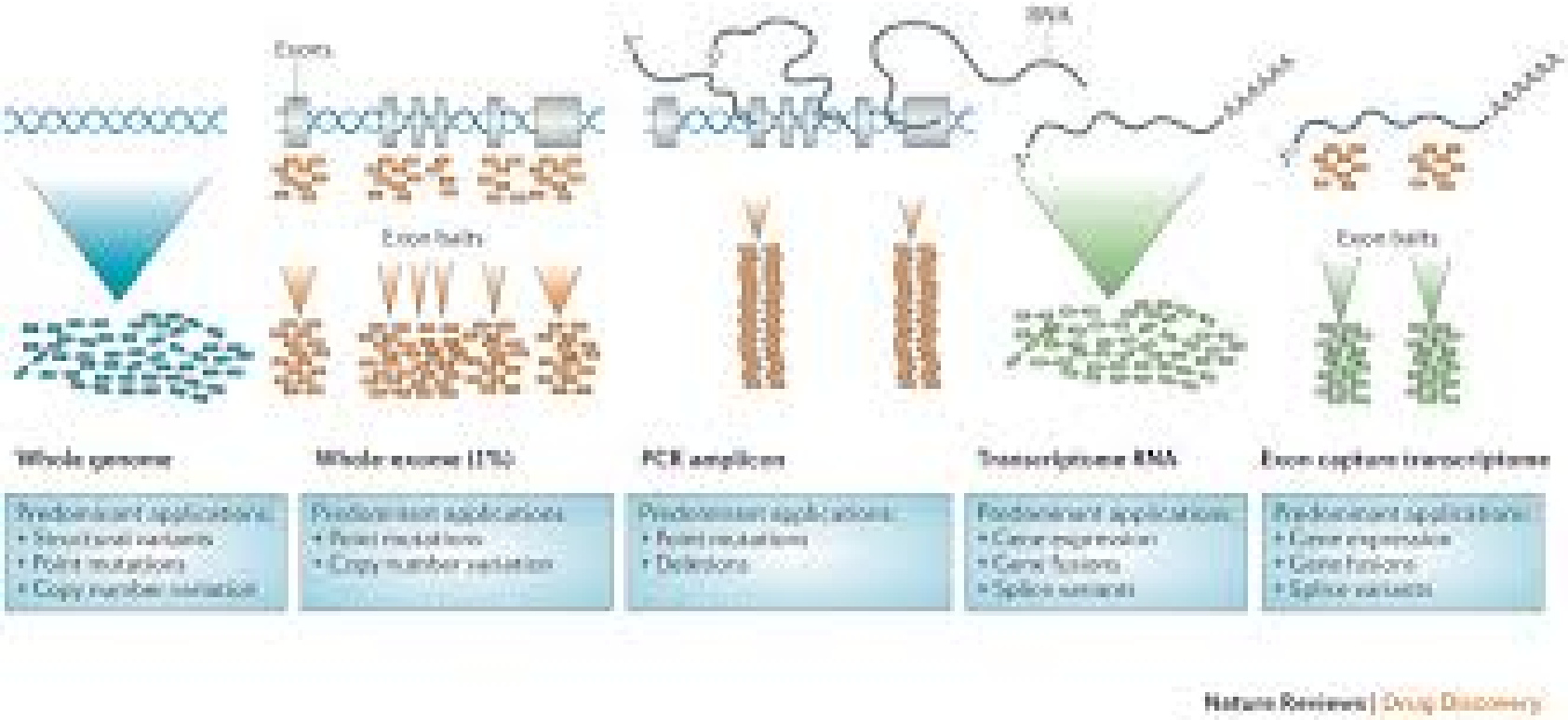
Directly estimating a map: GWAS. Computational tools: classical hypothesis testing, regularised regression

From genotype to phenotype



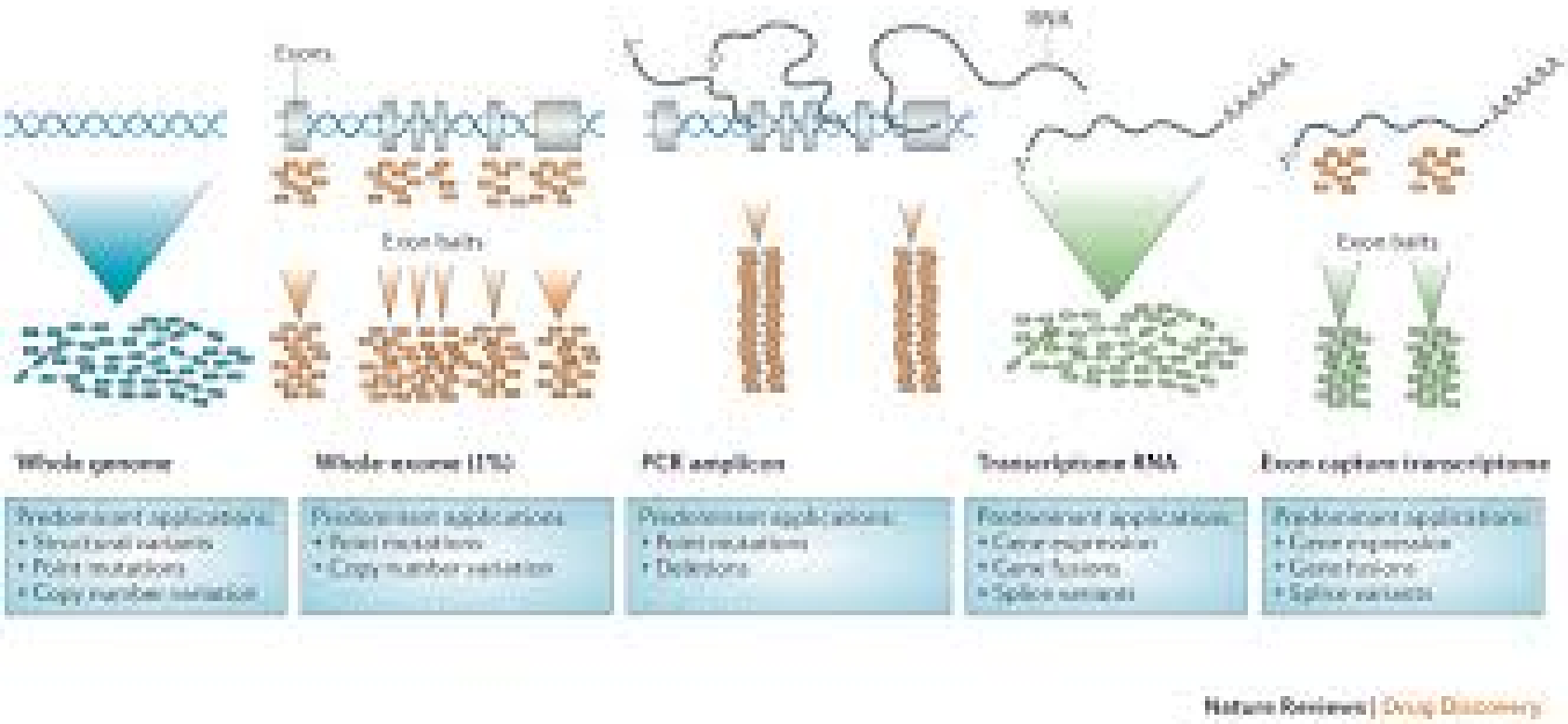
Open the box: targeted quantitative experiments on well-defined subsystems. Computational tools: dynamical systems

From genotype to phenotype



Blow-up the box: high-throughput experiments. Computational tools/ questions?

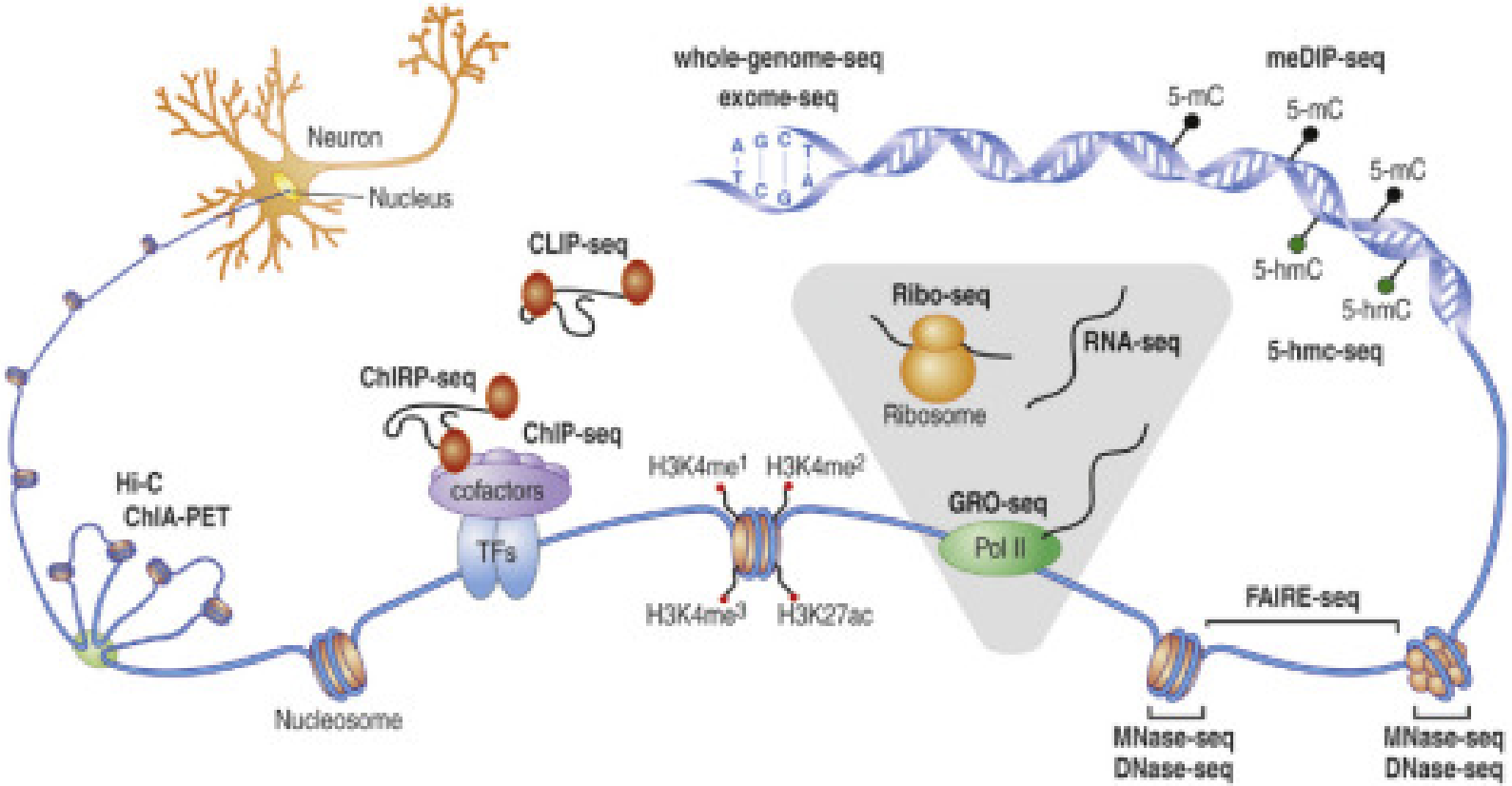
From genotype to phenotype



Blow-up the box: high-throughput experiments. Computational tools/ questions?

All of above, plus a good dose of high-dimensional statistics and machine learning, lots of open methodological problems $\rightarrow$ **model-based bioinformatics!**

NGS and gene expression



Each facet of gene expression measured by different NGS techniques.

What is Machine Learning?

- Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed [Wikipedia].

What is Machine Learning?

- Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed [Wikipedia].
- Really?

What is Machine Learning?

- Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed [Wikipedia].
- Really?
- My definition: ML algorithms aim at encoding mathematically predictive relationships hidden within the data
- Generally balances human input (prior model assumptions) and data
- Statistics ++ and optimisation

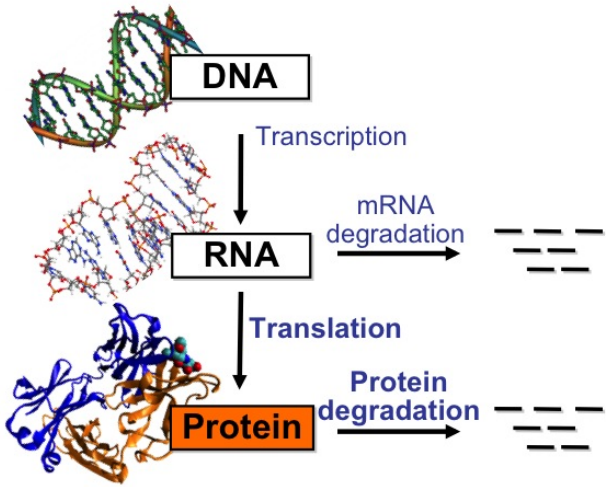
What is Machine Learning?

- Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed [Wikipedia].
- Really?
- My definition: ML algorithms aim at encoding mathematically predictive relationships hidden within the data
- Generally balances human input (prior model assumptions) and data
- Statistics ++ and optimisation
- Philosophy: taking a global look (i.e. integrating multiple sources of information) to see further in the detail

Talk outline

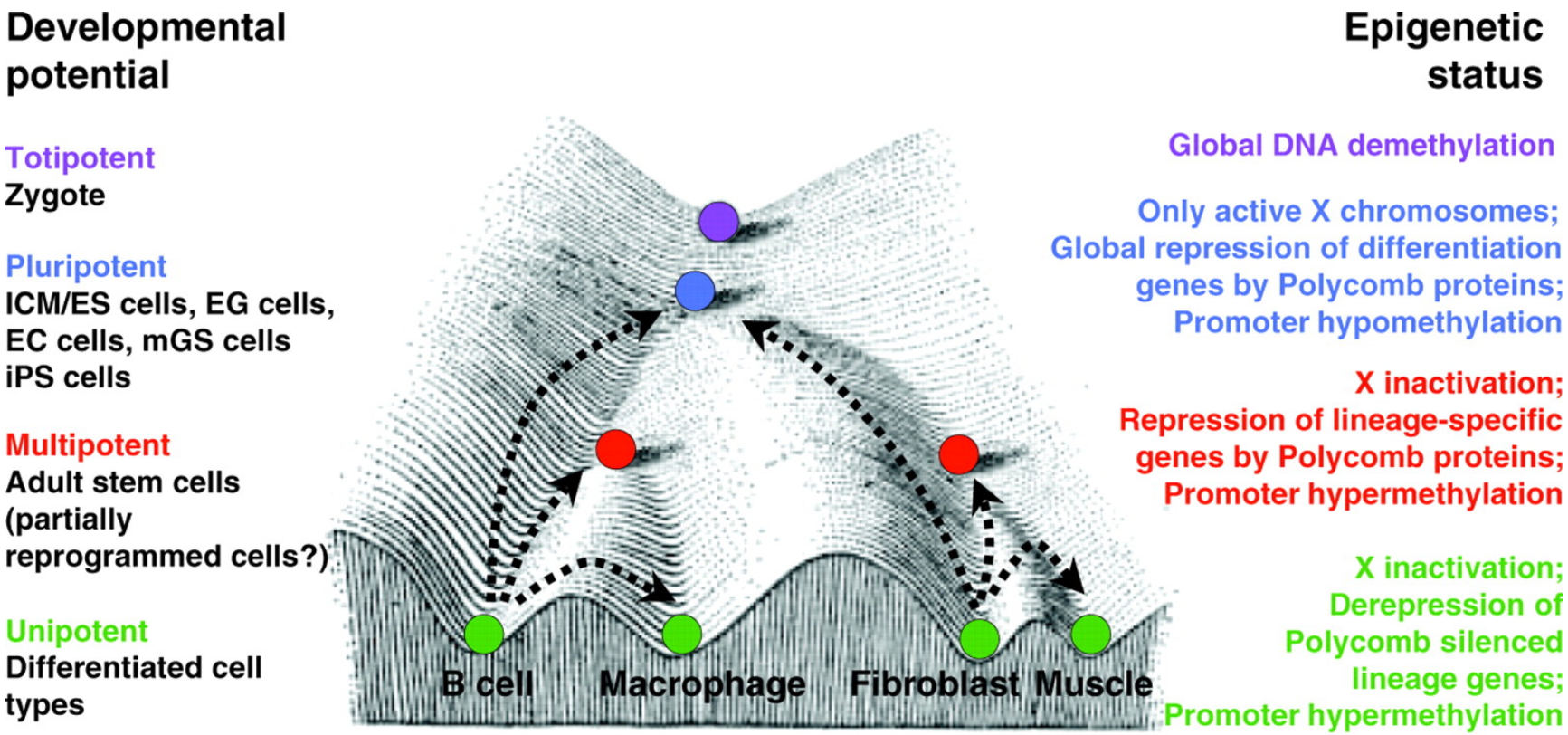
1. Spatial effects in epigenomic data
   - Statistical testing for epigenomic data
   - Transcription factors and histone modifications
   - Clustering and prediction from epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
   - Isoform quantification from time series RNA-seq
   - Splicing quantification in single cells
3. Conclusions

The central dogma



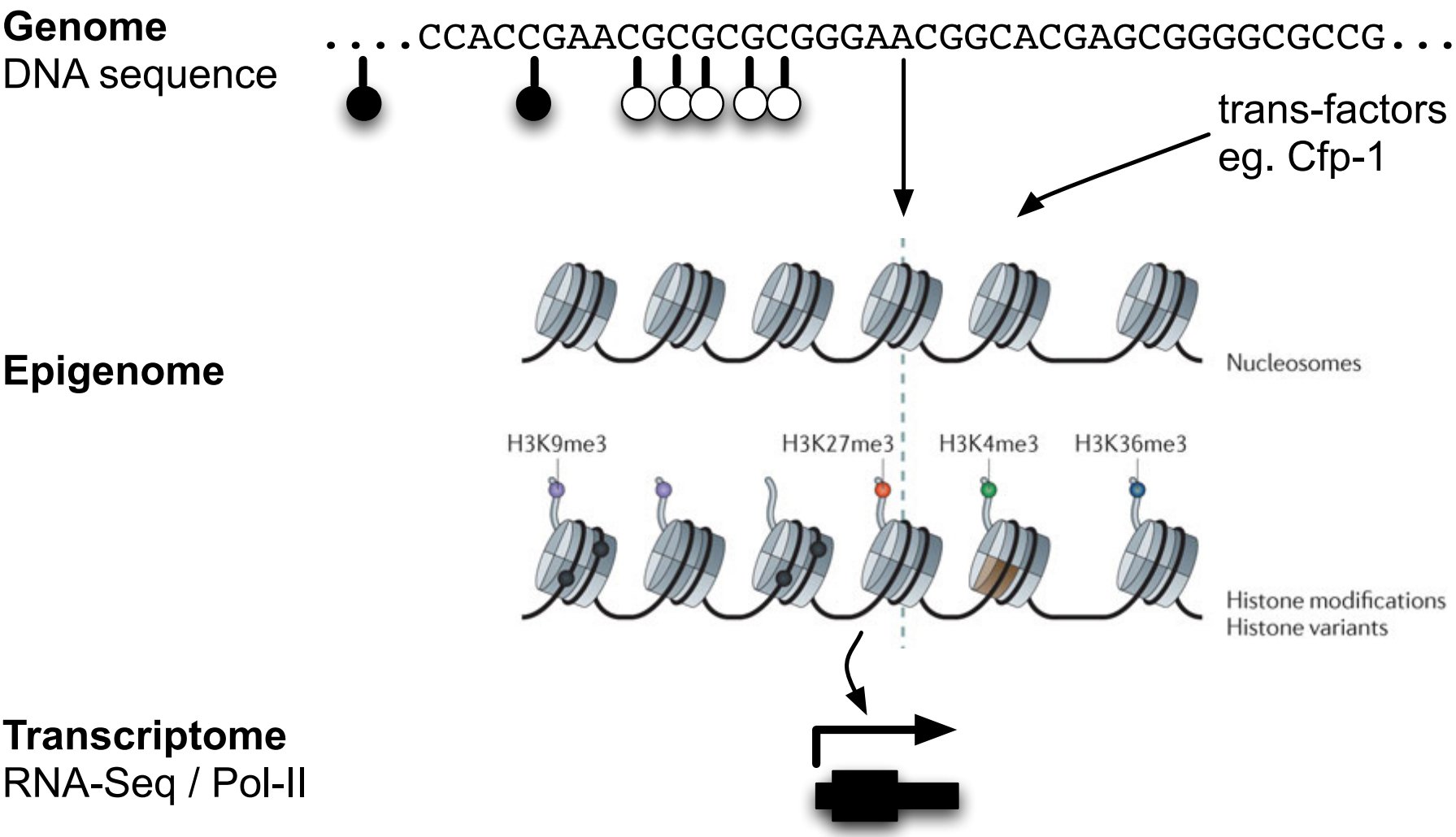
Where does variability come into play? What can we measure?

Epigenetics



A modeller's dream!

A more accurate picture?



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

The modelling cycle

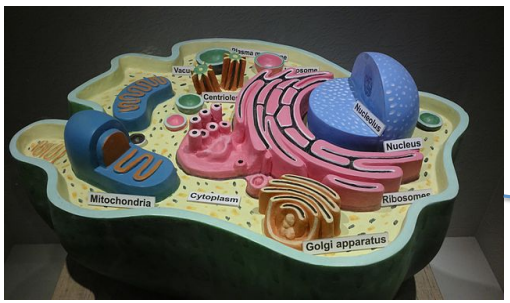
$$\frac{da}{dt} = A(ql - a)$$

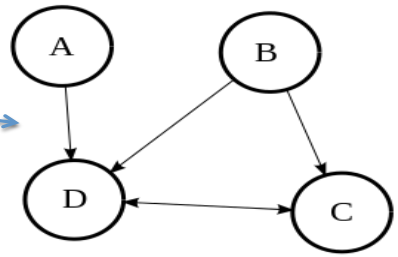
$$\frac{dt_m}{dt} = \frac{L_t l + R_a a}{1 + L_t + R_a a + (R_c c_N)^{H_t}} - Y_t t_m$$

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

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

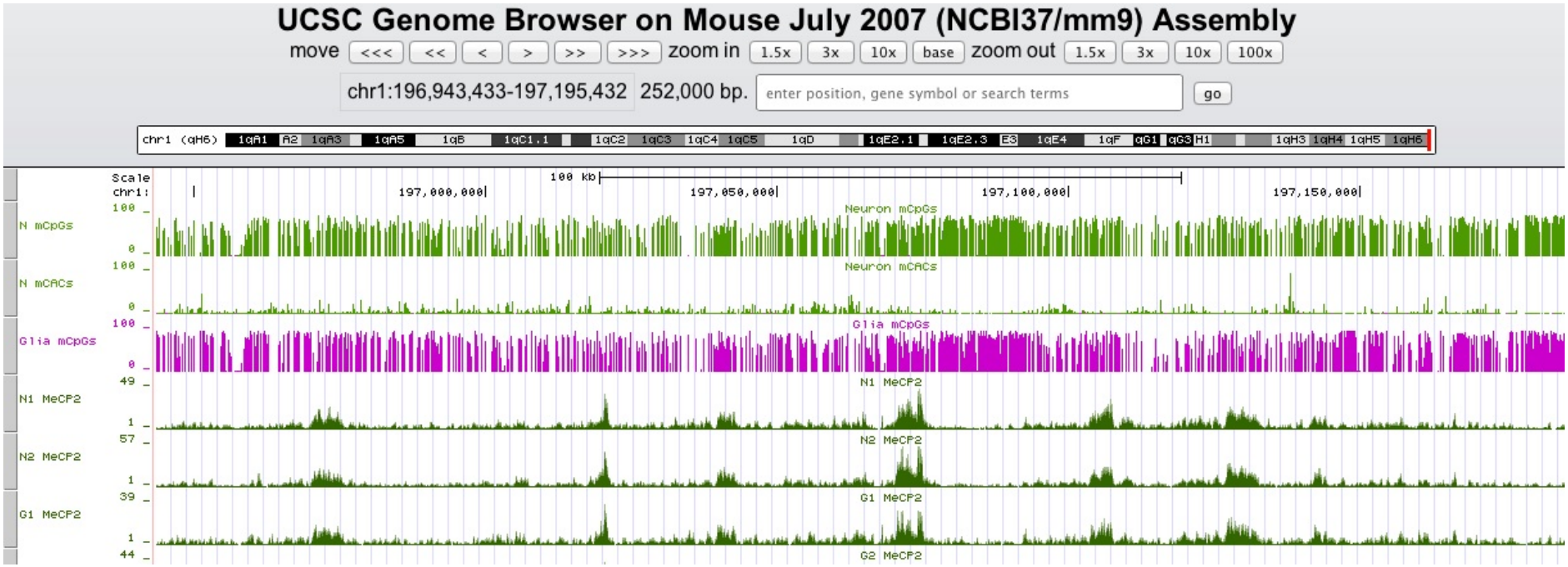
$$\frac{dc_m}{dt} = \frac{((lR_{c,l} + (1-l)R_{c,d})t_A)^{H_c}}{1 + ((lR_{c,l} + (1-l)R_{c,d})t_A)^{H_c}} - Y_c c_m$$





Informatics will provide the synthesis!

Epigenetics: what the data looks like

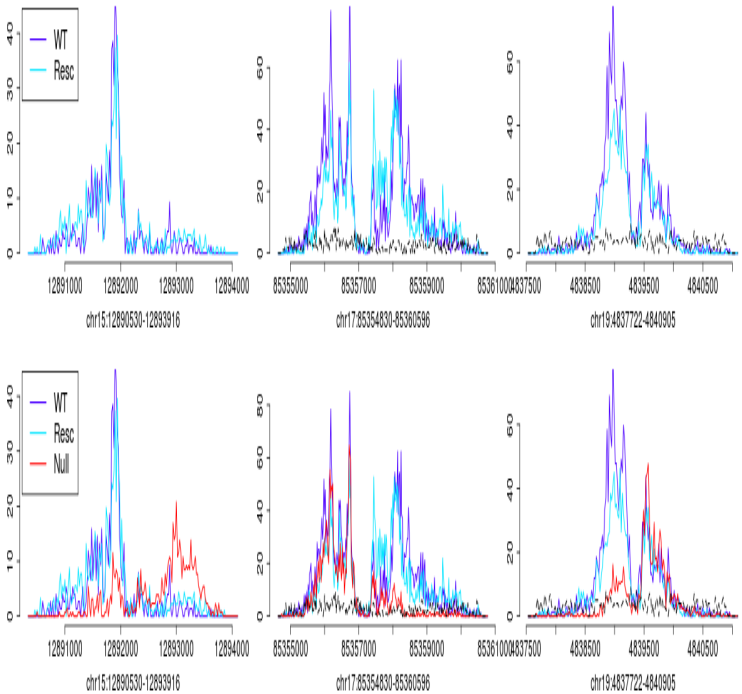


Each row is a tiny fraction of a next-generation sequencing experiment's data. Each row $\geq$1GB of data.

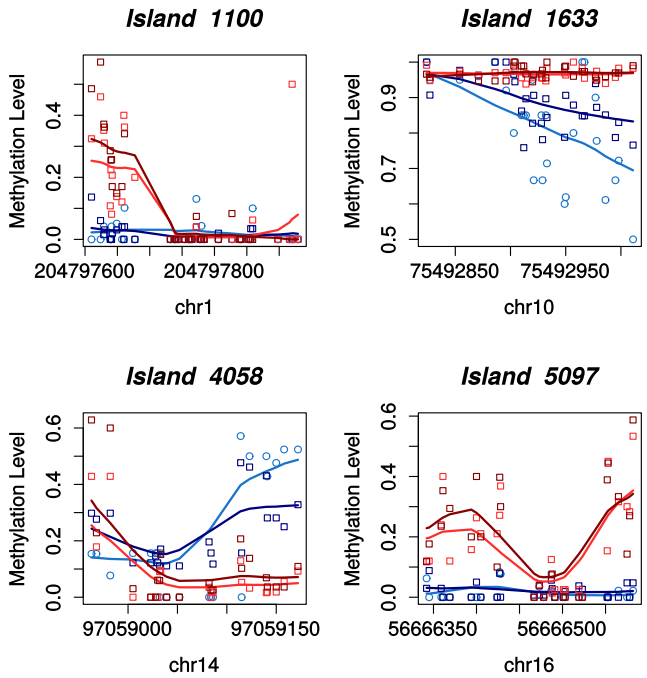
What the data looks like

. . . after QC, mapping, alignment, . . .

Histone modification data



DNA Methylation data

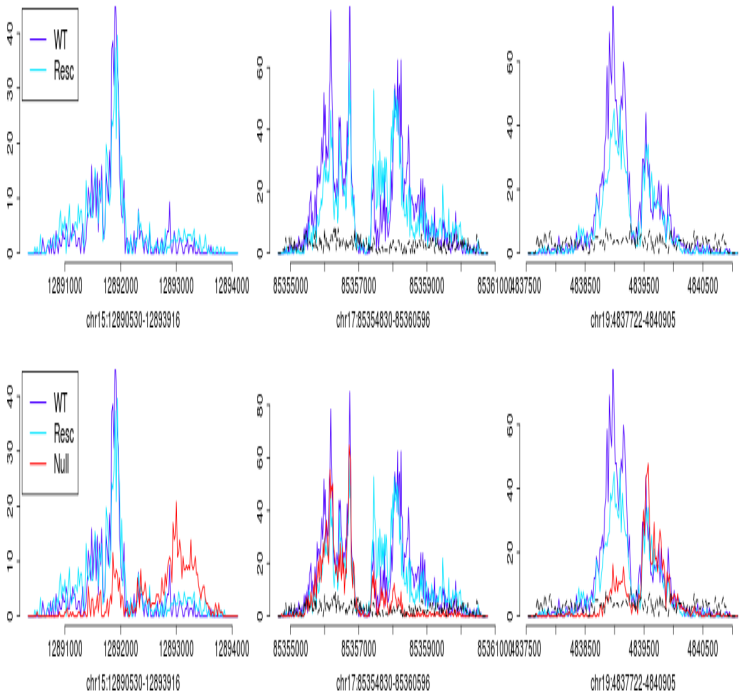


Even basic questions (significant difference?) are hard

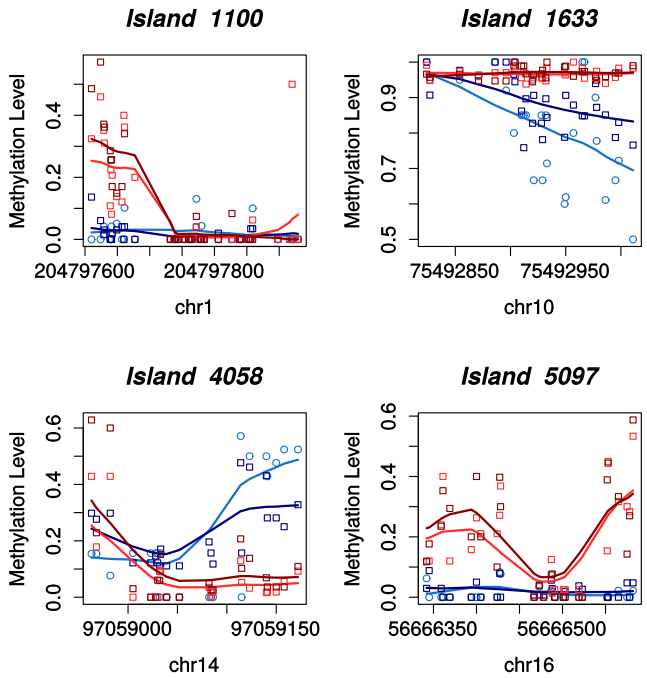
What the data looks like

. . . after QC, mapping, alignment, . . .

Histone modification data



DNA Methylation data



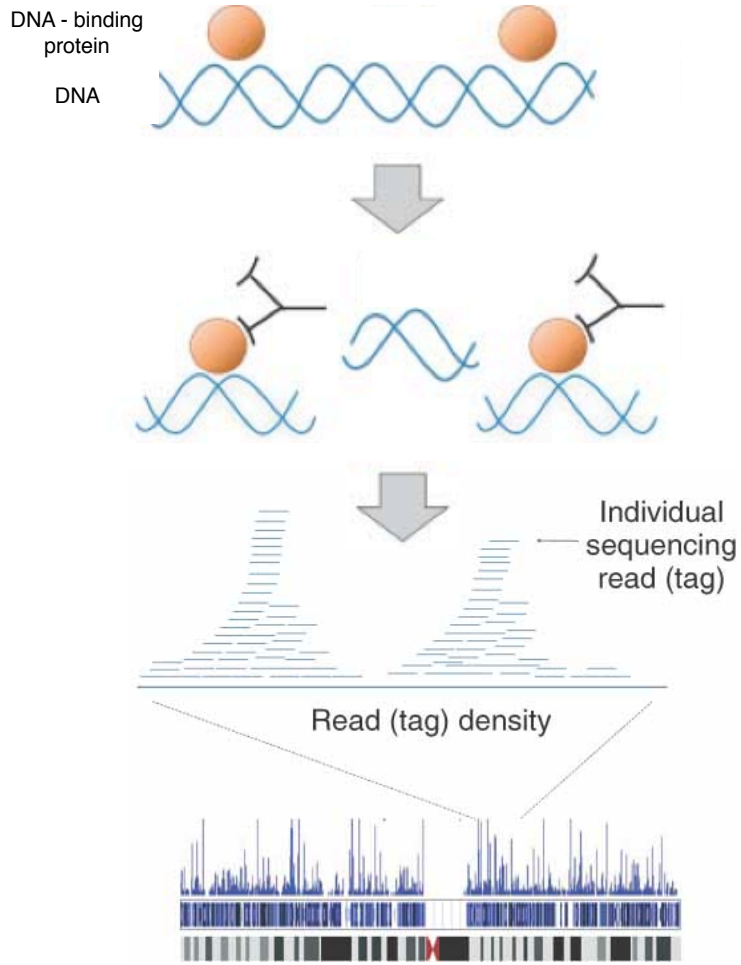
Even basic questions (significant difference?) are hard

Maybe local correlations can be exploited to borrow statistical strength?

Talk outline

1. Spatial effects in epigenomic data
   - Statistical testing for epigenomic data
   - Transcription factors and histone modifications
   - Clustering and prediction from epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
   - Isoform quantification from time series RNA-seq
   - Splicing quantification in single cells
3. Conclusions

Introduction: ChIP-Seq



- Cross-linking

- DNA fragmentation

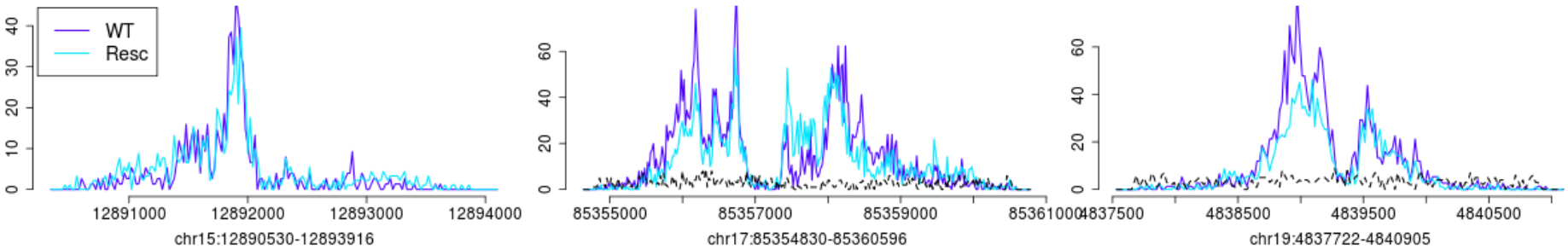
- Enrichment with specific antibody (ChIP)

- Profiling of enriched DNA (Seq)

Kim and Park, 2011

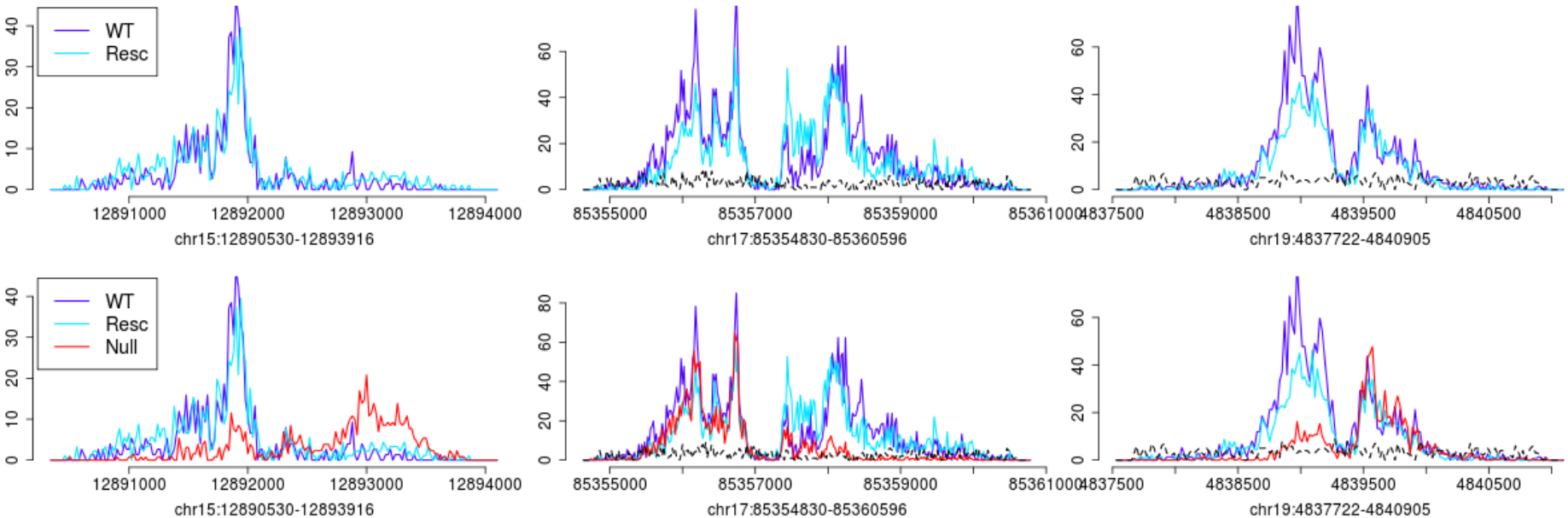
What the data looks like

. . . after QC, mapping, alignment,. . .



What the data looks like

. . . after QC, mapping, alignment,. . .



How do you (statistically) tell the difference?

Differential Peak Calling

Which Peaks are significantly different between WT and Null?

Differential Peak Calling

Which Peaks are significantly different between WT and Null?

→ Current approaches mostly adopted from RNA-Seq based methods e.g. DESeq (Anders and Huber, 2010)

- DBChIP (Liang and Keles, 2012)
- DiffBind (Ross-Innes et al., 2012)

Differential Peak Calling

Which Peaks are significantly different between WT and Null?

→ Current approaches mostly adopted from RNA-Seq based methods e.g. DESeq (Anders and Huber, 2010)

- DBChIP (Liang and Keles, 2012)
- DiffBind (Ross-Innes et al., 2012)

- Assumptions: Negative Binomial distribution
- Information sharing across peaks:
  Variance is a function of the mean.
- Test statistic: Total counts within a peak.

Differential Peak Calling

Which Peaks are significantly different between WT and Null?

→ Current approaches mostly adopted from RNA-Seq based methods e.g. DESeq (Anders and Huber, 2010)

- DBChIP (Liang and Keles, 2012)
- DiffBind (Ross-Innes et al., 2012)

- Assumptions: Negative Binomial distribution
- Information sharing across peaks:
  Variance is a function of the mean.
- Test statistic: Total counts within a peak.

→ Draw-back: Peaks are represented by a single value: total counts

→ Peaks are extended over several kb.
changes in their shape are not considered

Formulate the test question

Suppose for a peak i we are given

- n observations (i.e. reads) 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

Suppose for a peak i we are given

- n observations (i.e. reads) 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'.

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

Formulate the test question

Suppose for a peak i we are given

- n observations (i.e. reads) 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'.

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

Define test statistic:

- should summarize the data, preferably in a single number
- should capture higher order moments

$\rightarrow$ use the MMD kernel method (Gretton et al 2012)

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

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

MMDiff: testing for shape changes in ChIP-Seq (Schweikert et al, BMC Gen 2013)

- MMD values are computed for each peak independently
- Every time, we compare two sets of observations:
  e.g. WT vs Null, WT vs Resc etc.
- Each read mapping to a given peak is considered an observation
- The feature we use is the 5' end of the alignment

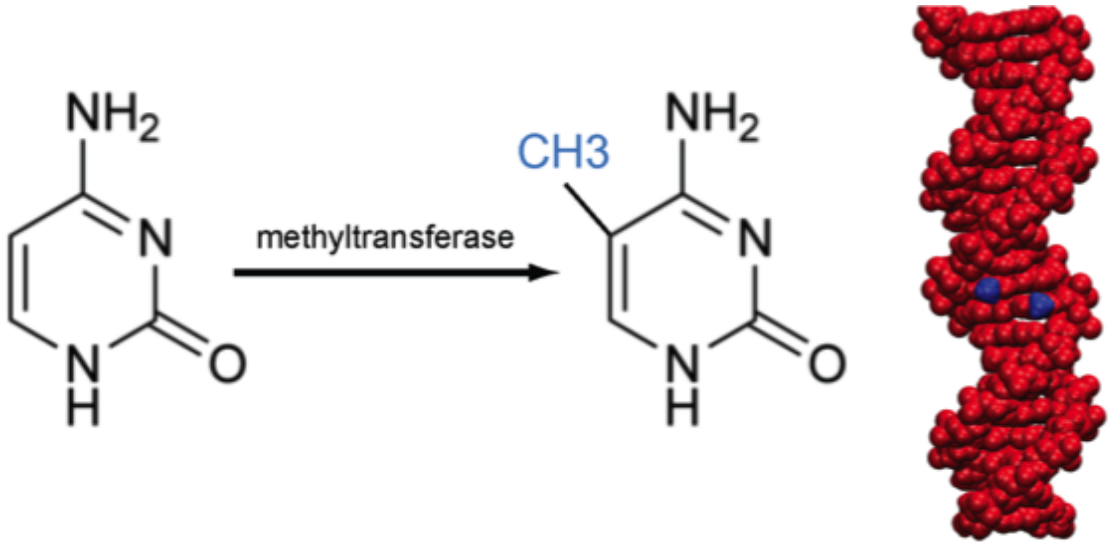
MMDiff: testing for shape changes in ChIP-Seq (Schweikert et al, BMC Gen 2013)

- MMD values are computed for each peak independently
- Every time, we compare two sets of observations:
  e.g. WT vs Null, WT vs Resc etc.
- Each read mapping to a given peak is considered an observation
- The feature we use is the 5' end of the alignment
- We use RBF Kernels to capture neighbourhood information
- The Kernel width is chosen to be the median distance between all observations

MMDiff: testing for shape changes in ChIP-Seq (Schweikert et al, BMC Gen 2013)

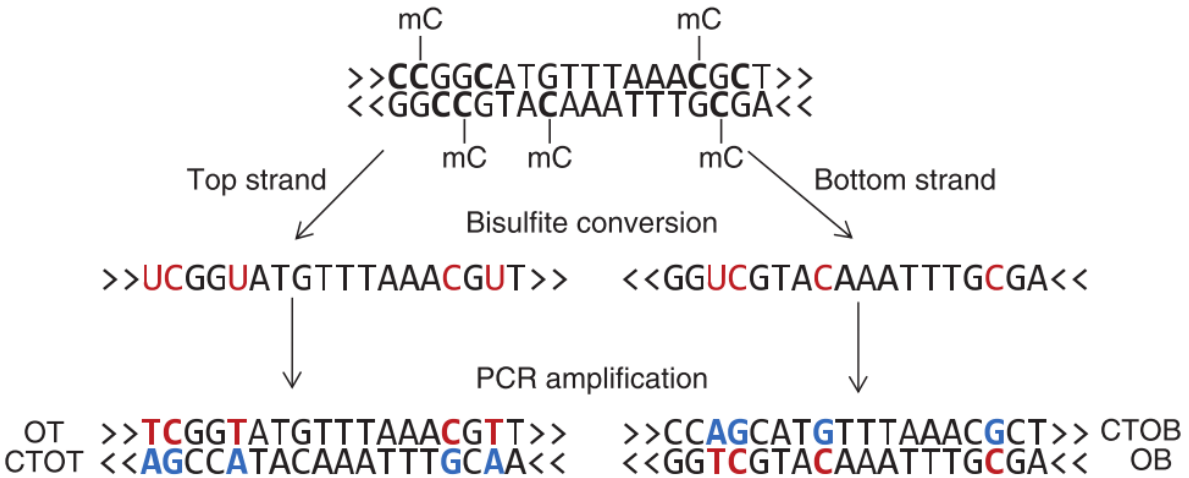
- MMD values are computed for each peak independently
- Every time, we compare two sets of observations:
  e.g. WT vs Null, WT vs Resc etc.
- Each read mapping to a given peak is considered an observation
- The feature we use is the 5' end of the alignment
- We use RBF Kernels to capture neighbourhood information
- The Kernel width is chosen to be the median distance between all observations
- Empirical p-Values are determined on peaks with similar total counts

DNA Methylation



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

Methylation Data



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

Testing for differential methylation: M^3D (Mayo et al, Bioinformatics 2015)

- DNA methylation as measured by BS-seq also presents some of the same features as ChIP-seq
- Strong spatial correlations between neighbouring CpG sites
- Insufficient replication for testing individual CpGs (if at all meaningful)

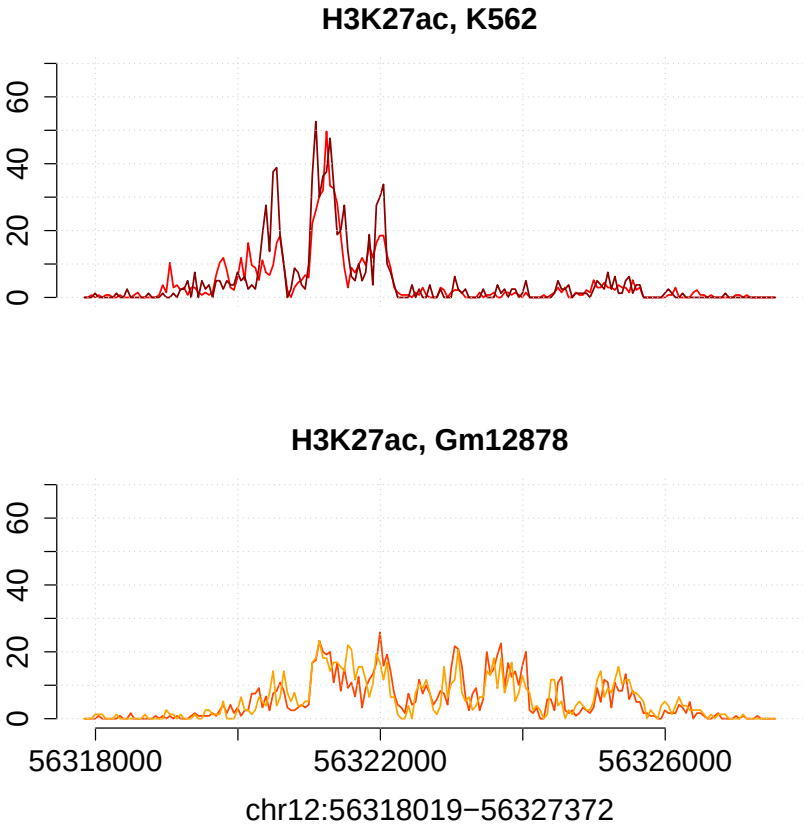
Testing for differential methylation: M^3D (Mayo et al, Bioinformatics 2015)

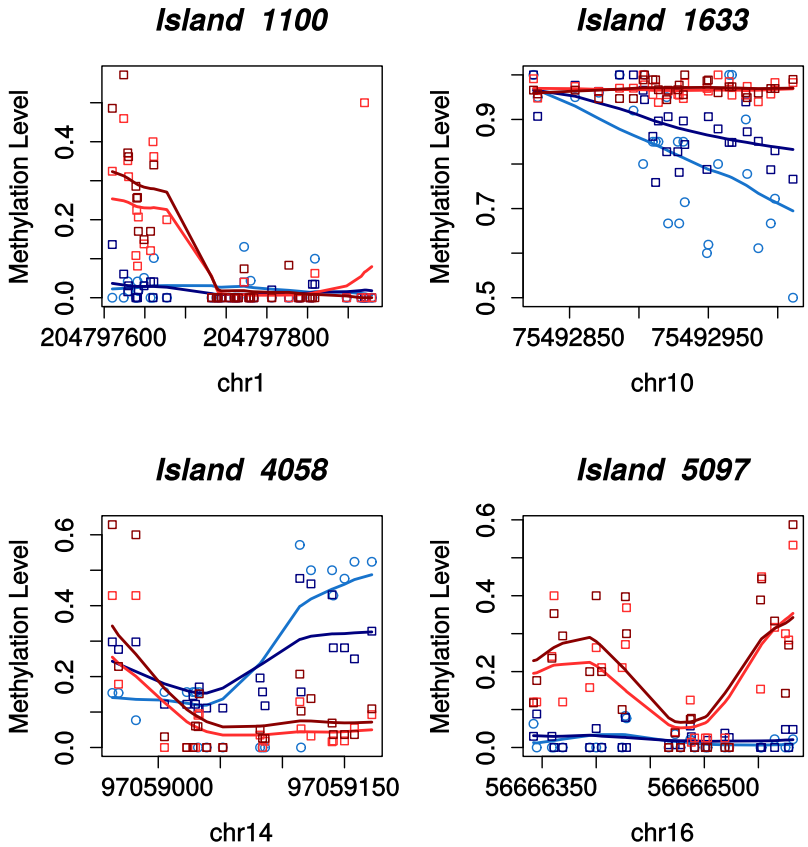
- DNA methylation as measured by BS-seq also presents some of the same features as ChIP-seq
- Strong spatial correlations between neighbouring CpG sites
- Insufficient replication for testing individual CpGs (if at all meaningful)
- We adapted the MMD metric to devise a non-parametric test for differentially methylated regions, M^3D (Maximum Mean Methylation Discrepancy)
- Technically, a bit more involved due to fractional nature of methylation measurements

Experiments on ENCODE data

- Compare ChIP-Seq/ RRBS-Seq marks across different cell types
- Studied two different marks: broad histone mark H3K27ac
- Cell types: human K562 (leukaemia) vs GM12878 for H3K27ac, mouse brain cortex, cerebellum and liver for CTCF

ENCODE results





Both called by MMDiff/ M^3D and not competing methods

Talk outline

1. Spatial effects in epigenomic data
   - Statistical testing for epigenomic data
   - Transcription factors and histone modifications
   - Clustering and prediction from epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
   - Isoform quantification from time series RNA-seq
   - Splicing quantification in single cells
3. Conclusions

Genetics strikes back

- MMDiff gives a powerful test for changes in ChIP-Seq, yet the basis of its power are not obvious: why is shape so conserved among replicates?
- Results in Cfp1 data set suggest a role for transcription factors (TFs) in regulation of histone methylation, which could explain shape conservation
- How general is TF regulation in the establishment of histone marks?

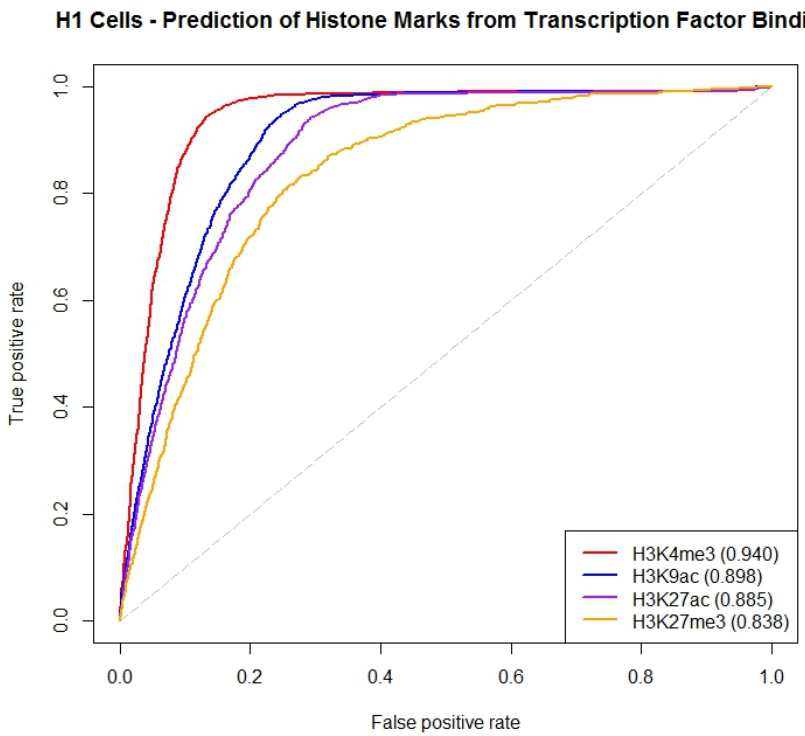
Mechanistic traces in big data?

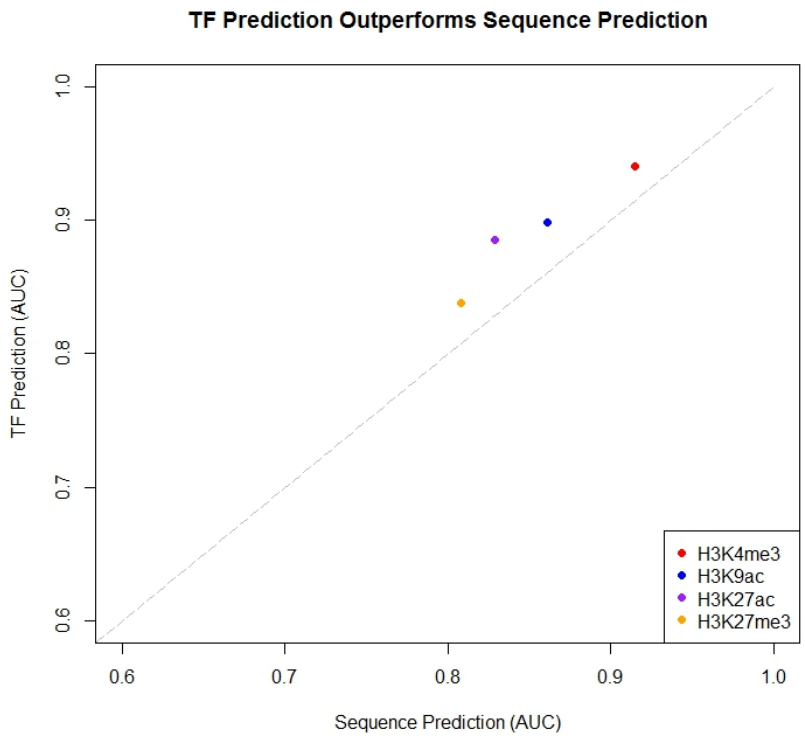
- If TF binding determined the histone mark signal, it should be possible to predict histone modifications from TF binding
- At a simpler level, it should be possible to predict the presence/absence of marks from TF ChIP-Seq data
- This does NOT provide a mechanistic proof; rather it is a necessary but not sufficient condition
- Isolated examples of interactions between TFs and histone modifiers are known

Testing the hypothesis: data

- We interrogated the ENCODE data sets in the three Tier I cell lines (GM12878, K562, H1 hESC)
- Outputs: five histone modifications, H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K27me3 found near transcription start sites. Genomic regions defined positive if they intersect with a histone peak
- Inputs: normalised read counts for ALL TF chipped in the Tier I cell lines
- Prediction method: logistic regression. Probabilistic predictor which computes relative importance of input features as a weight vector

TFs can predict very accurately





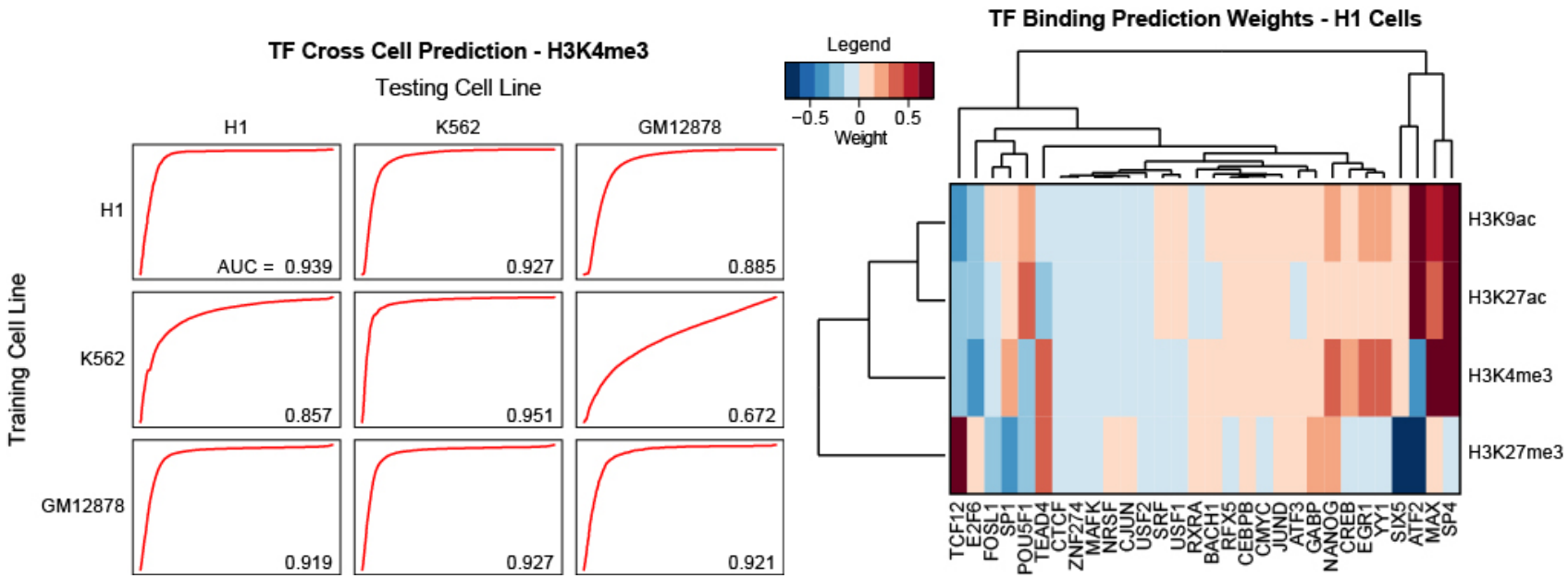
ROC curves for predictions of histone modifications at promoters in H1 cells (left). TF-based predictions vs sequence based predictions in H1 cells (right)

TFs can predict genome wide

Table: Predictions of histone modification presence in H1 cells

Mark	Seq. (promoters)	TF promoters	DNase	Enhancers F5
H3K4me1	N.D.	N.D.	0.854 $\pm$ 0.001	0.842$\pm$ 0.003
H3K4me3	0.918 ($\pm$ 0.001)	0.950 ($\pm$ 0.001)	0.974 ($\pm$ 0.001)	0.962 ($\pm$ 0.001)
H3K9ac	0.867 ($\pm$ 0.001)	0.921 ($\pm$ 0.001)	0.976 ($\pm$ 0.001)	0.961 ($\pm$ 0.001)
H3K27ac	0.828 ($\pm$ 0.002)	0.909 ($\pm$ 0.001)	0.968 ($\pm$ 0.001)	0.950($\pm$ 0.001)
H3K27me3	0.808 ($\pm$ 0.002)	0.877 ($\pm$ 0.002)	0.916 ($\pm$ 0.001)	0.918 ($\pm$ 0.002)

Cross-cell and interpretable predictions



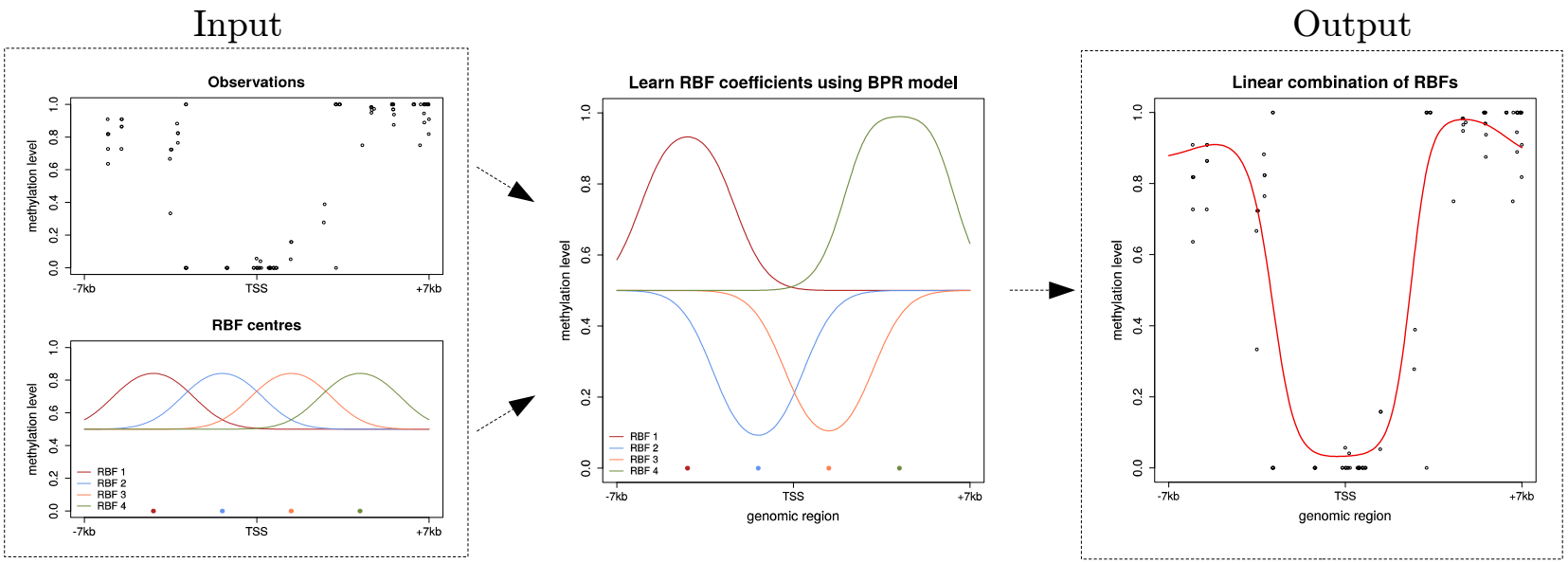
Left: TF features enables cross-cell predictions with high accuracy. Right: LR weights for TF-based histone modification predictions in H1 cells. Many known TF-histone modifier interactions are represented.

Talk outline

Why does shape matter?

- Spatial patterns of epigenomic modifications seem highly reproducible
- One possibility: the "shape" of the mark is a readout of what else is happening to the chromatin
- MMDiff differential peaks were enriched for TF binding sites
- Generally, histone modifications are systematically predictable from TF binding (more than from sequence, Benveniste et al, PNAS 2014)
- More functional implications of shape?

Quantifying methylation profiles (Kapourani and G.S, Bioinformatics 2016)



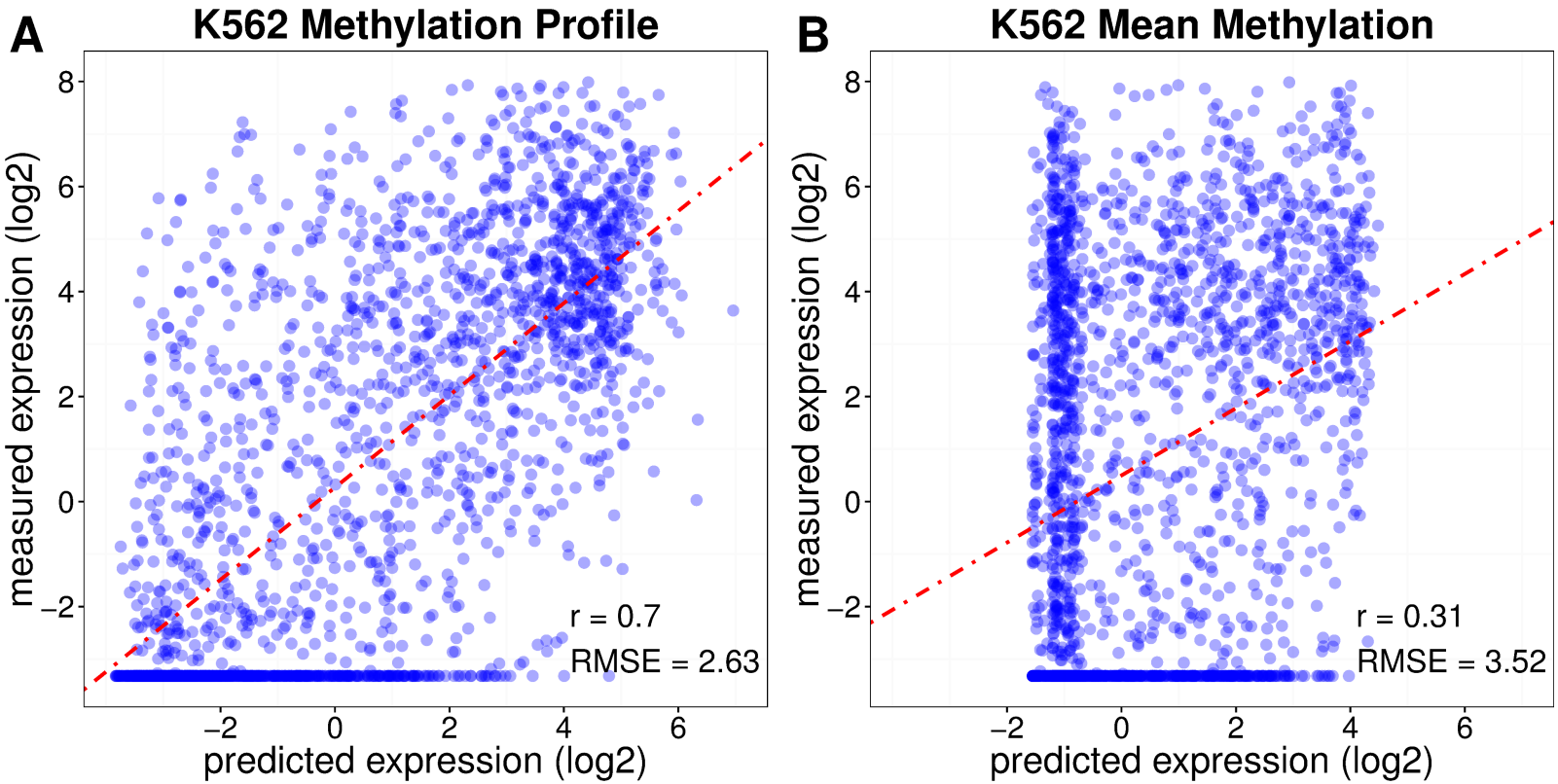
The BPRM model

- We assume the methylation pattern of a region to be determined by an unobserved methylation function $f(x) = \Phi\big(g(x)\big)$, where Φ 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 ξ_j (e.g. RBF)
- The actual number of methylated reads at position i is binomial distributed

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

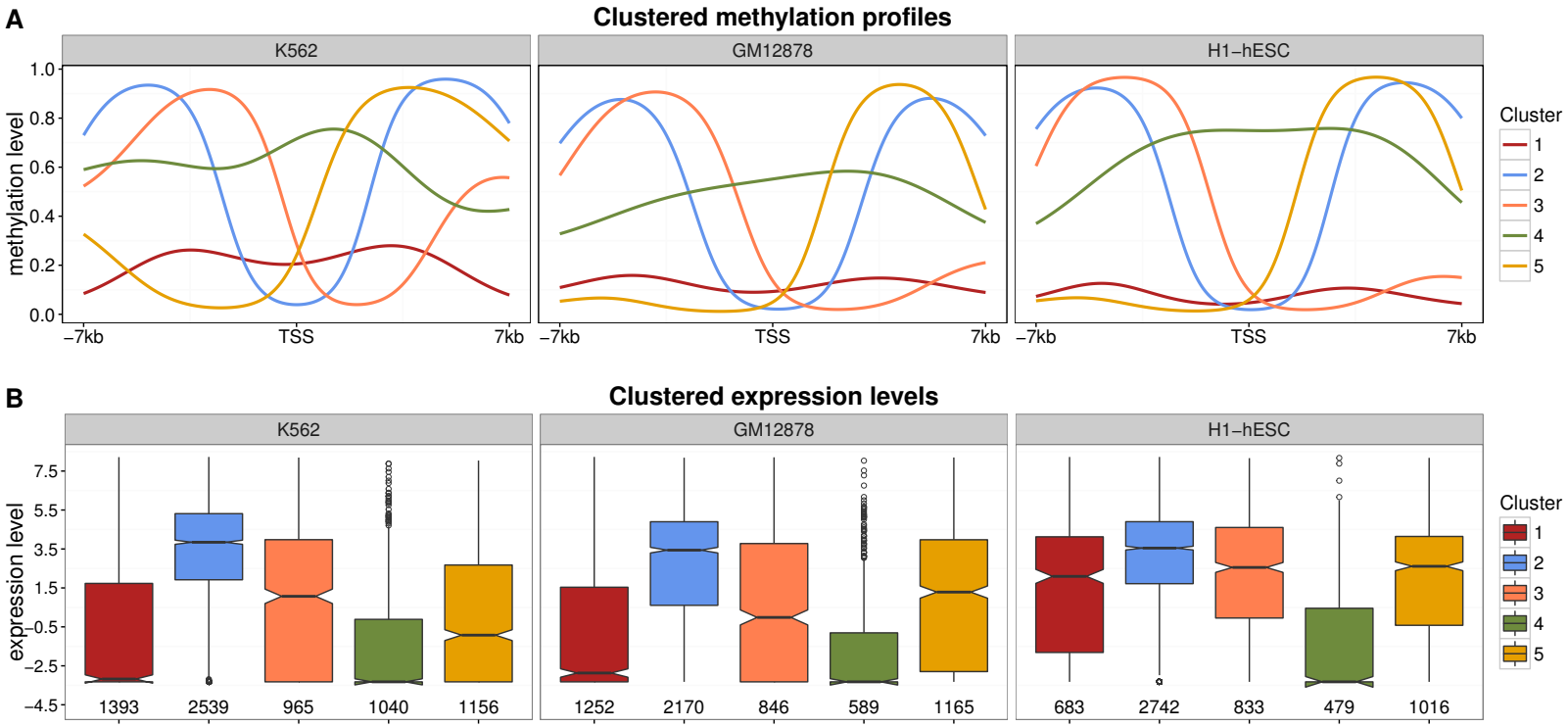
  with m_i the coverage at position i.
- Optimising the likelihood given by (1) w.r.t. the weights $\mathbf{w}$ associates each region with *methylation profile features*

Predicting gene expression



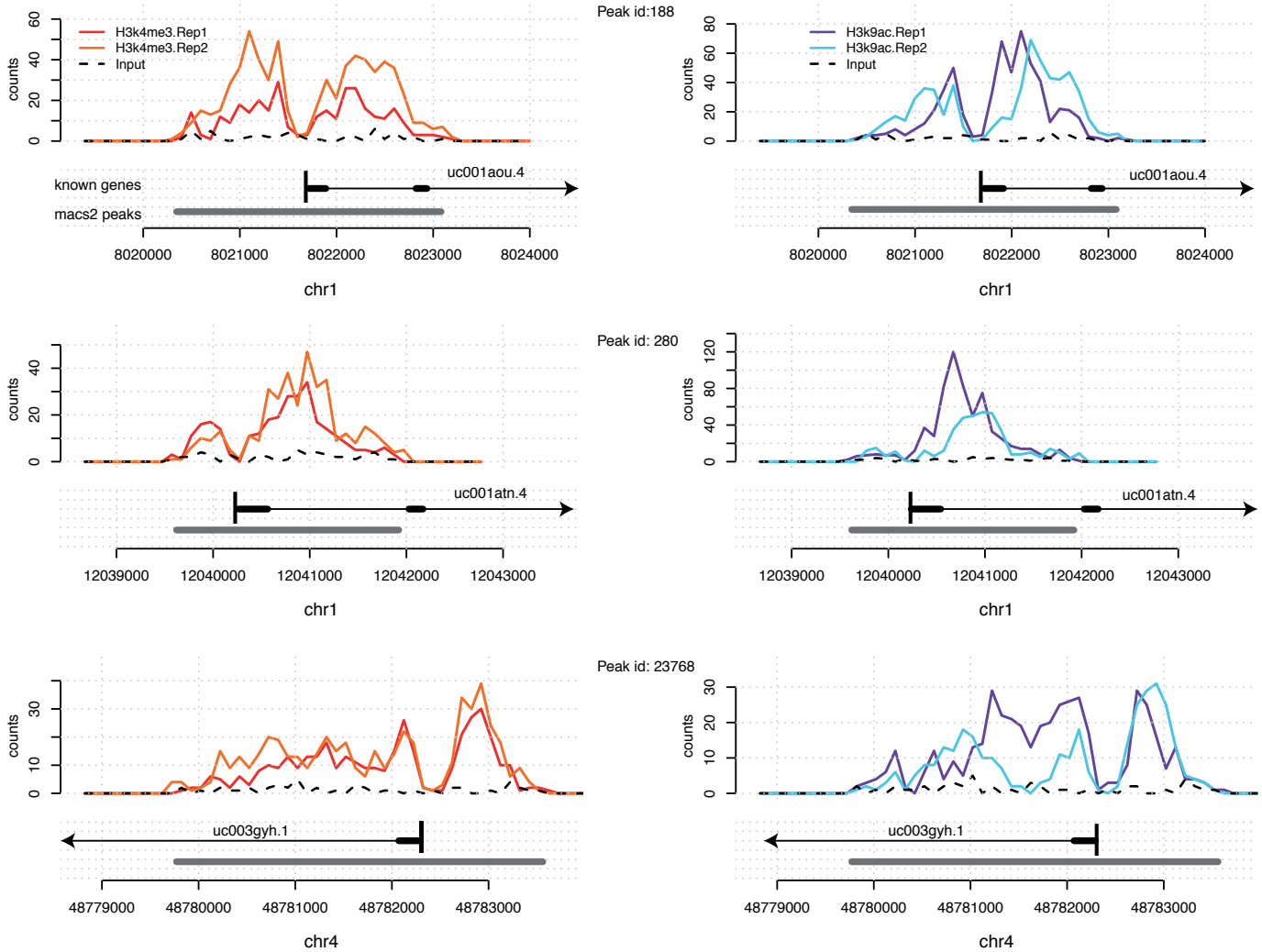
Predicting gene expression from methylation profiles (left) or mean methylation levels (right). Overall improvement in Pearson r from 0.31 to 0.72.

Clustering methylation profiles at promoters



BPRM identifies five prototypical methylation profiles across different cell lines, with distinctive expression levels

Mining similarities across epigenomic patterns?



Can we find prototypical combinations of marks at genes?

Mining similarities across epigenomic patterns?

Problems:

- Enriched Regions are detected by Peak callers such as Macs
- Each Peak has a different length
  (which might be determined by an underlying gene structure)
- Peaks are not anchored / aligned, for example to the transcription start site.
- Epigenomic marks may have some local variation, which might not be relevant for the their function

Mining similarities across epigenomic patterns?

Problems:

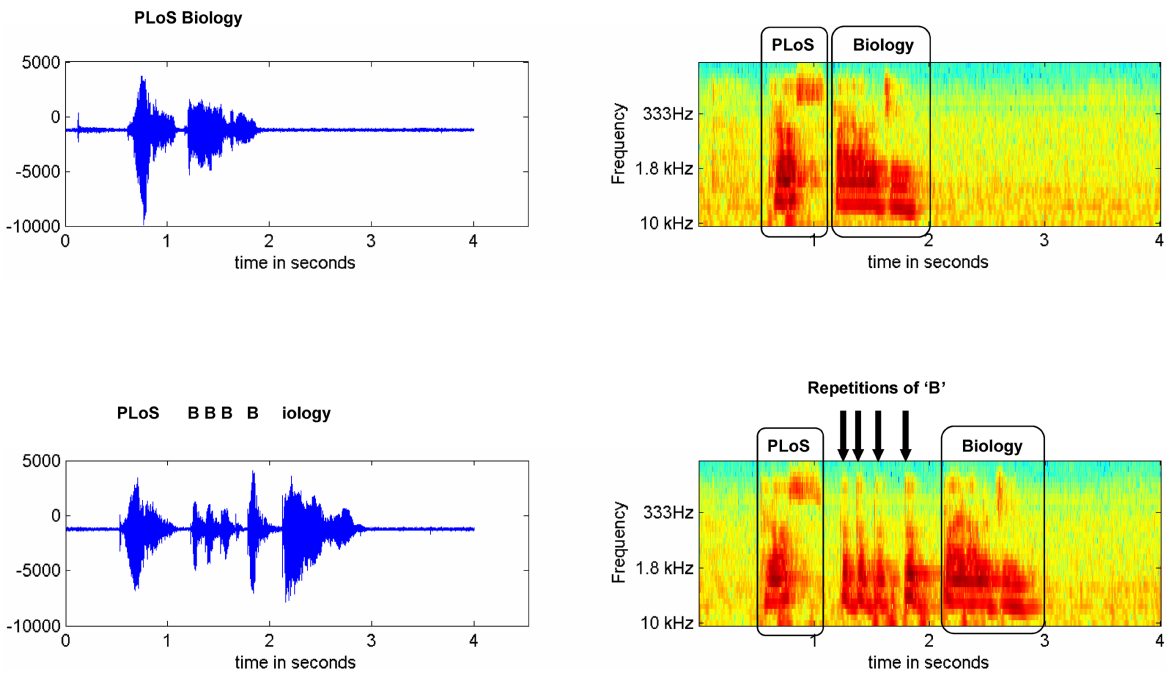
- Enriched Regions are detected by Peak callers such as Macs
- Each Peak has a different length
  (which might be determined by an underlying gene structure)
- Peaks are not anchored / aligned, for example to the transcription start site.
- Epigenomic marks may have some local variation, which might not be relevant for the their function

Our approach: Dynamic Genome Warping (DGW)

- Flexible alignment using Dynamic Time Warping (DTW) allowing *local* stretching or shrinking, subject to constraints.
- Computing pairwise distances between warped peaks
- Hierarchical clustering

Analogy: speech recognition

Situation similar to what speech engineers faced in '70s: identify robustly spectral / temporal patterns
regardless of speed of elocution



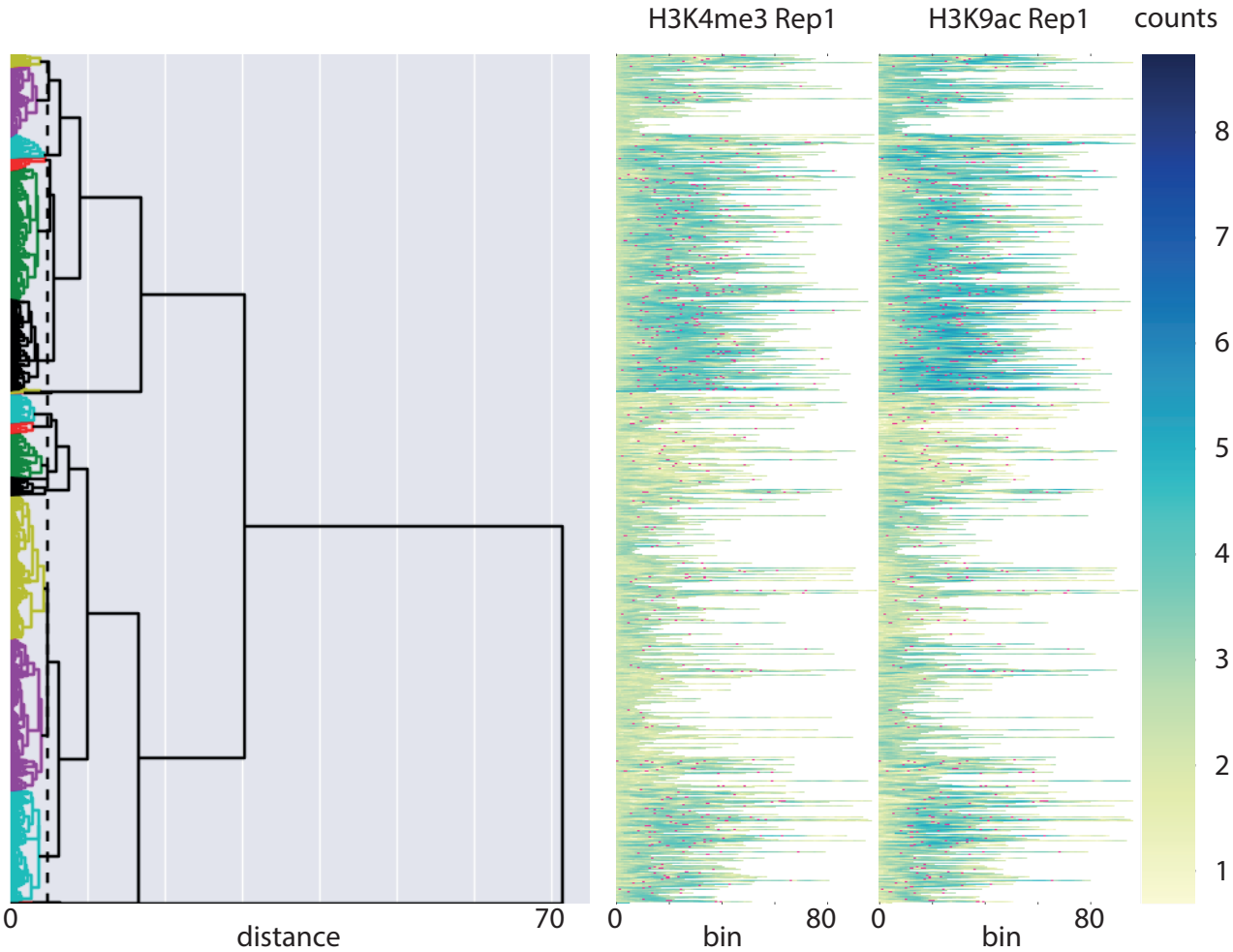
Buechel,Plos Biology, 2004

Solution: adaptive rescaling of time: Dynamic Time warping, DTW)

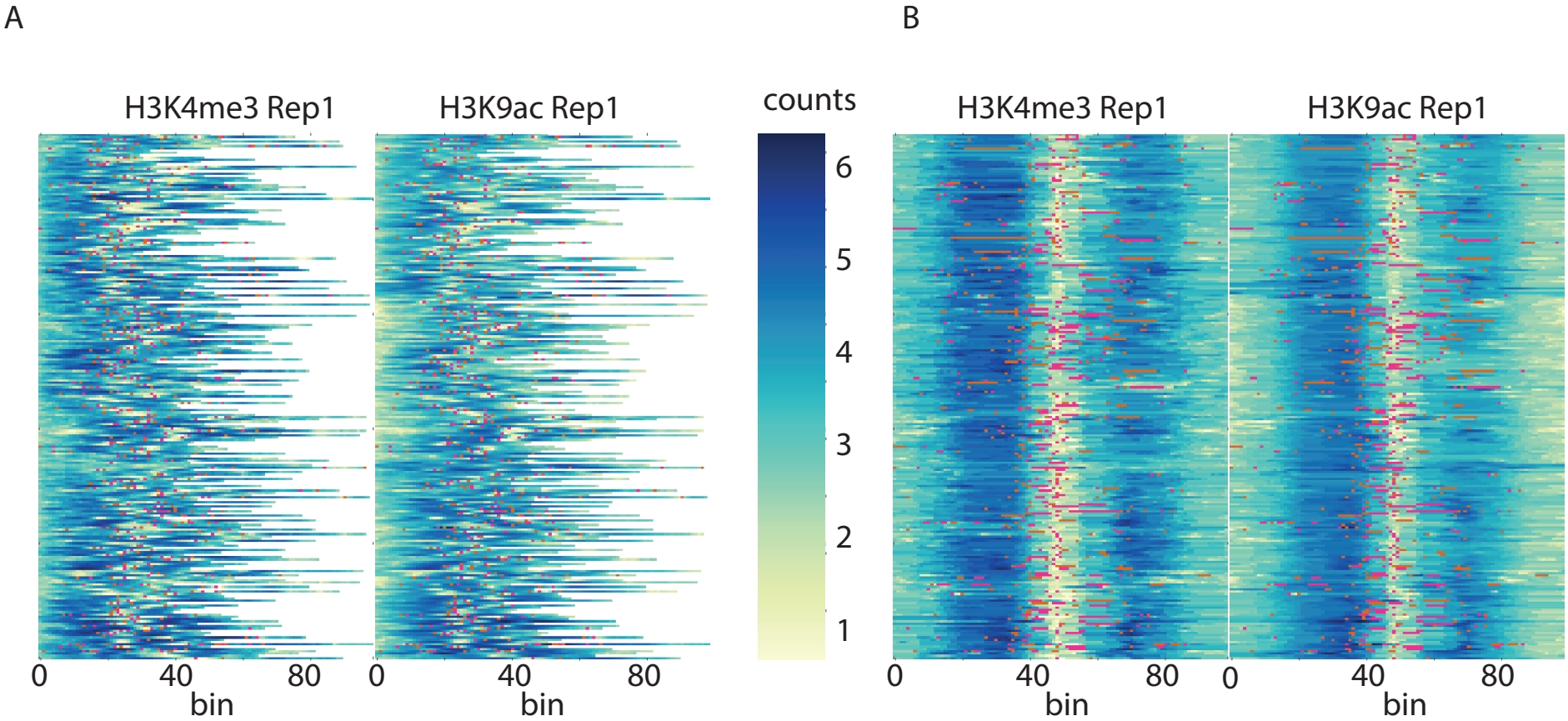
Example: H3K4me3 and H3K9ac

- Epigenomic marks H3K4me3 and H3K9ac
- Data from ENCODE K562 leukaemia cell line (accession code wgEncodeBroadHistoneK562)
- These marks have been shown to accumulate at transcription start sites and splicing sites

Results: Encode Data

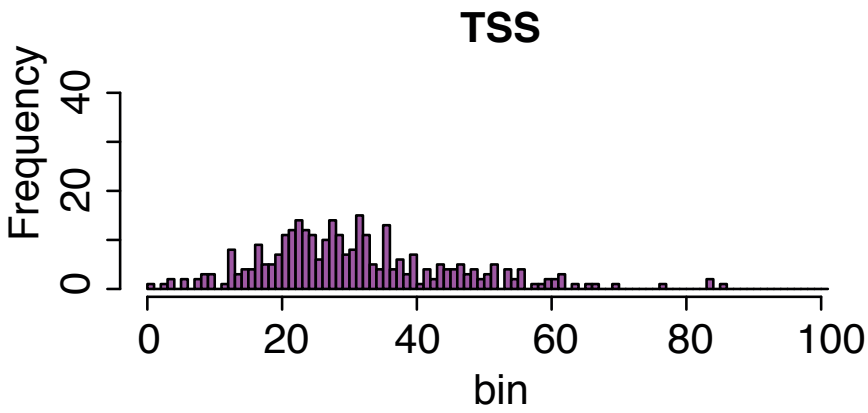


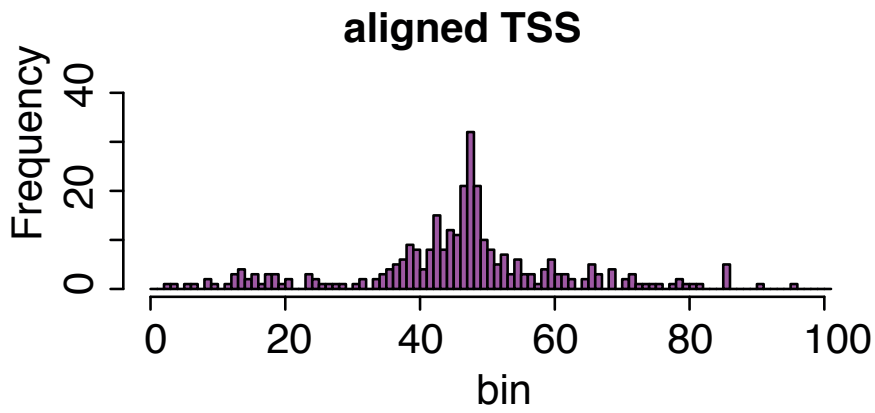
Results: Cluster 7



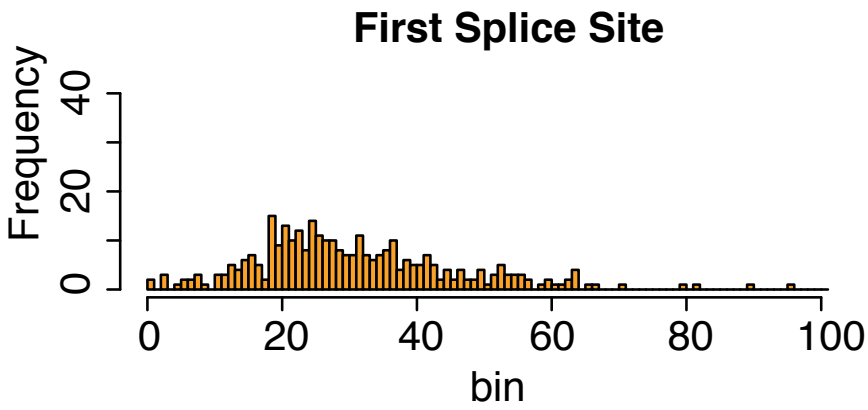
Results: DGW aligns genomic landmarks

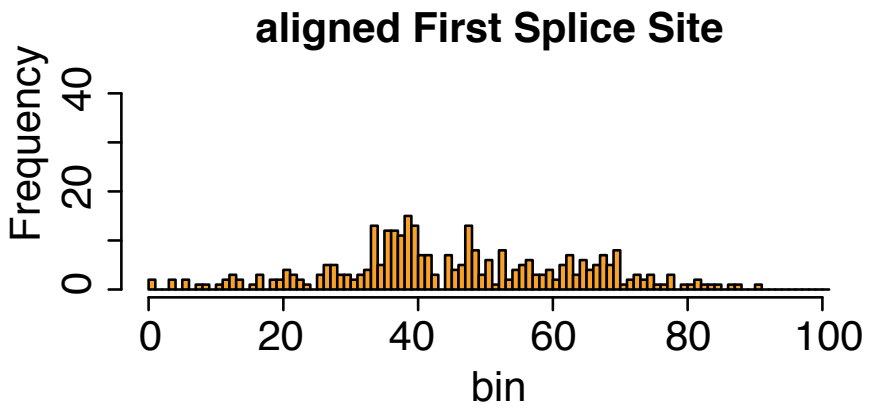
C





D





Talk outline

1. Spatial effects in epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
3. Conclusions

RNA splicing and alternative isoform

- RNA splicing: exon, intron
- Alternative isoforms:
  1) exon1-exon3;
  2) exon1-exon2-exon3
- Common in Eukaryotes:
  $\sim$ 20,000 human genes $\Rightarrow$
  $\sim$ 200,000 proteins
- Important in biological processes, etc

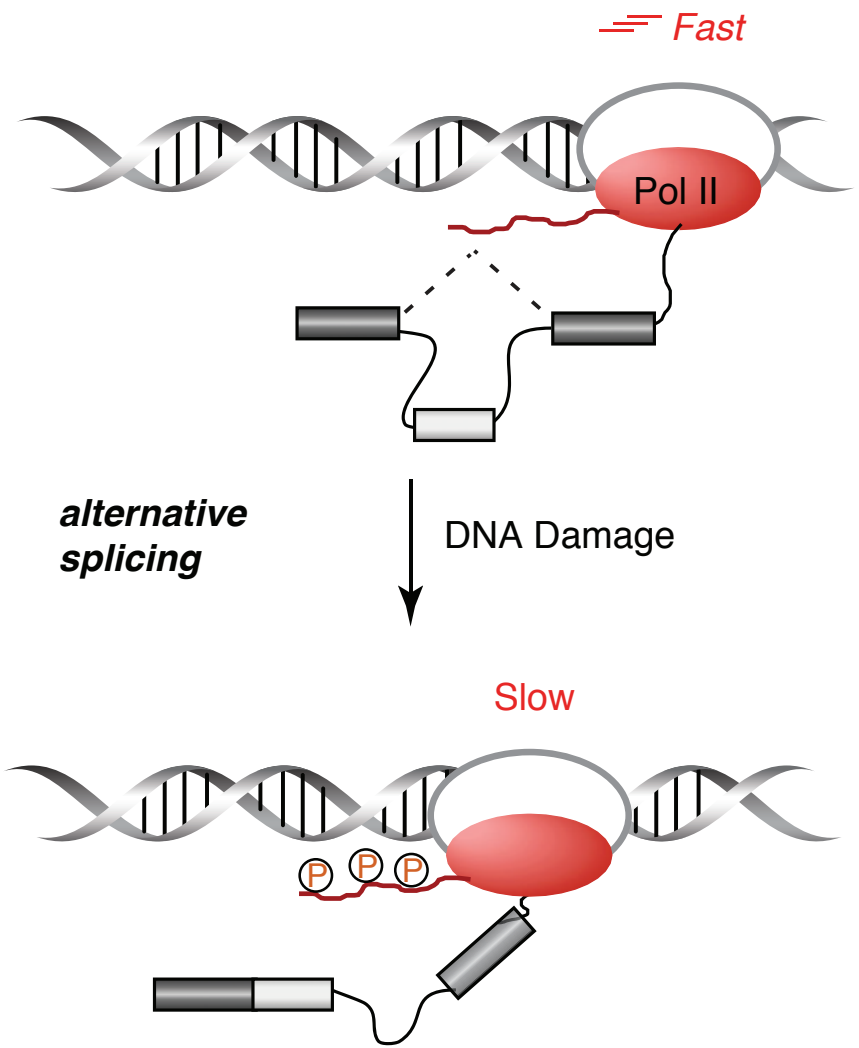


Figure modified from Heyd & Lynch, Tren. Bioche. Sci., 2011.

RNA-seq for isoform quantification

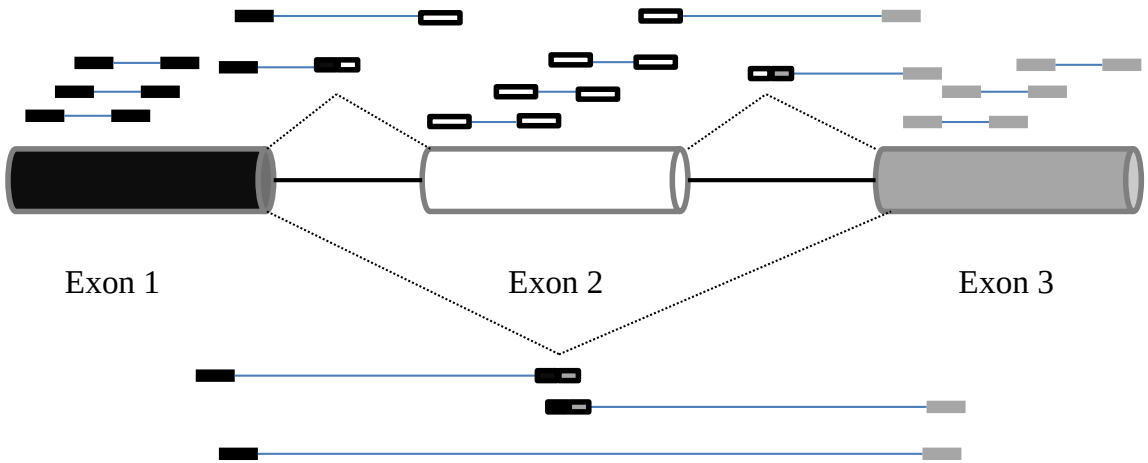


Figure: A gene with two isoforms from three exons, and aligned RNA-seq reads.

Challenges in isoform quantification

- Isoforms have shared information
- RNA-seq reads are short: 30$\sim$150bp

$\Rightarrow$ ambiguous reads identity $I_n = 1$ or $I_n = 2$

Isoforms quantification (Counting or Inferring)

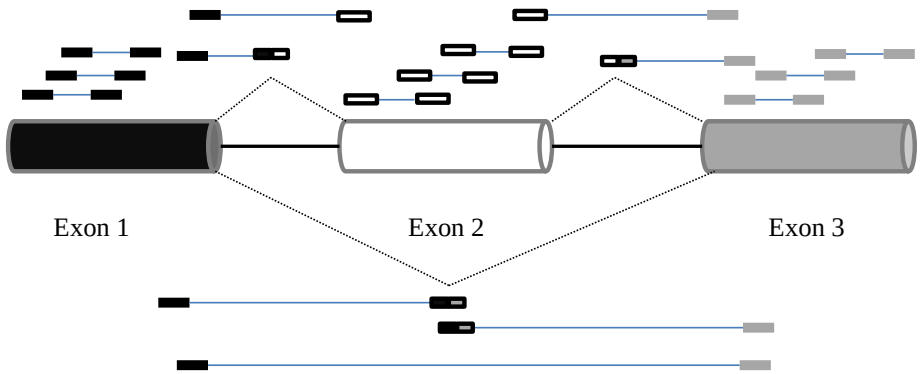


Figure: A gene with two isoforms from three exons, and aligned RNA-seq reads.

Measure isoforms ratios Ψ from reads set $R_{1:N}$

- Direct method: junction reads, counting
  $\psi_1 : \psi_2 =$ exon1_exon2 : exon1_exon3 $= 2:3$
- Probabilistic method: all reads, approximating
  $\mathcal{L}(\Psi) = \prod_{n=1}^{N} P(R_n|\Psi) = \prod_{n=1}^{N} \sum_{I_n=1}^{2} \{P(R_n|I_n)P(I_n|\Psi)\}$

  $I_n|\Psi$: multinomial distribution.
  $R_n|I_n$: denoted by the reads position; pre-computed

Isoforms quantification (Counting or Inferring)

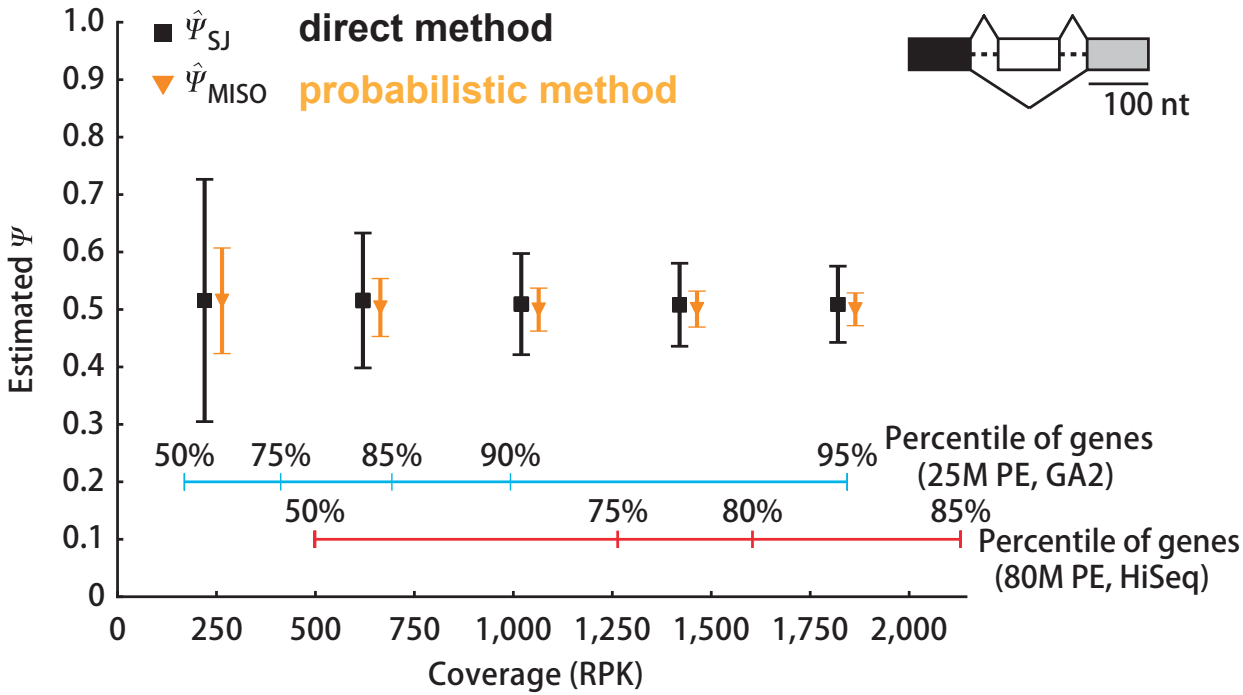


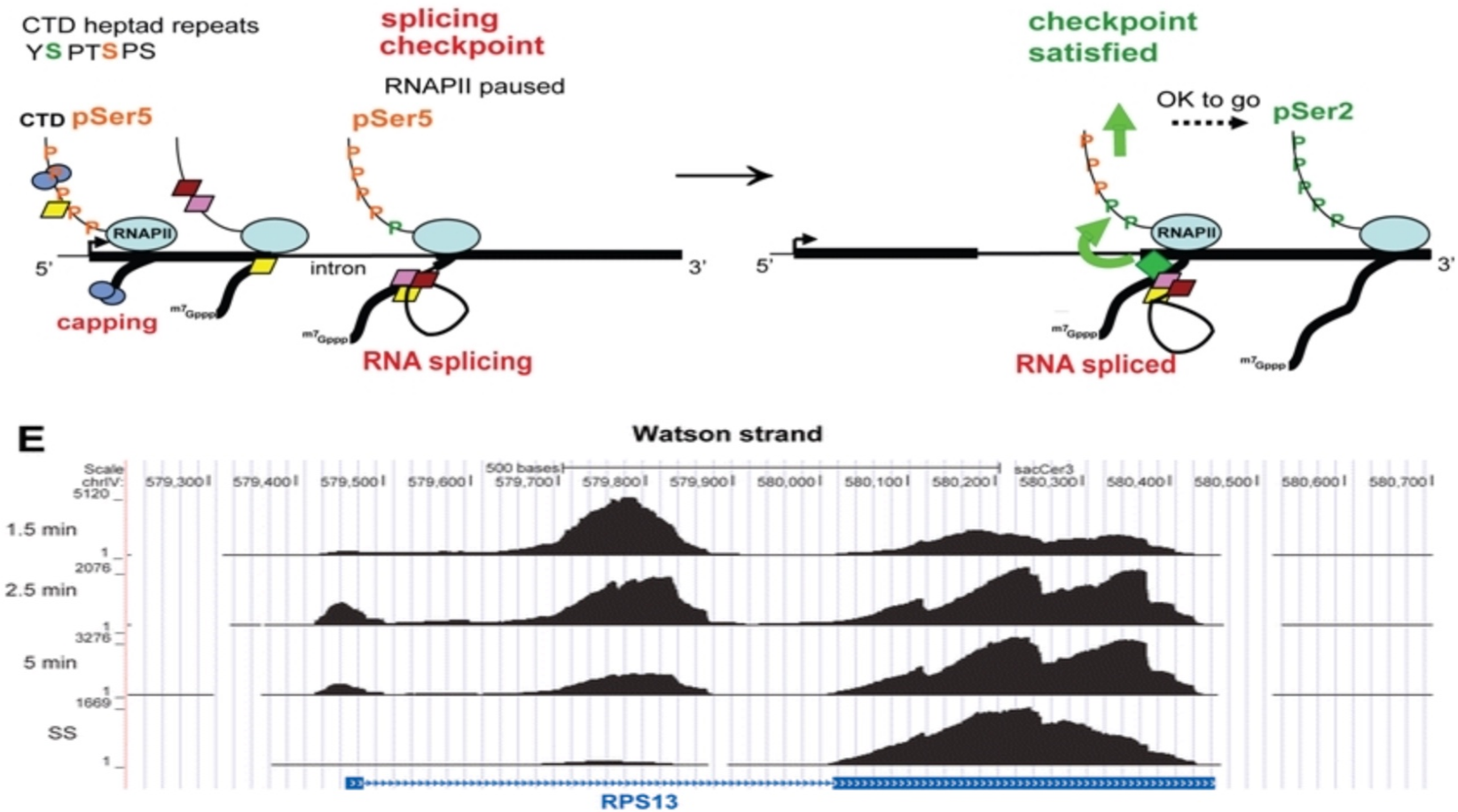
Figure: Benefits of probabilistic methods (Katz *et al* Nature Methods, 2010)

How about very low expression? Use side information.

Talk outline

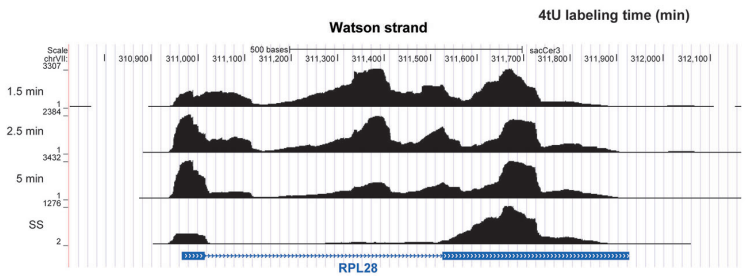
1. Spatial effects in epigenomic data
   - Statistical testing for epigenomic data
   - Transcription factors and histone modifications
   - Clustering and prediction from epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
   - Isoform quantification from time series RNA-seq
   - Splicing quantification in single cells
3. Conclusions

The biological question

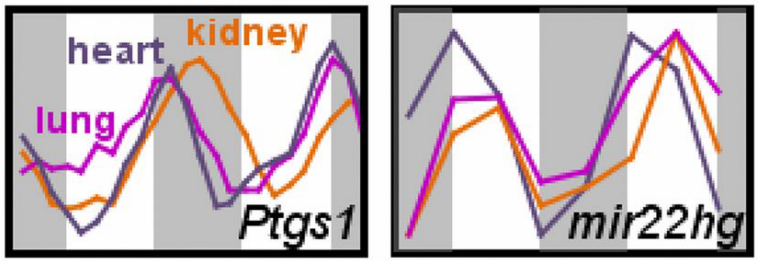


Studying splicing kinetics through time-series of labelled RNA-seq

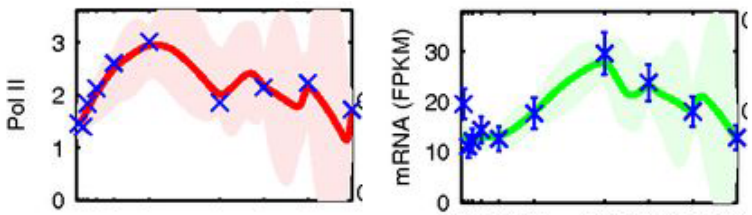
Increasing time series RNA-seq experiments



- Barrass *et al* Genome Bio. 2015.
- 4tU-labelling RNA-seq for splicing kinetics
- Estimate pre-mRNA and mRNA



- Zhang *et al* PNAS. 2014.
- Circadian RNA-seq expression
- Estimate mRNA ratios and dynamics



- Honkela *et al* PNAS. 2015.
- Time series RNA-seq and ChIP-seq
- Estimate pre-mRNA and many mRNA

Many more time series RNA-seq experiments...

Time-series isoform quantification: separate or joint?

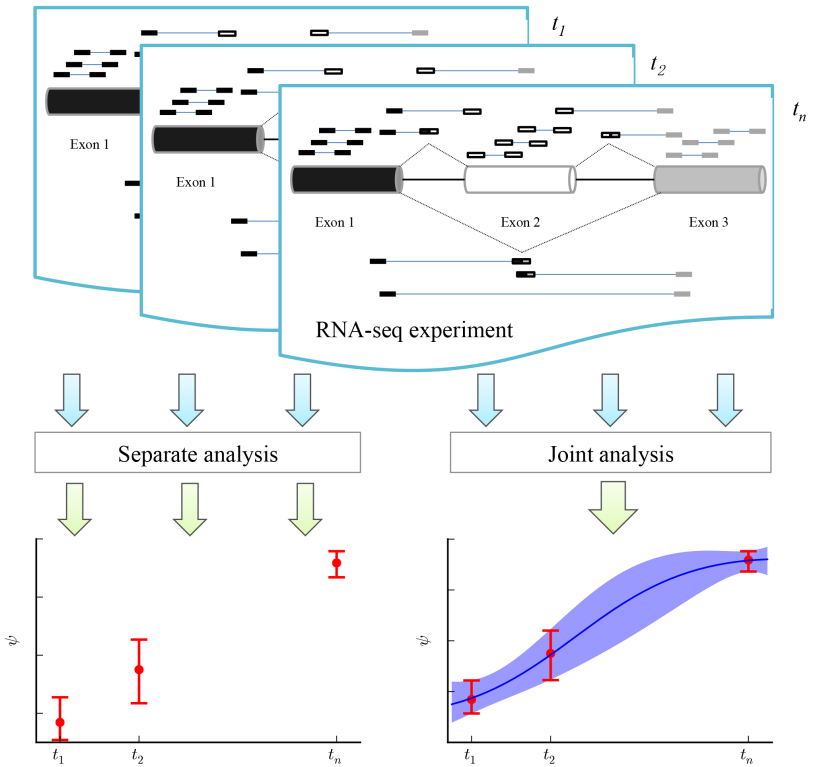


Figure: **The temporal correlation may be useful.**

Bayesian method: prior for temporal correlation

Likelihood

$$\mathcal{L}(\Psi) = \prod_{t=1}^{T} P(R_{1:N_t}^{(t)}|\Psi^{(t)}) = \prod_{t=1}^{T}\prod_{n=1}^{N_t}\sum_{I_n^{(t)}=1}^{C} P(R_n^{(t)}|I_n^{(t)})P(I_n^{(t)}|\Psi^{(t)}) \quad (2)$$

- T: number of time point; Ψ: isoform ratio; R: reads; I: identity.
- $I_n|\Psi$: multinomial distribution.
- $R_n|I_n$: pre-computed, with modelling biases, etc.

Posterior

$$P(\Psi|\Theta, \boldsymbol{R}) \propto P(\Theta)P(\Psi|\Theta) \times \prod_{t=1}^{T} P(R_{1:N_t}^{(t)}|\Psi^{(t)}) \quad (3)$$

Posterior **Joint Prior** **Likelihood**

GP prior: modelling the temporal correlation

Joint Prior: Gaussian process (GP)

$$P(\Psi|\Theta) = P(\boldsymbol{Y}|\boldsymbol{M}, \boldsymbol{K}) = \prod_{c=1}^{C-1} \mathcal{N}(Y_c|\boldsymbol{m}_c, K_c) \quad (4)$$

Assumptions

- latent variable $Y = \texttt{Softmax}^{-1}(\Psi)$, i.e., $\psi_c = e^{y_c} / \sum_{i=1}^{C} e^{y_i}$, and $y_c = 0$.
- latent vector $Y_c = [y_c^{(1)}, ..., y_c^{(T)}]$ follows a GP: $Y_c|T, \boldsymbol{\theta}_c \sim \mathcal{N}(\boldsymbol{m}_c, K_c), c \in [1, ..., C-1]$ where $\boldsymbol{m}_c$: vector of mean, K: covariance matrix
- K is defined by the hyperparameters $\boldsymbol{\theta}_c = (\theta_{c,1}, \theta_{c,2})$: $\text{Cov}(y_c^{(t_1)}, y_c^{(t_2)}) = \theta_{c,1} \exp(-\frac{1}{2\theta_{c,2}}(t_1 - t_2)^2)$

Note: $\theta_{c,2}$ **governing the strength of temporal correlation.**

Inference: Metropolis-Hastings sampler

Algorithm 1 Metropolis-Hastings sampler for posterior of latent $\boldsymbol{Y}$

Require: $T, \boldsymbol{R}, \boldsymbol{\Theta}, \lambda$
Initialize: $\boldsymbol{Y}^{(0)}$
Calculate: $\boldsymbol{\Psi}^{(0)} = \texttt{Softmax}(\boldsymbol{Y}^{(0)})$; $\boldsymbol{K} = \texttt{GPcov}(\Theta, T)$
for $i = 0$ to H **do**
  Sample: $\mu \sim U(0, 1)$
  Sample: $\boldsymbol{Y}^* \sim Q_y(\boldsymbol{Y}^* | \boldsymbol{Y}^{(i)}, \lambda \boldsymbol{K})$
  Calculate: $\boldsymbol{\Psi}^* = \texttt{Softmax}(\boldsymbol{Y}^*)$
  if $\mu < \texttt{min}\left\{ \dfrac{P(\boldsymbol{\Psi}^* | \boldsymbol{R}) \times Q_y(\boldsymbol{Y}^{(i)} | \boldsymbol{Y}^*, \lambda \boldsymbol{K})}{P(\boldsymbol{\Psi}^{(i)} | \boldsymbol{R}) \times Q_y(\boldsymbol{Y}^* | \boldsymbol{Y}^{(i)}, \lambda \boldsymbol{K})}, 1 \right\}$ **then**
    $\boldsymbol{Y}^{(i+1)} \leftarrow \boldsymbol{Y}^*$; $\boldsymbol{\Psi}^{(i+1)} \leftarrow \boldsymbol{\Psi}^*$
  else
    $\boldsymbol{Y}^{(i+1)} \leftarrow \boldsymbol{Y}^{(i)}$; $\boldsymbol{\Psi}^{(i+1)} \leftarrow \boldsymbol{\Psi}^{(i)}$
  end if
end for

- fixed $\theta_1 = 3.0$ and θ_2 covers 1/3 duration (user setting).
- The proposal distribution Q_y: $\mathcal{N}(Y_c^{(i)}, \lambda K_c)$ with $\lambda = (5\sigma_y^2)/(CT\theta_{c,1})$ to ensure 30-50% acceptance ratio.

Methods performance in simulation (human)

Mean absolute error **95% confidence interval**

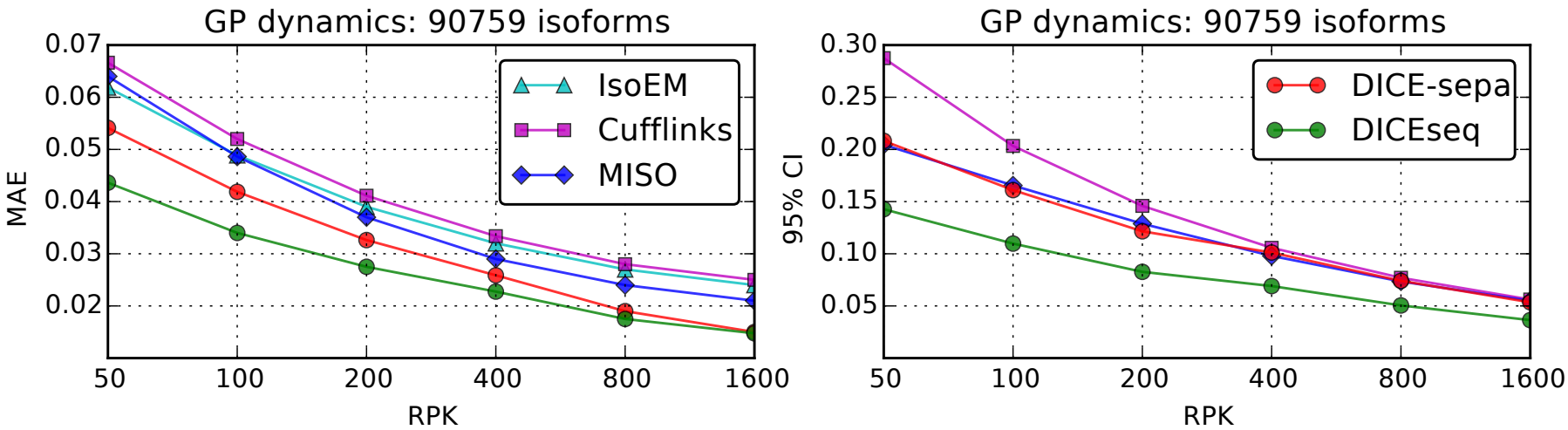


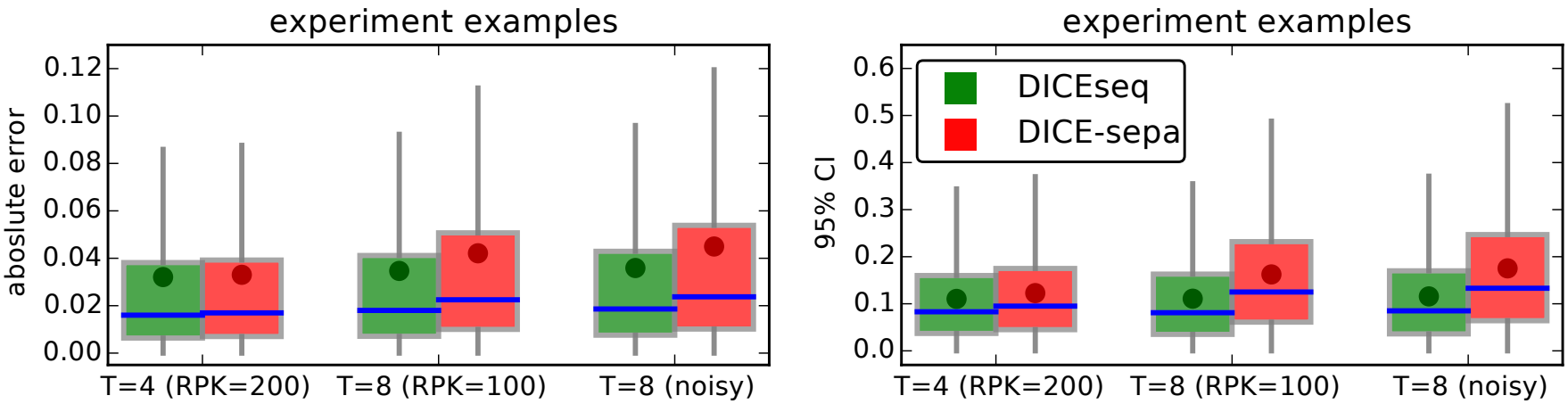
Figure legends shared in two panels

- Simulation on 11,462 human genes with different RPK
- RPK: reads per kilo base-pair

- DICEseq has clear advantages at lower coverage
- More confident in estimate by using temporal information

Time series experiment design (human)

Q: how many time points to use and how deep of the sequencing?



Two cases with similar total reads counts (similar costs)

- 4 time points with RPK=200 (left)
- 8 time points with RPK=100 (middle)

What if a time point with higher noise (right)

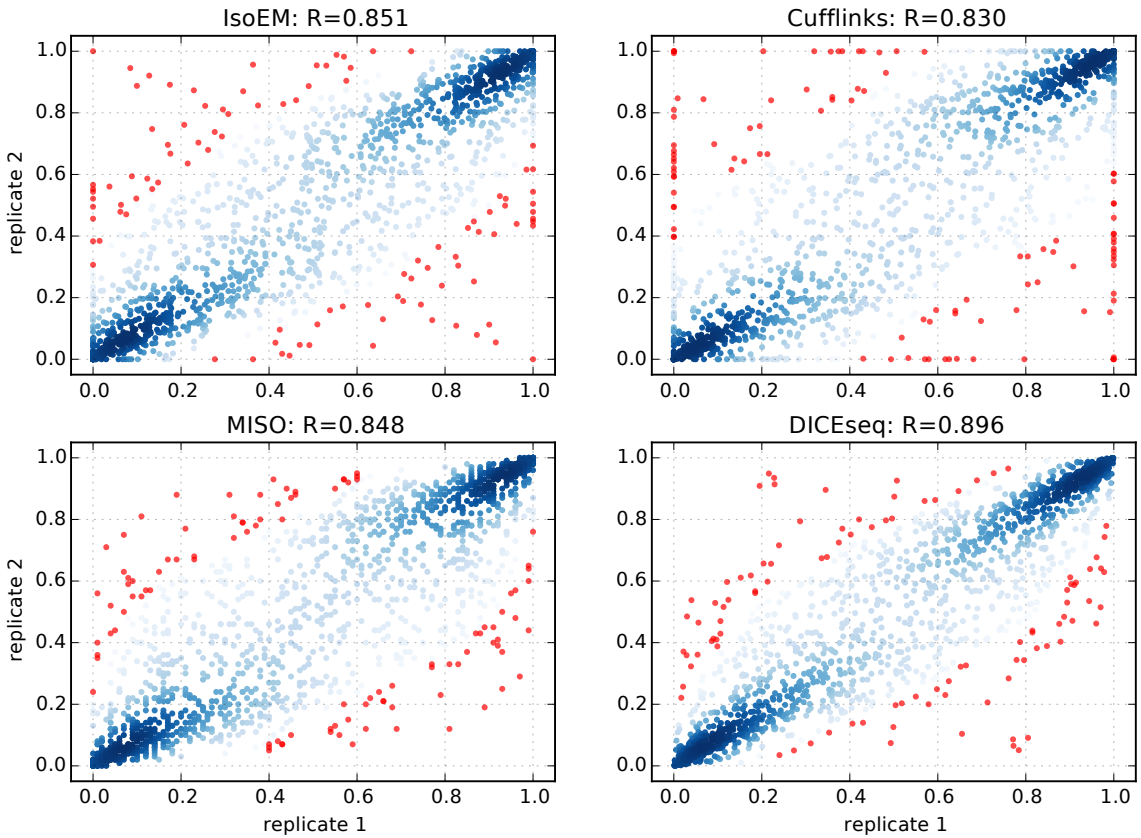
Application 1: 4tU-seq for RNA splicing kinetics (yeast)

Experiments

- labelling: newly transcribed mRNA
- Isoforms: pre- and mature mRNA
- Time: 1.5, 2.5, 5 min

No 95%CI

N(95%CI<0.3)=174



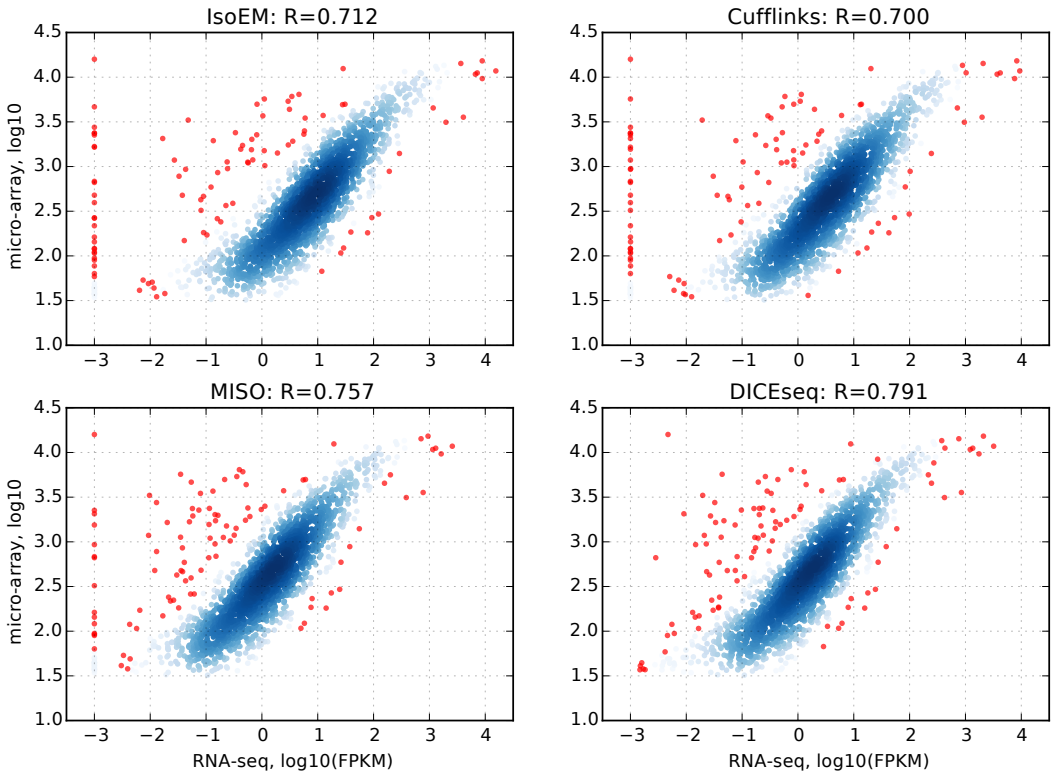
N(95%CI<0.3)=169

N(95%CI<0.3)=213

Application 2: Circadian gene expression (mouse)

Experiments

- Circadian gene expression
- Isoforms: $\geqslant$2, each gene
- Time: 6 hours x 8 points



Application 1&2: Lowly expressed genes

- Pearson's correlation: two replicates (4tU-seq) and two techniques (circadian)
- 1/3 bottom (4tU-seq) and 1/3 middle (circadian) genes in expression

	IsoEM	Cufflinks	MISO	DICEseq
4tU-seq, all	0.851	0.830	0.848	0.896
4tU-seq, 1/3 low	0.775	0.657	0.757	0.860
circadian, all	0.712	0.700	0.757	0.791
circadian, 1/3 mid	0.336	0.296	0.408	0.513

Improvement is more significant for lowly expressed genes.

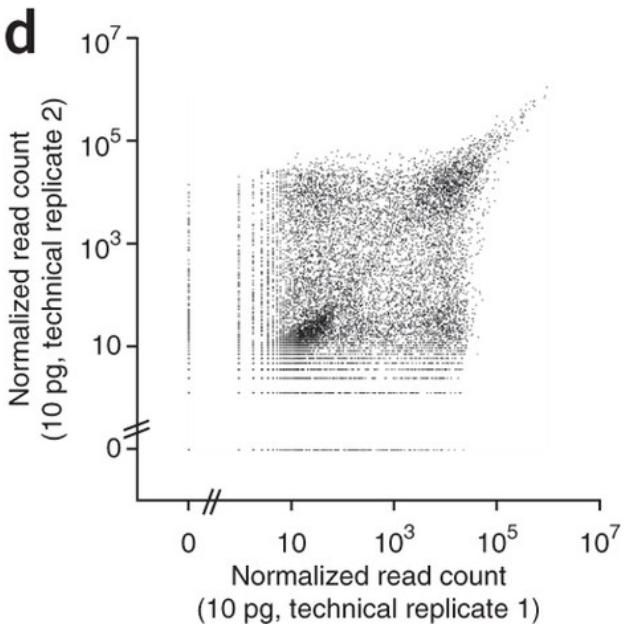
Talk outline

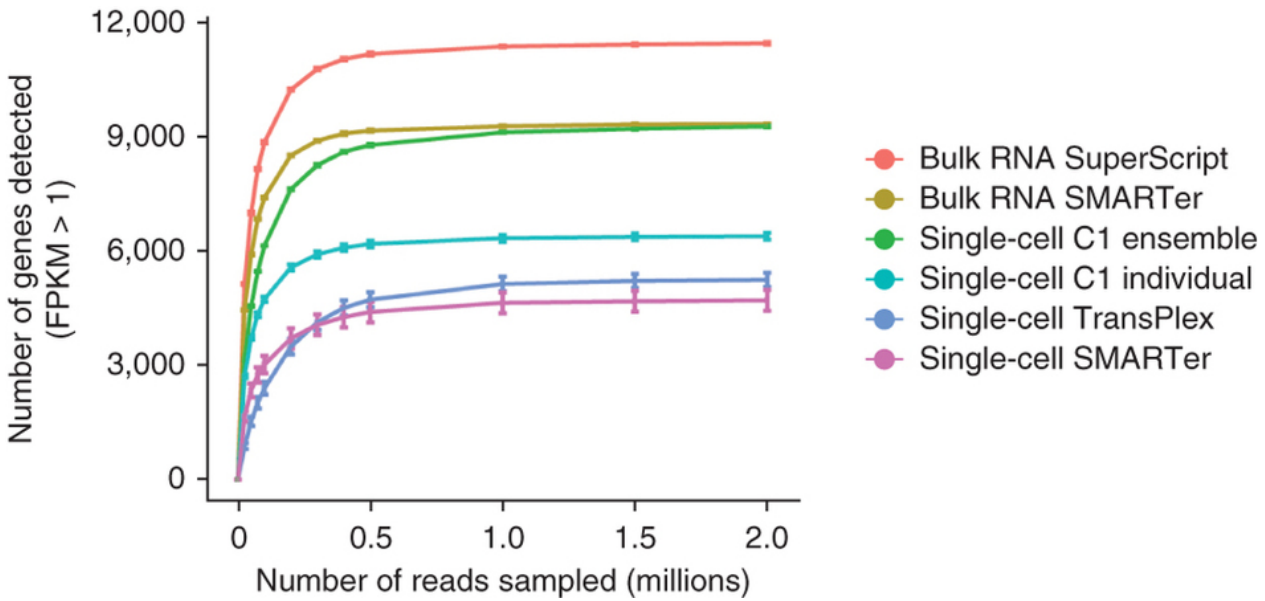
1. Spatial effects in epigenomic data
   - Statistical testing for epigenomic data
   - Transcription factors and histone modifications
   - Clustering and prediction from epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
   - Isoform quantification from time series RNA-seq
   - Splicing quantification in single cells
3. Conclusions

RNA-seq in single-cell experiment: technical noise

Very limited original RNA

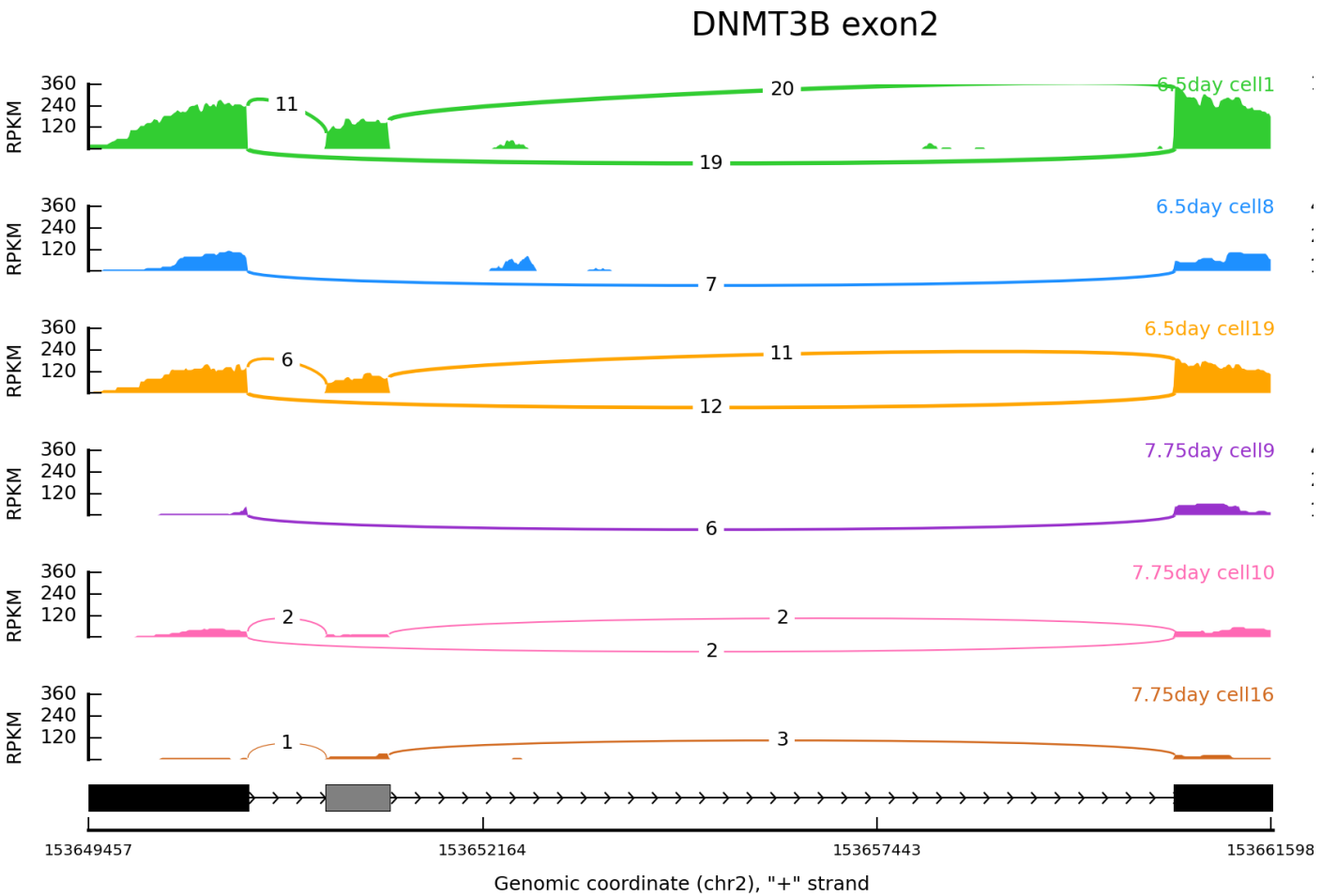
- large technical noise (low correlation between technical replicates)
- around 40% expressed genes detected (high drop-out)





Left: Philip Brennecke et al., Nat Met. 2013. Right: Angela R Wu et al., Nat Met. 2014.

Splicing in single cell: example



What to do with missing data?

- A glance at the data shows that, at least for some genes, there may be evidence for alternative splicing in single cells
- Large fraction of missing data means we cannot say anything for a majority of genes

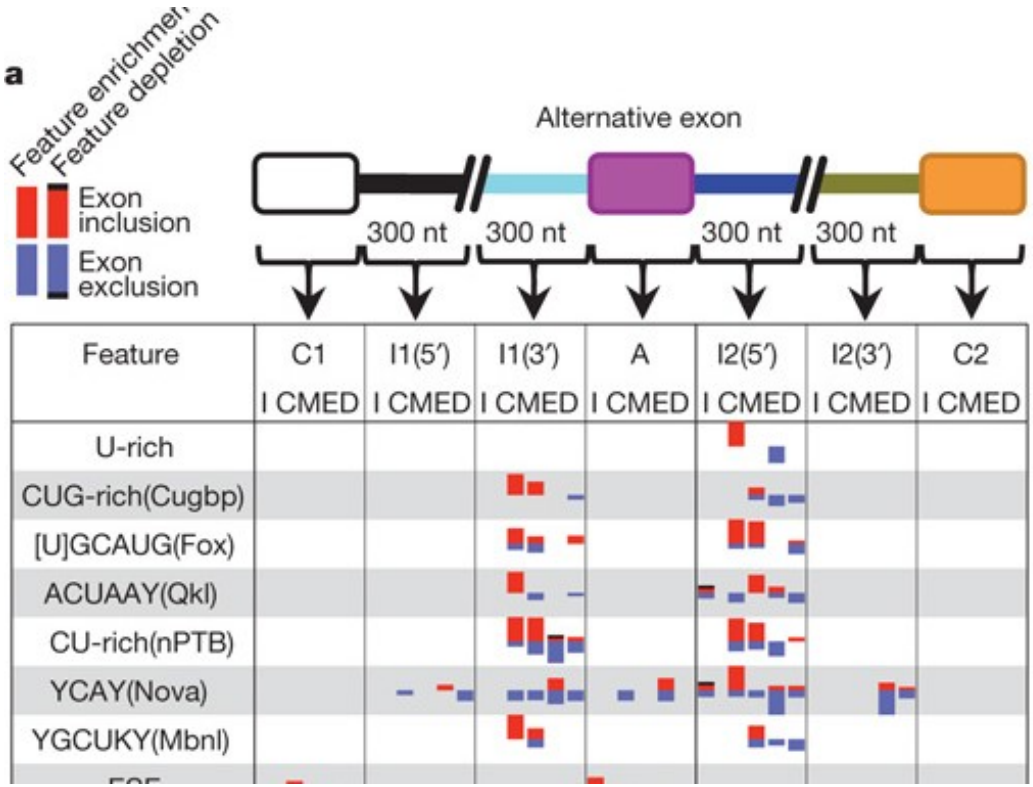
What to do with missing data?

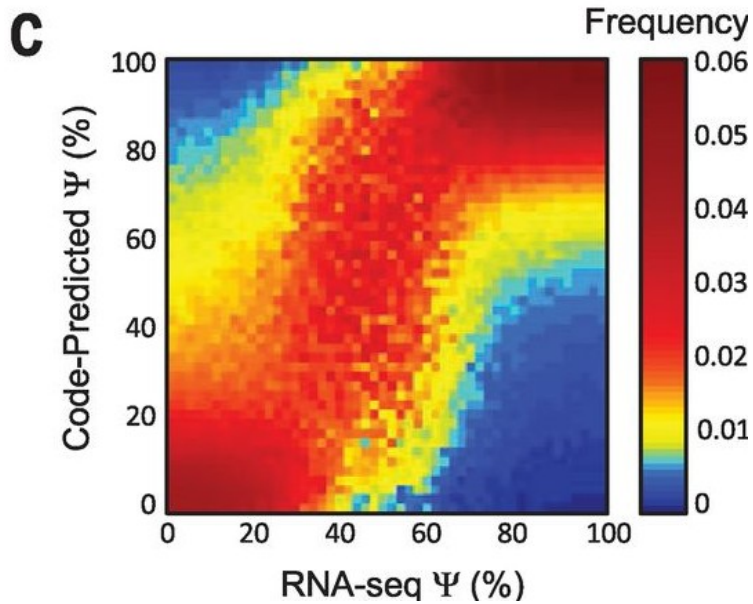
- A glance at the data shows that, at least for some genes, there may be evidence for alternative splicing in single cells
- Large fraction of missing data means we cannot say anything for a majority of genes
- **Idea:** learn an informative prior distribution from data!
- When missing data, impute. When plenty, let data speak. In between, do Bayesian inference!

Genetic regulatory code for splicing

Genetic sequence motifs can predict skipping event well

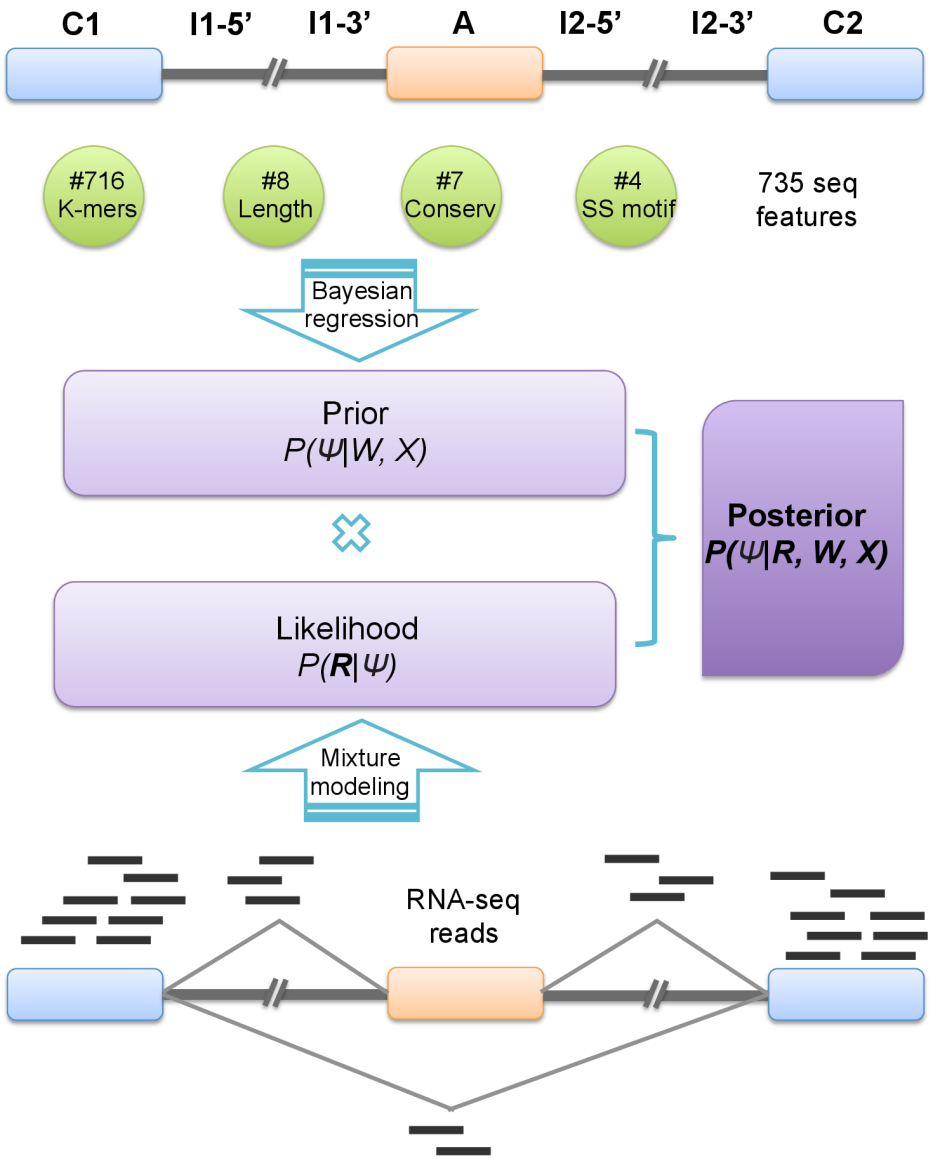
Exon triplets: 7 bins and over 1300 sequence features



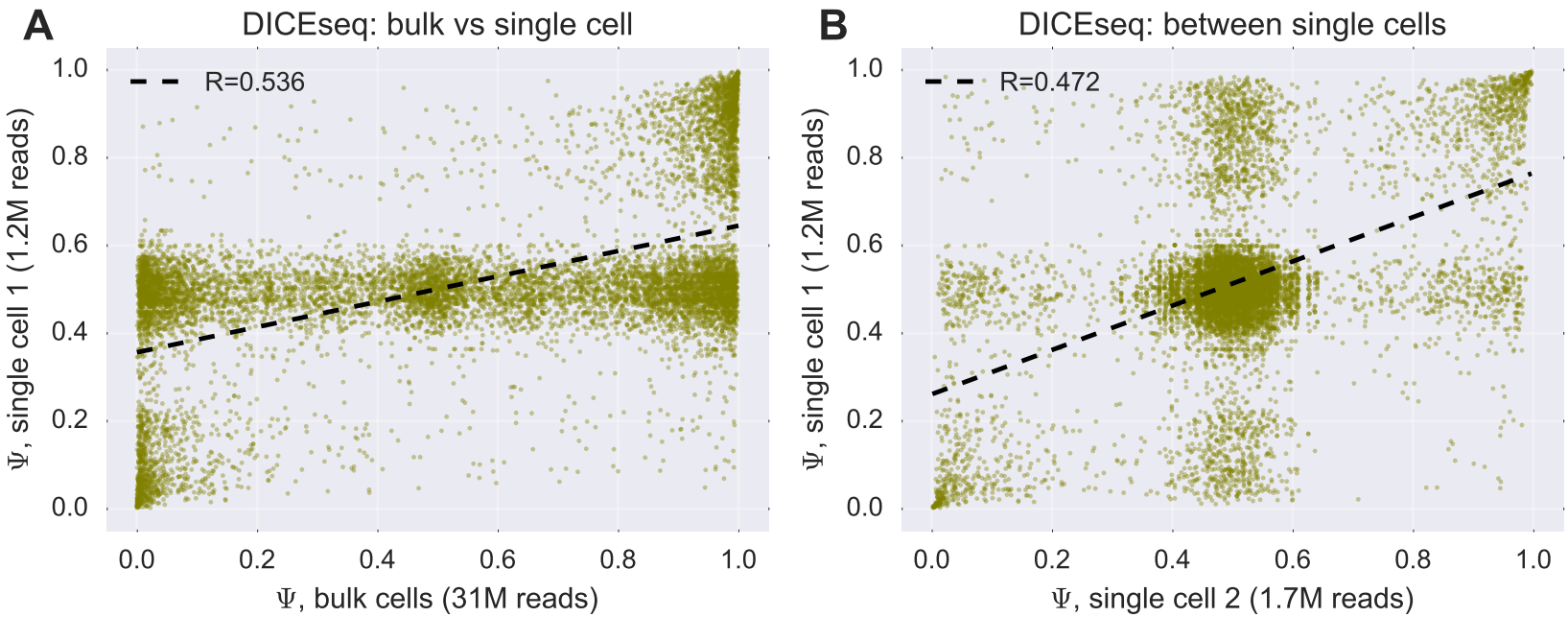


Left: Yoseph Barash et al., Nature. 2010. Right: Hui Y. Xiong et al., Science. 2014.

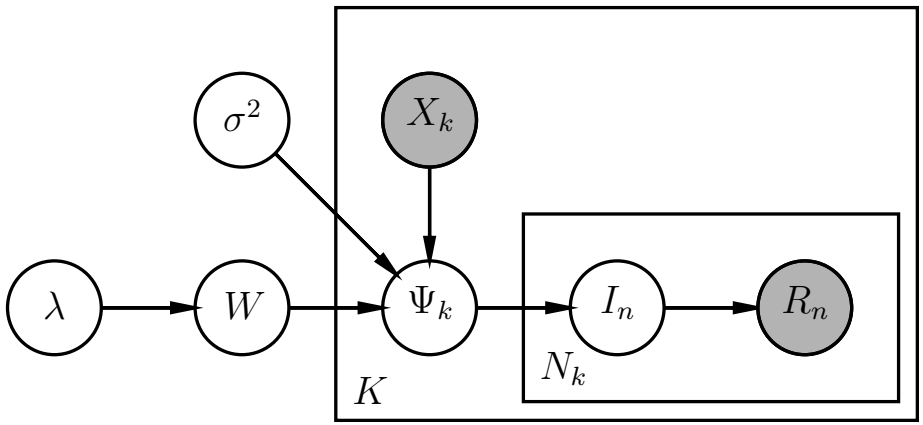
BRIE: Bayesian regression for isoform estimate



Benchmark single cell RNA-seq data (Wu et al, no prior info)



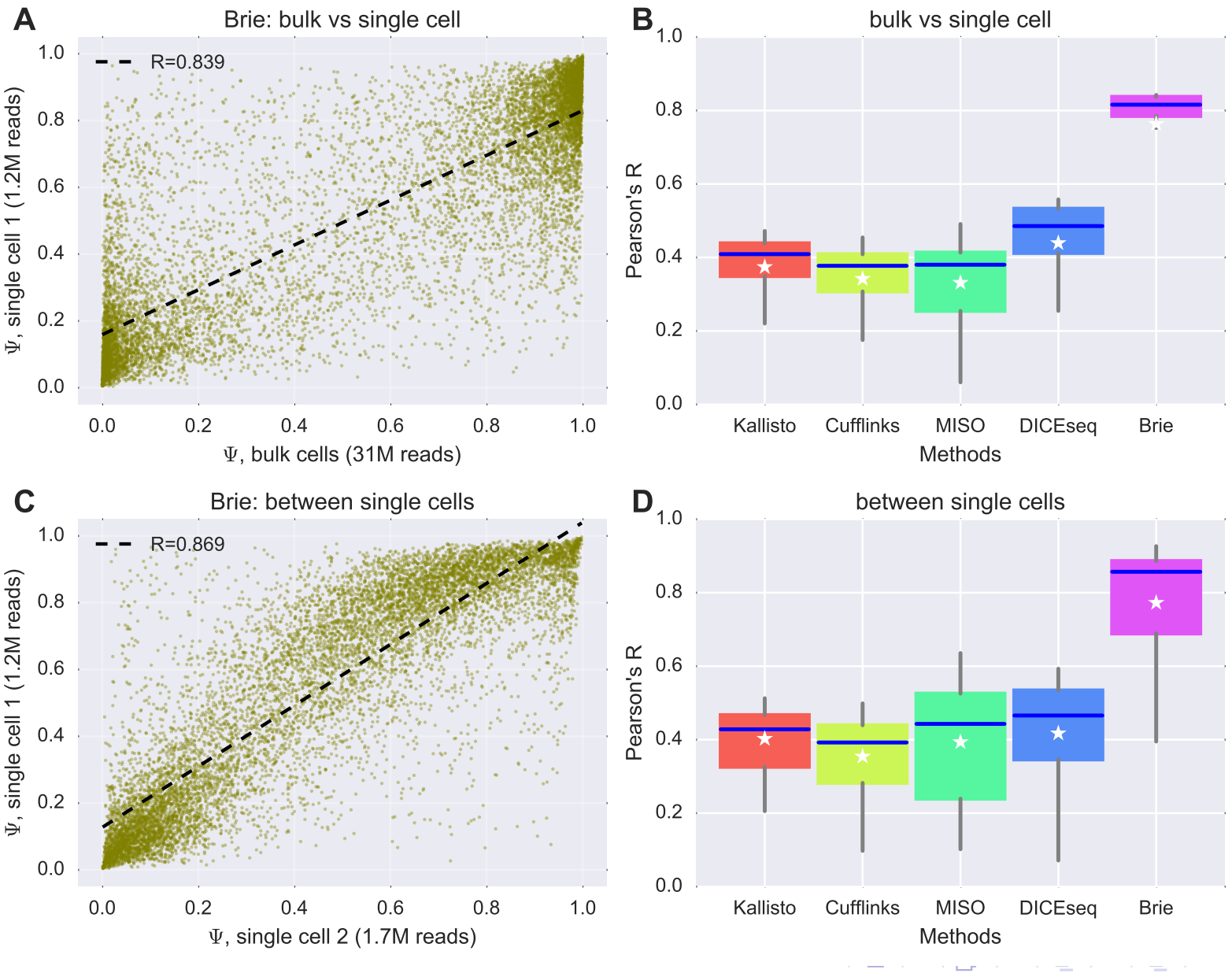
BRIE graphical presentation



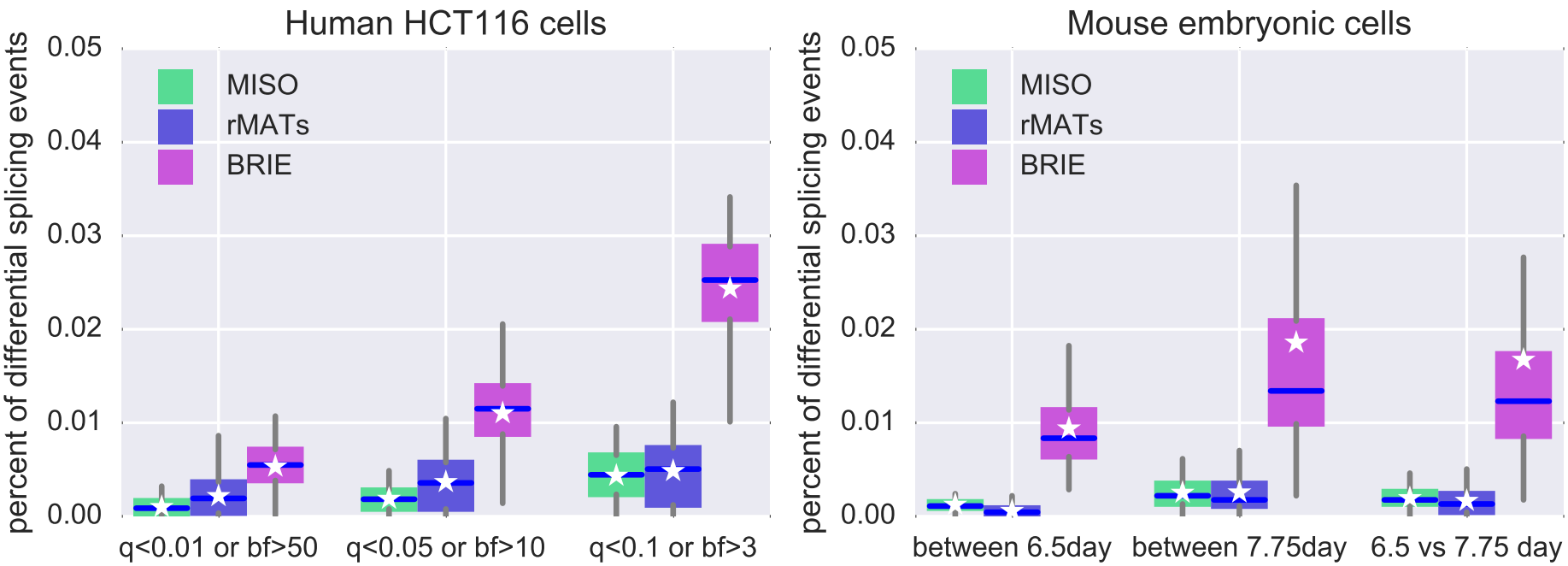
Graphical presentation of brie model

- X_k: feature vector for the kth isoform
- θ_k: expression level for the kth isoform
- I_n: isoform identity for the nth read
- R_n: variable for the nth read
- Bayesian regression: $P(\theta_k|W, X_k, \sigma) = \mathcal{N}(\theta_k|W^\top X_k, \sigma^2)$
- Mixture model: $P(R_{1:N}|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(R_n|I_n = k)P(I_n = k|\Theta)$

BRIE on Wu et al): 11478 skipping exons



BRIE for differential splicing (Wu et al and Brennecke et al)



Talk outline

1. Spatial effects in epigenomic data
   - Statistical testing for epigenomic data
   - Transcription factors and histone modifications
   - Clustering and prediction from epigenomic data
2. Isoform quantification at very low coverage (Y. Huang)
   - Isoform quantification from time series RNA-seq
   - Splicing quantification in single cells
3. Conclusions

Conclusions

- High-throughput biology poses challenges that are beyond classical statistics
- Machine learning can help extract patterns from high-throughput data and suggest biological functions/ clarify links between different data types

Outstanding challenges

- Systematic integration of data sources
- Translational applications: heterogeneity
- How to make causal/ mechanistic inferences

Thanks

School of Informatics

- Gabriele Schweikert
- Dan Benveniste
- Tom Mayo
- Andreas Kapourani
- Yuanhua Huang

Wellcome Trust Centre for Cell Biology

- Adrian Bird
- Jean Beggs

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
- T. Mayo et al, M^3D: a kernel-based test for spatially correlated changes in methylation profiles, Bioinformatics 31(6), 809-816, 2015
- S. Lukauskas et al, DGW: an exploratory data analysis tool for clustering and visualisation of epigenomic marks, BMC Bioinformatics 17(Suppl 16):447, 2016
- A. Kapourani and G.S., Higher order methylation features for clustering and prediction in epigenomic studies, ECCB16/ Bioinformatics 32(17), i405-i412
- Y. Huang and G.S, Statistical modeling of isoform splicing dynamics from RNA-seq time series data, Bioinformatics 32 (19): 2965-2972, 2016
- Y. Huang and G.S, Transcriptome-wide splicing quantification in single cells, http://www.biorxiv.org/content/early/2017/01/05/098517

Software packages

- MMDiff2

  `http://www.bioconductor.org/packages/release/bioc/html/MMDiff2.html`
- M3D `http://www.bioconductor.org/packages/devel/bioc/html/M3D.html`
- BPRMeth

  `https://bioconductor.org/packages/release/bioc/html/BPRMeth.html`
- DGW `https://pypi.python.org/pypi/dgw`
- DICEseq `https://pypi.python.org/pypi/diceseq/0.2.6`
- BRIE `https://pypi.python.org/pypi/brie/0.1.0`