Some loose ends:

- data along genome: peak finding (ChIP-seq) versus HMMs versus segmentation
- functional category analysis / gene-set testing
- robustness

Final lecture: 17 Dec 2018

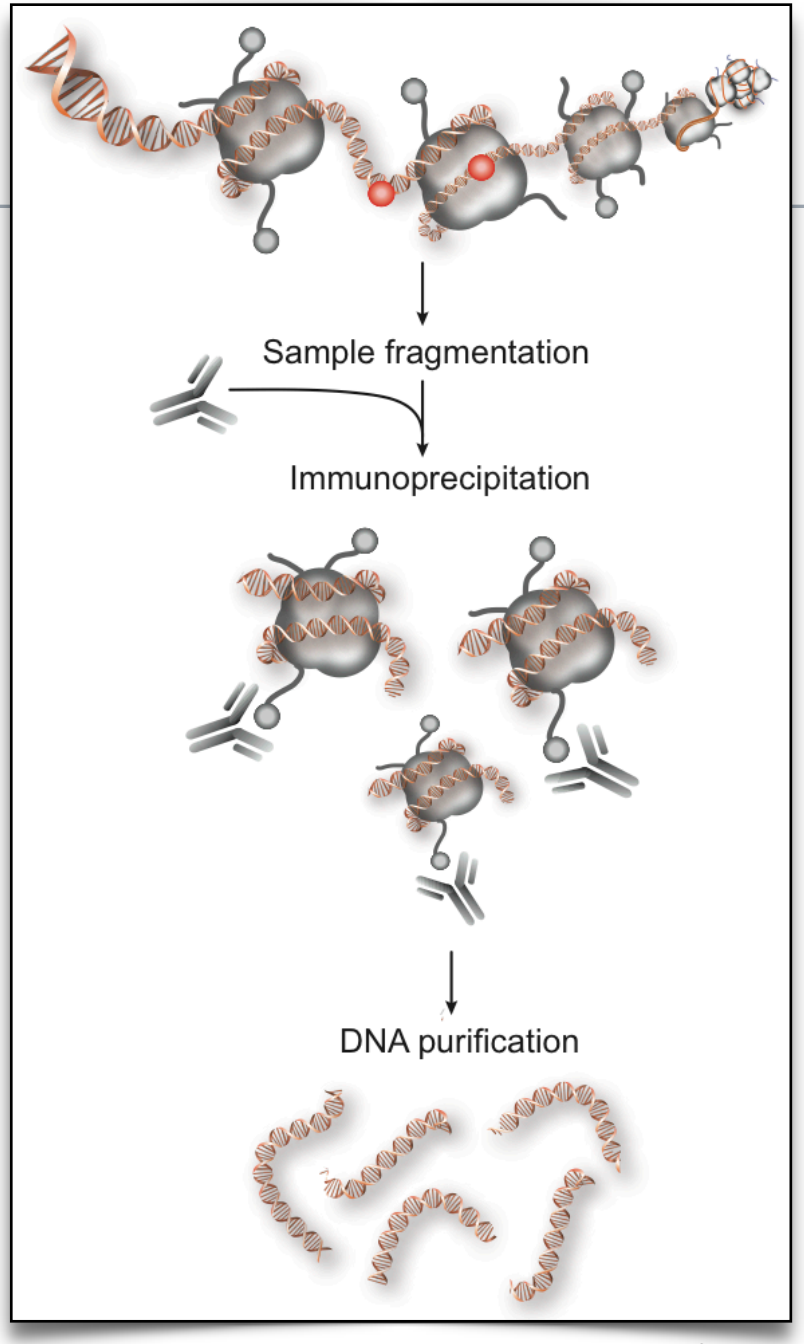
– There is no Journal Club next week; we will start in 01-F-50 with the final workflow day (single cell RNA-seq preprocessing)

Chromatin immunoprecipitation for protein-DNA interactions

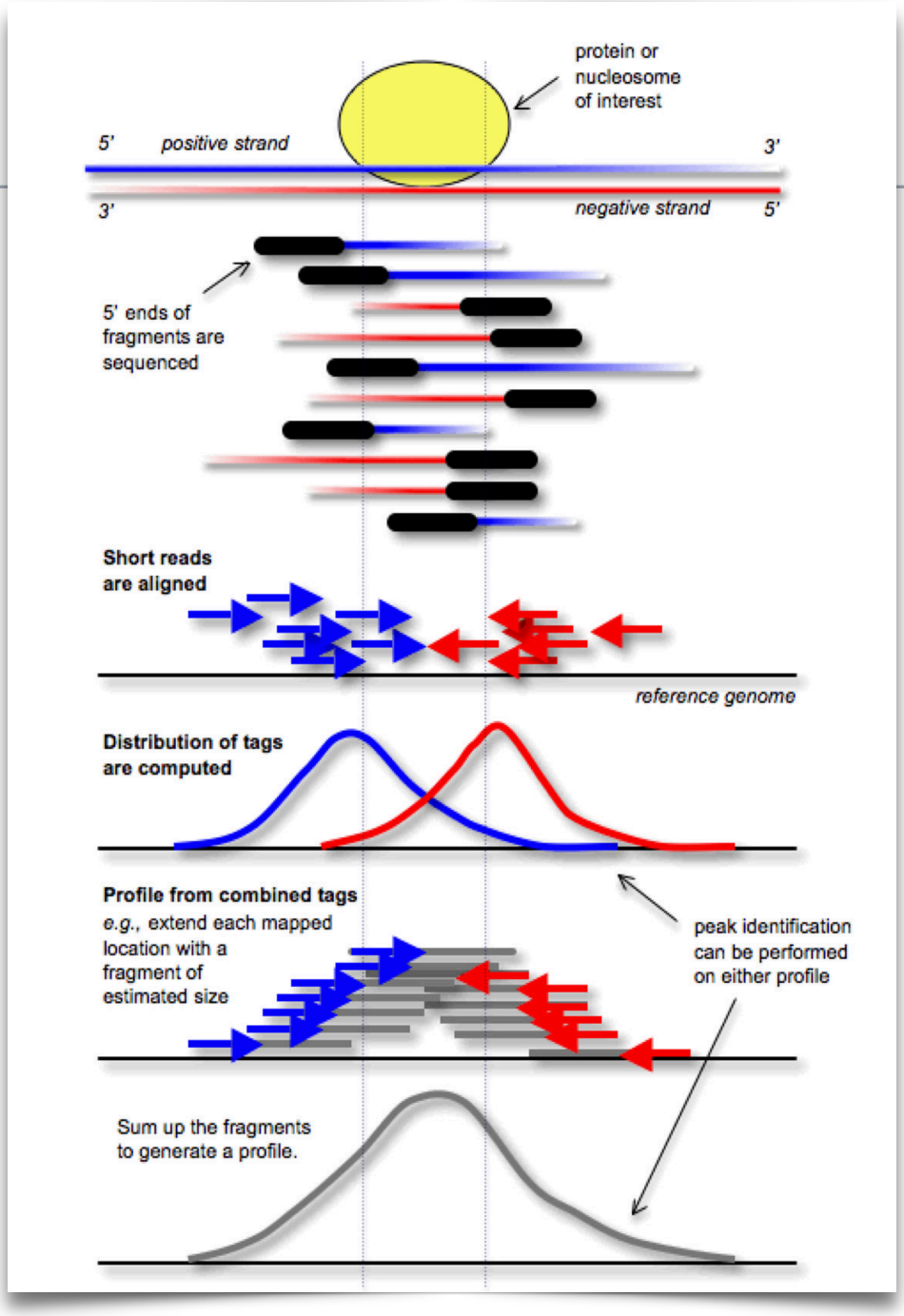
A very basic summary of the histone code for gene expression status is given below (histone nomenclature is described here):

Type of modification	Histone						
	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
mono-methylation	activation[6]	activation[7]		activation[7]	activation[7][8]	activation[7]	activation[7]
di-methylation		repression[3]		repression[3]	activation[8]		
tri-methylation	activation[9]	repression[7]		repression[7]	activation,[8] repression[7]		repression[3]
acetylation		activation[9]	activation[9]				

- H3K4me3 is found in actively transcribed promoters, particularly just after the transcription start site.
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.



Peak/region detection for ChIP-seq data



Region/peak finding depends on the association (epigenetic mark versus transcription factor)

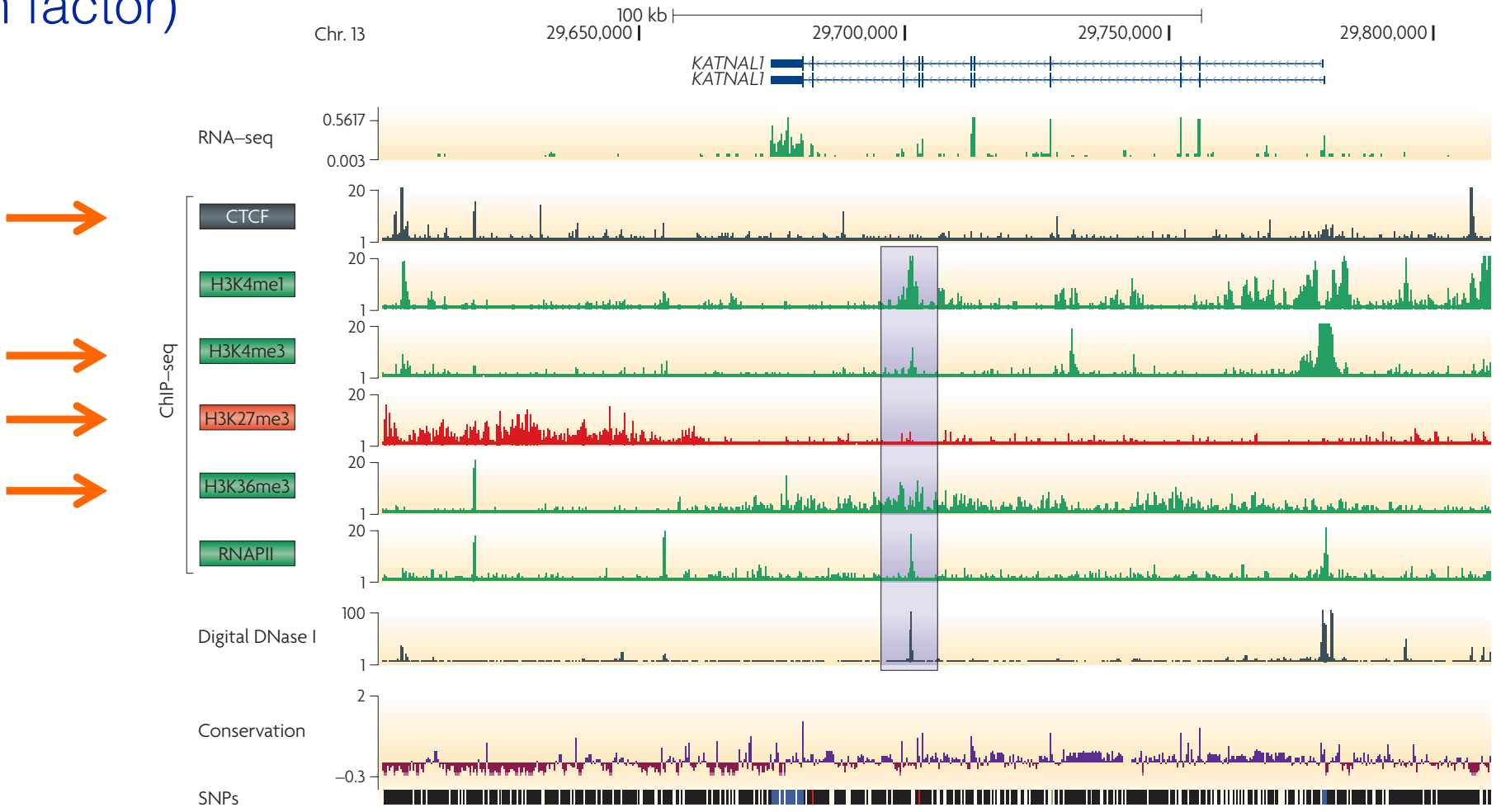
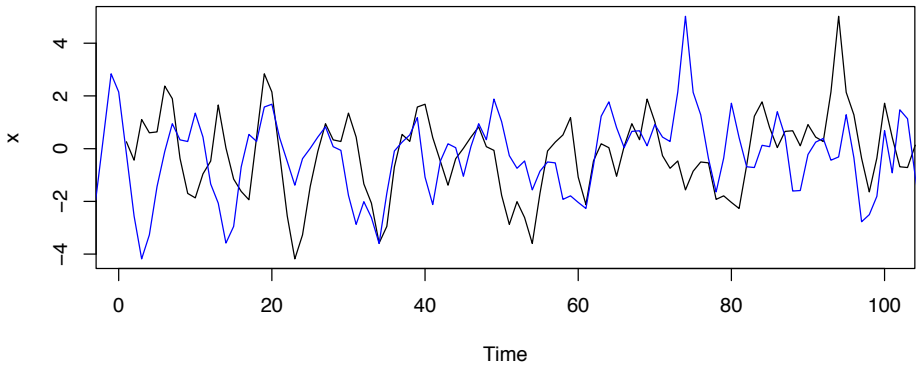


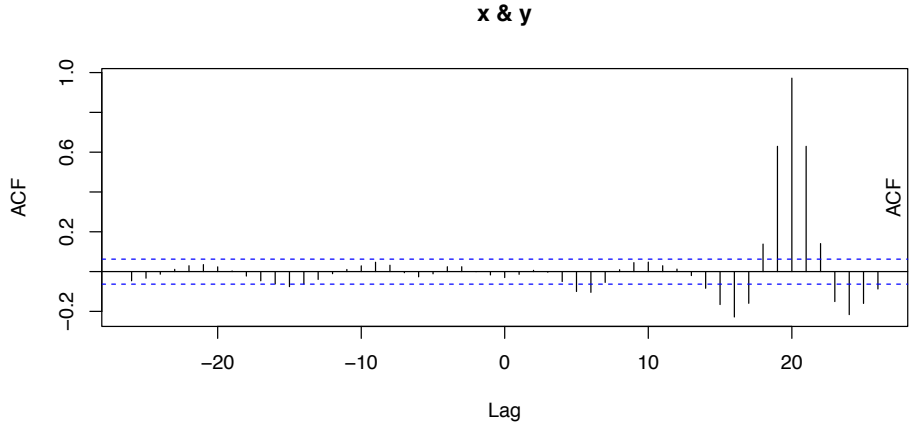
Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing

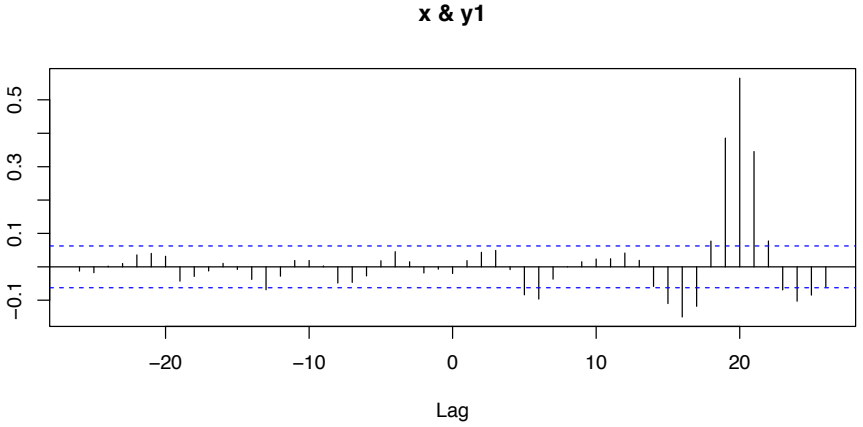
$$(f \star g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f^*[m]\ g[n+m].$$

Cross correlation

```
x <- arima.sim(model=list(ar=c(.99,-.5)),n=1000)
y <- lag(x,20)

plot(x, type="l",xlim=c(1,100))
lines(y, col="blue")


ccf(x,y)

y1 <- lag(x,20)+rnorm(100,sd=2)
ccf(x,y1)
```



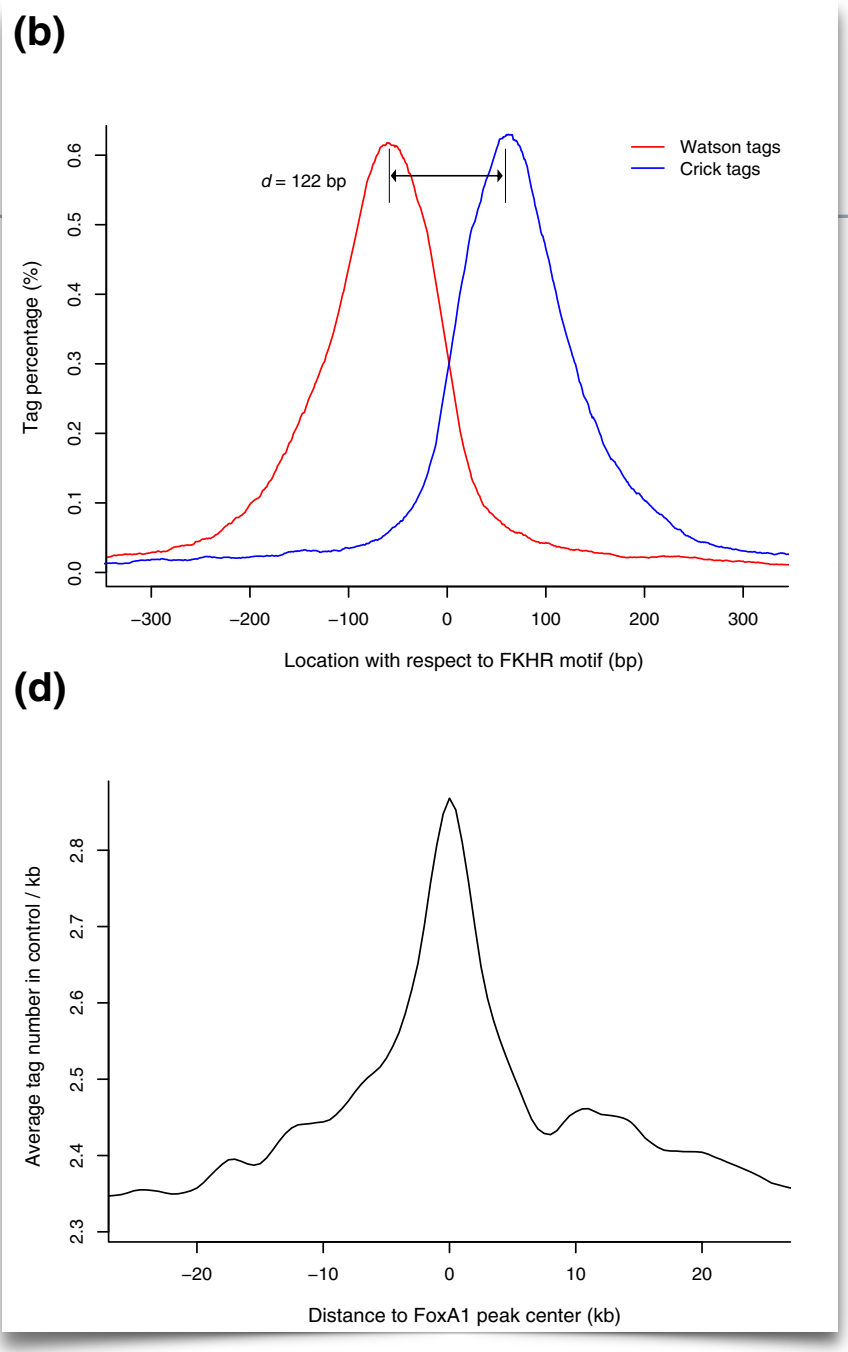




MACS – model-based analysis of ChIP-seq

Simple but effective algorithm:

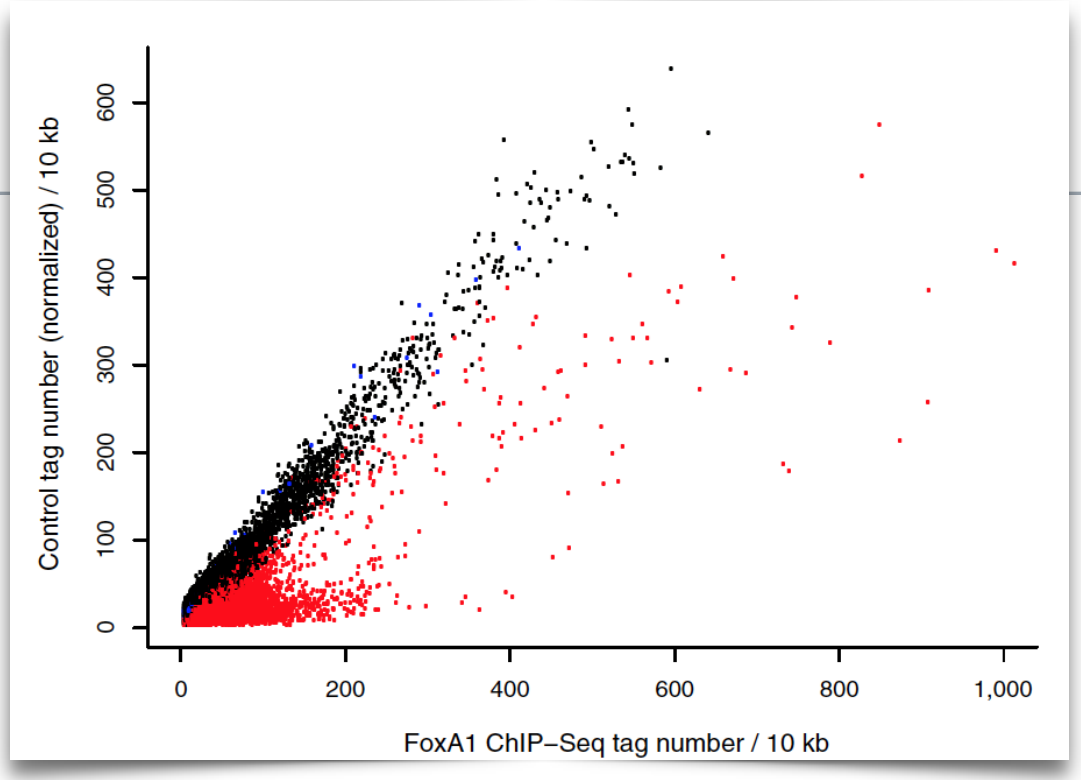
1. Estimate average fragment size ‘d’ (cross-correlation)
2. Adjust reads by d/2
3. From control sample, estimate local background (if control sample used)
4. For each window, calculate Poisson P-value (probability of more extreme than local rate)
5. Estimate empirical FDR



MACS – model-based analysis of ChIP-seq

Simple but effective algorithm:

1. Estimate average fragment size 'd' (cross-correlation)
2. Adjust reads by d/2
3. From control sample, estimate local background (if control sample used)
4. For each window, calculate Poisson P-value (probability of more extreme than local rate)
5. Estimate empirical FDR



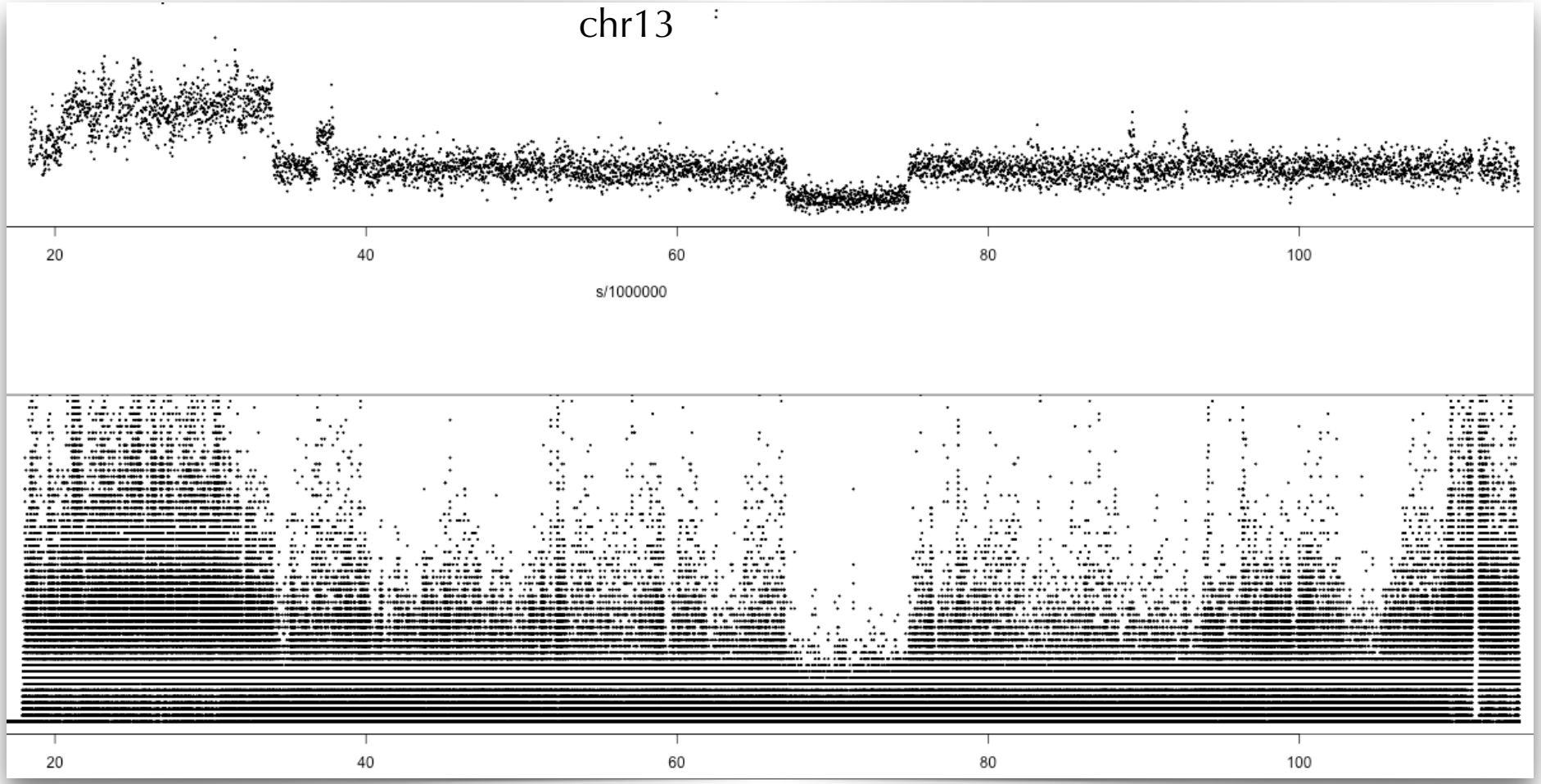
$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$$

For a ChIP-Seq experiment with controls, MACS empirically estimates the false discovery rate (FDR) for each detected peak using the same procedure employed in the previous ChIP-chip peak finders MAT [13] and MA2C [14]. At each *p*-value, MACS uses the same parameters to find ChIP peaks over control and control peaks over ChIP (that is, a sample swap). The empirical FDR is defined as Number of control peaks / Number of ChIP peaks. MACS can also be applied to

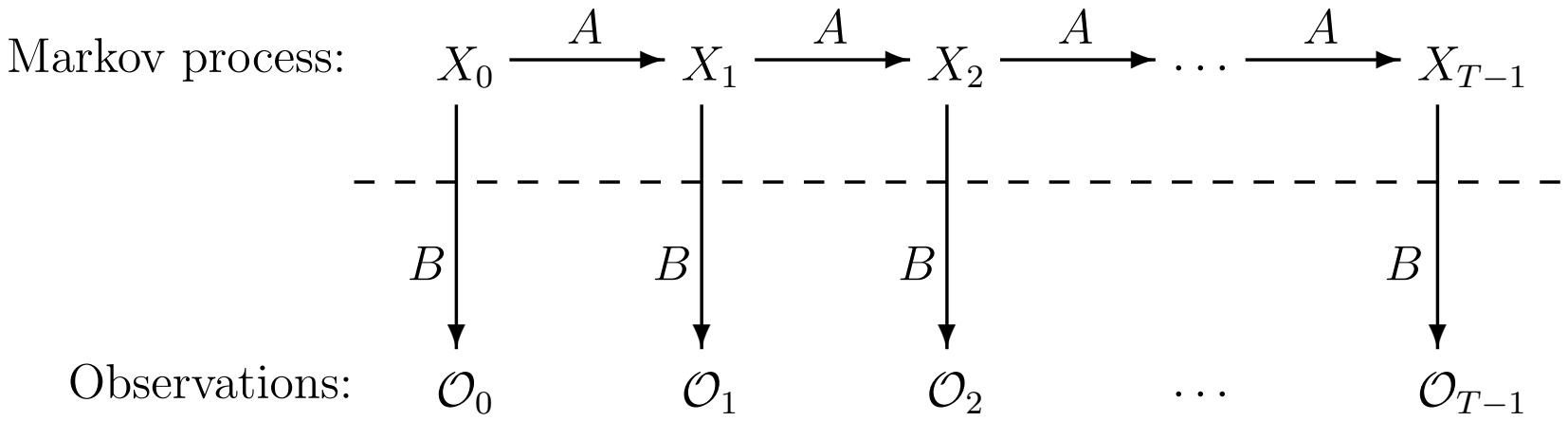
quantitative DNA-seq signal
= biology (copy number, enrichment) + technical effects + noise

Copy number (normalized read depth)

ChIP-seq (K27me3)



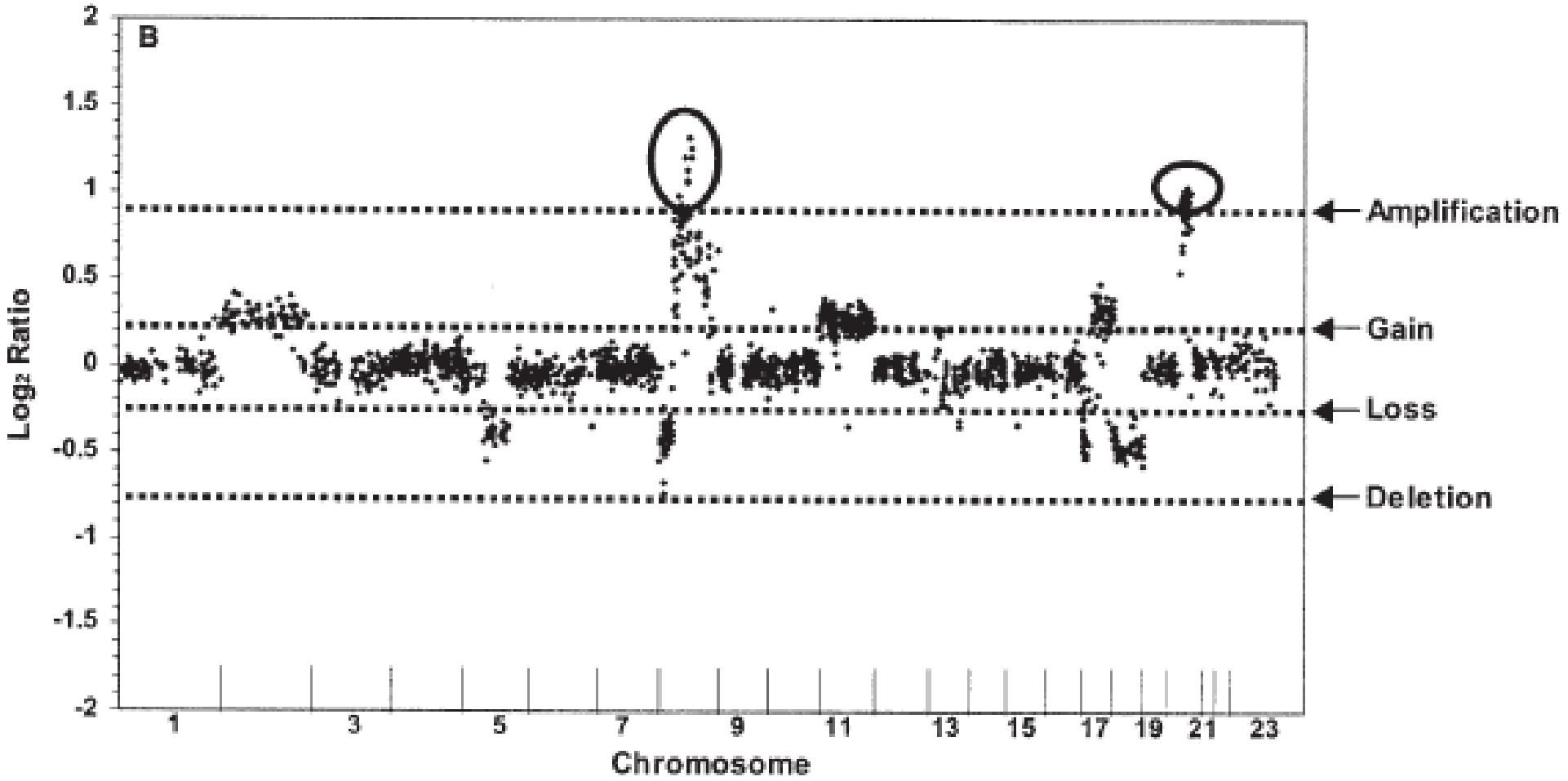
Introduction to Hidden Markov Models



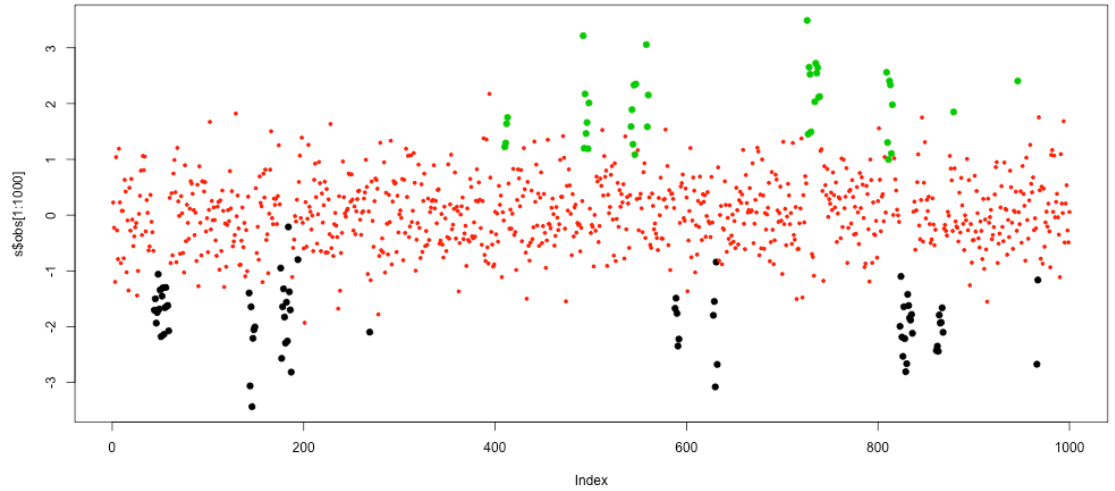
X_i – hidden ("latent", unobserved state)
O_i – "emitted" observation

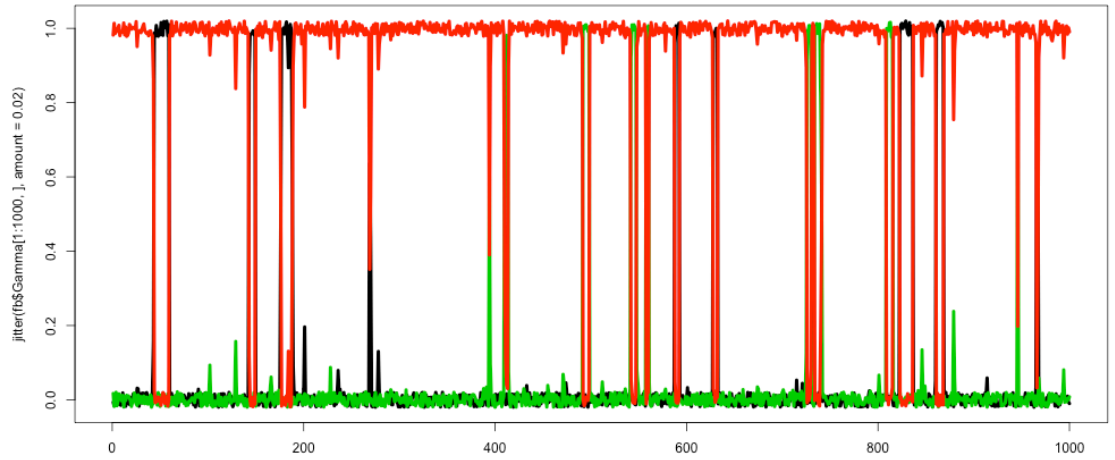
A – transition probabilities
B – emission probabilities/distributions

Examples: HMMs in genomics – copy number



A “vanilla” HMM: Normal emission distributions



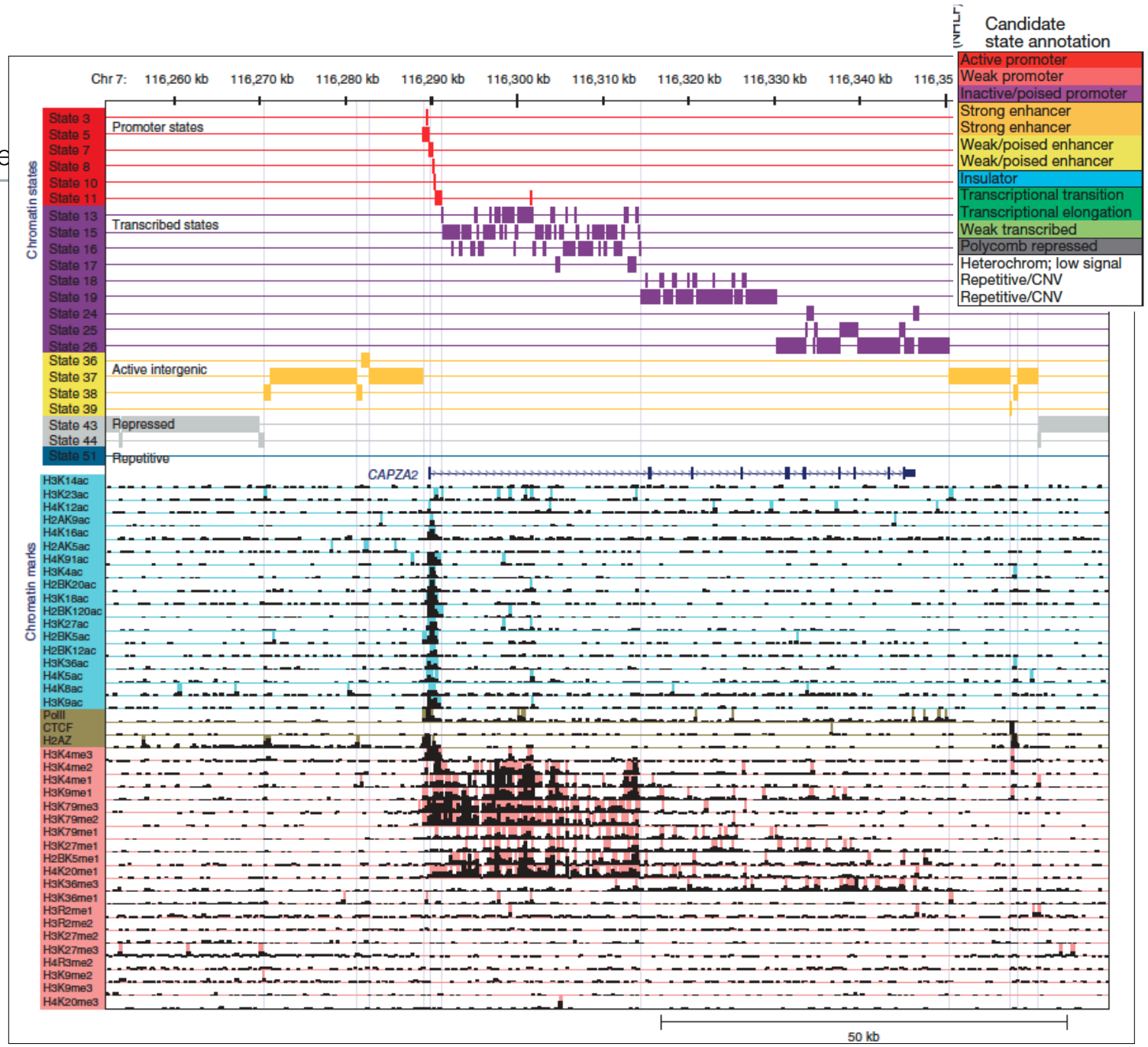



```
library(RHmm)
h <- distributionSet("NORMAL", mean=c(-2, 0, 2),
                                var=c(.4, .4, .4))
ip <- c(0,1,0)
tr <- rbind(c(.8,.2,0), c(.01, 0.98, .01), c(0,.2,.8) )


hs <- HMMSet(ip, tr, h)
s <- HMMSim(5000, hs)

hf <- HMMFit(s$obs, nStates=3)
fb <- forwardBackward(hf, s$obs)

r <- rank(hf$HMM$distribution$mean)

par(mfrow=c(2,1))
plot( s$obs[1:1000], col=(1:3)[s$states], pch=19,
      cex=c(1,.5,1)[s$states] )
matplot(jitter(fb$Gamma[1:1000,],amount=.02), col=r,
        type="l", lwd=4, lty=1)

> tr
     Bl    R    G
Bl 0.80 0.20 0.00
R  0.01 0.98 0.01
G  0.00 0.20 0.80
```

Exploratory analyses with HMMs

Every 200bp region of the genome is binarized based on a background model

Multivariate HMM is trained; genome is partitioned into 15 states

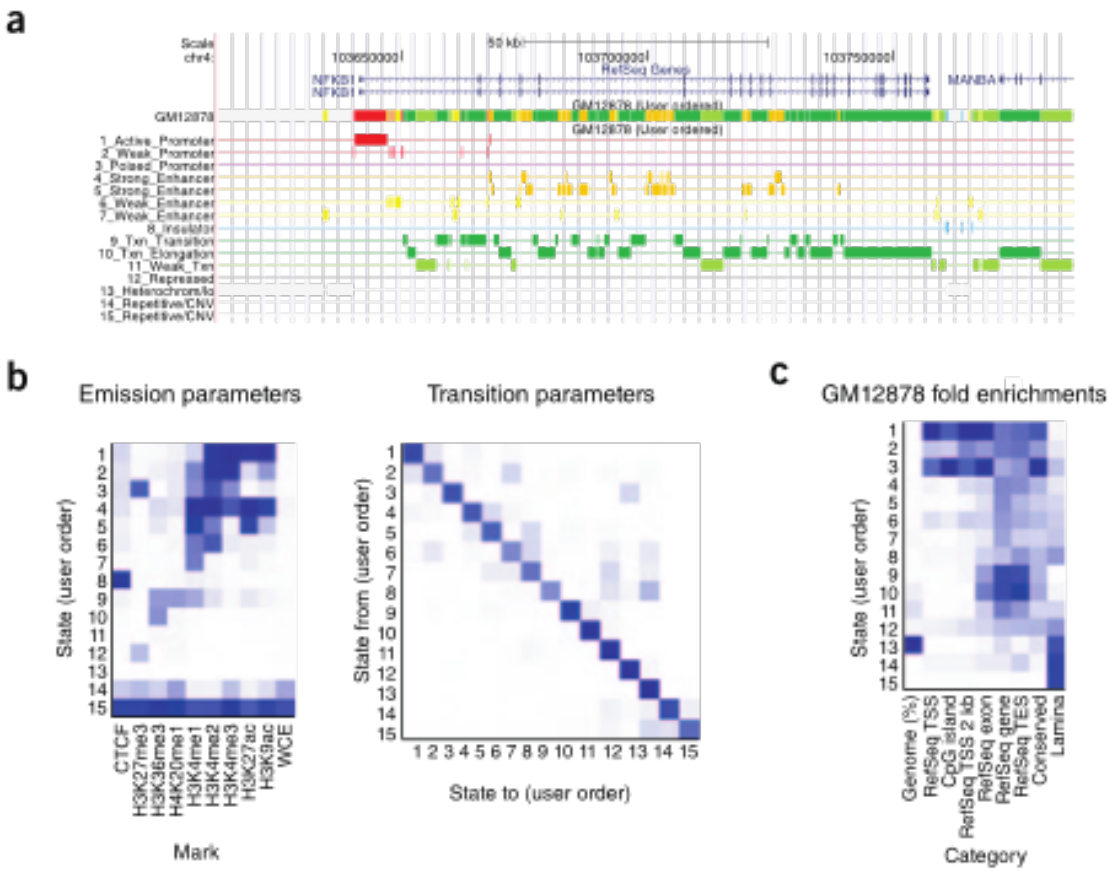


Ernst et al., Nature 2010

Ernst and Kellis, Nature Biotech 2010

ChromHMM

ChromHMM is based on a multivariate hidden Markov model that models the observed combination of chromatin marks using a product of independent Bernoulli random variables[2], which enables robust learning of complex patterns of many chromatin modifications. As input, it receives a list of aligned reads for each chromatin mark, which are automatically converted into presence or absence calls for each mark across the genome, based on a Poisson background distribution. One can use an optional addi-



Ernst and Kellis, Nature Methods (March) 2012

BayesPeak:

$$
\begin{aligned}
Y_t^+, Y_{t+1}^- \mid Z_t = 0 &\sim \text{Poisson}(\lambda_0 \gamma^{w_t}) \\
Y_t^+, Y_{t+1}^- \mid Z_t = 1,2,3 &\sim \text{Poisson}((\lambda_0 + \lambda_1)\gamma^{w_t}) \\
\lambda_0 &\sim \Gamma(\alpha_0, \beta_0) \\
\lambda_1 &\sim \Gamma(\alpha_1, \beta_1)
\end{aligned}
$$

$$
Z_t = \begin{cases} 0 & \text{if} \quad (S_t, S_{t+1}) = (0,0) \\ 1 & \text{if} \quad (S_t, S_{t+1}) = (0,1) \\ 2 & \text{if} \quad (S_t, S_{t+1}) = (1,0) \\ 3 & \text{if} \quad (S_t, S_{t+1}) = (1,1) \end{cases}
$$

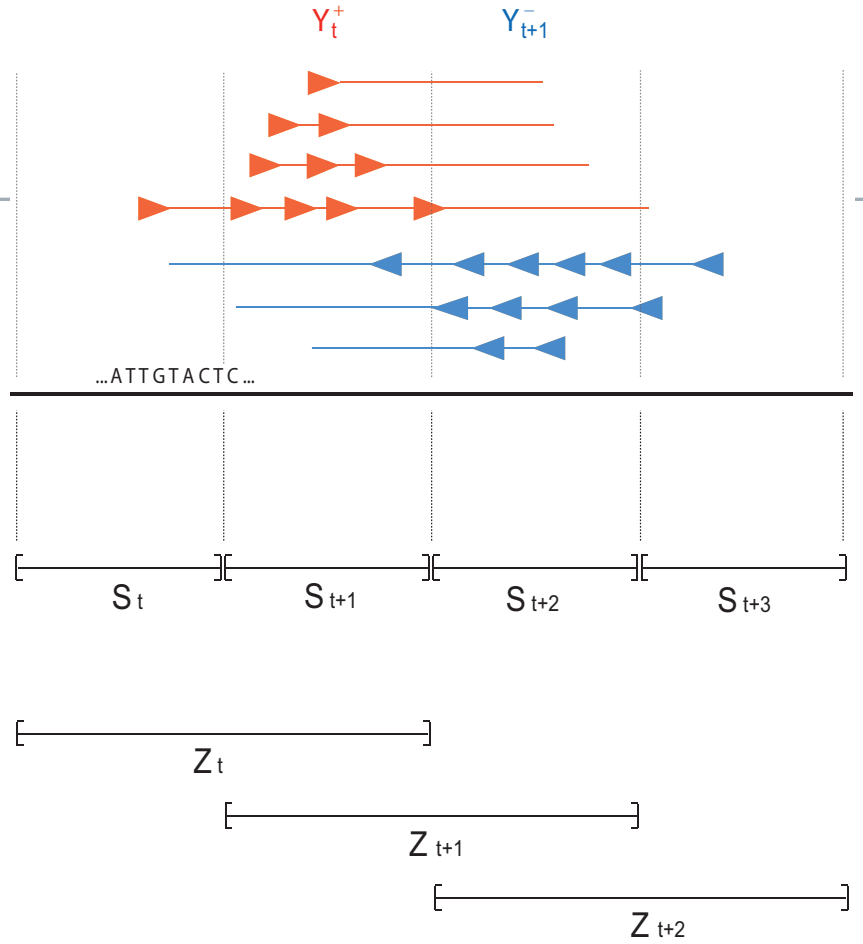


Figure 1
Illustration of the model. This figure shows how the reads (arrows) on the forward and reverse strand, indicated by red and blue respectively, are counted as Y_t^+ and Y_{t+1}^- and depend on the nature of the underlying regions *t* and *t* + 1 when their full length is taken into consideration. Moreover, this figure shows how each Z_t state corresponds to the underlying ones S_t and S_{t+1}.

BayesPeak models +/- strands directly

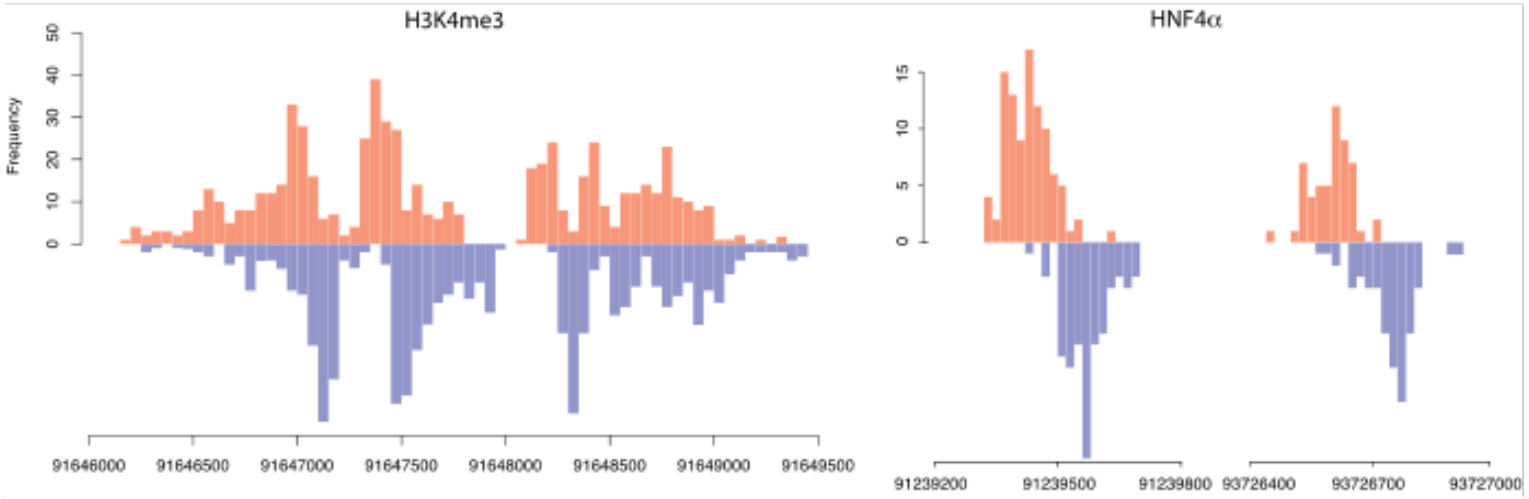


Figure 3
A closer view of some HeK4me3 and HNF4α peaks. These histograms present the counts of the 5' ends of the reads from the H3K4me3 and the HNF4α data, forming peaks on the forward (red) and reverse (blue) strand. The offset between them shows how the enclosed area corresponds to an enriched region. The plots are on a different scale to show the density of reads clearly and highlight the difference between the peaks formed by a histone mark and a transcription factor.

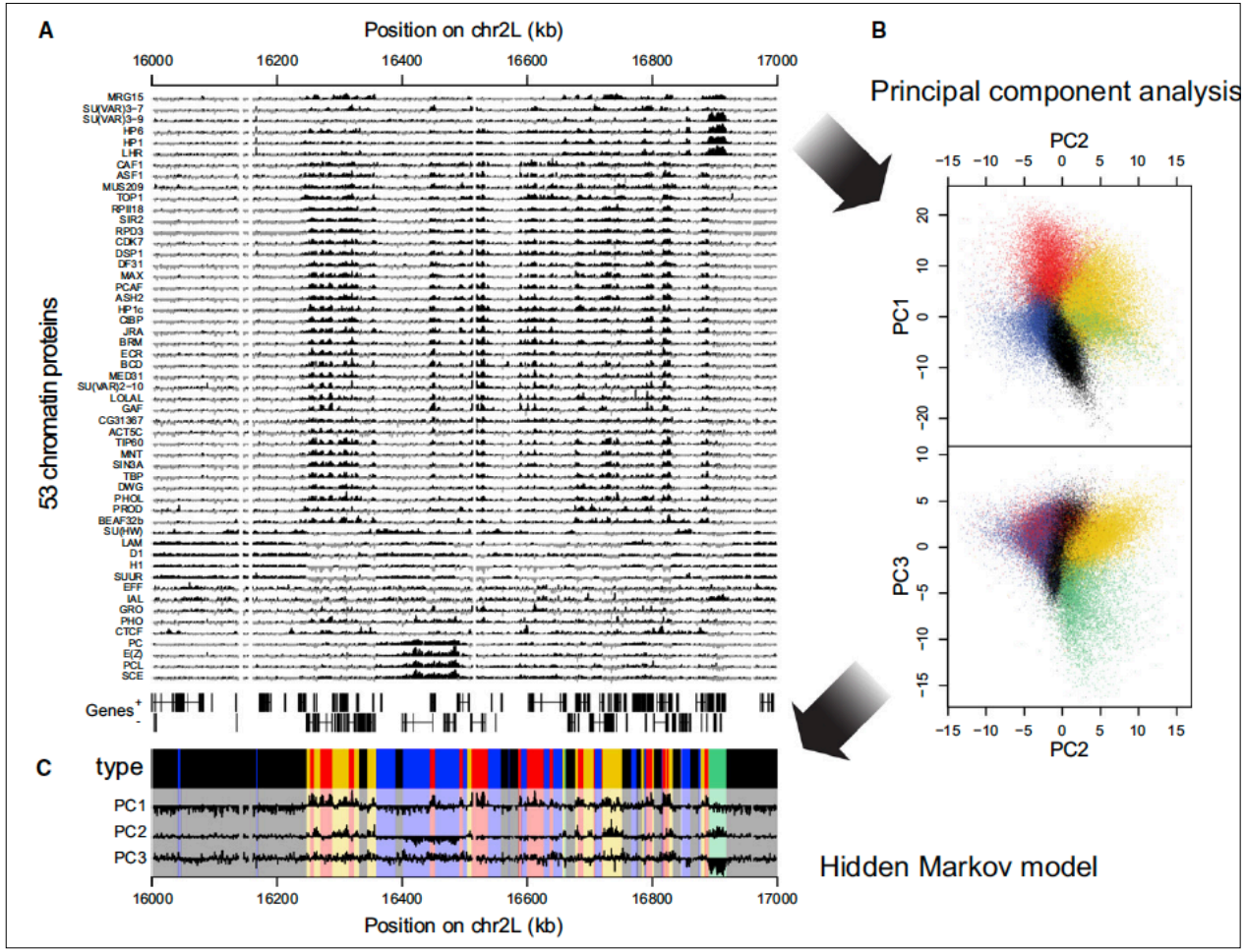
Exploratory analyses with HMMs

53 chromatin factors (ChIP-seq)

Compression to 3 principal components

Learn HMM

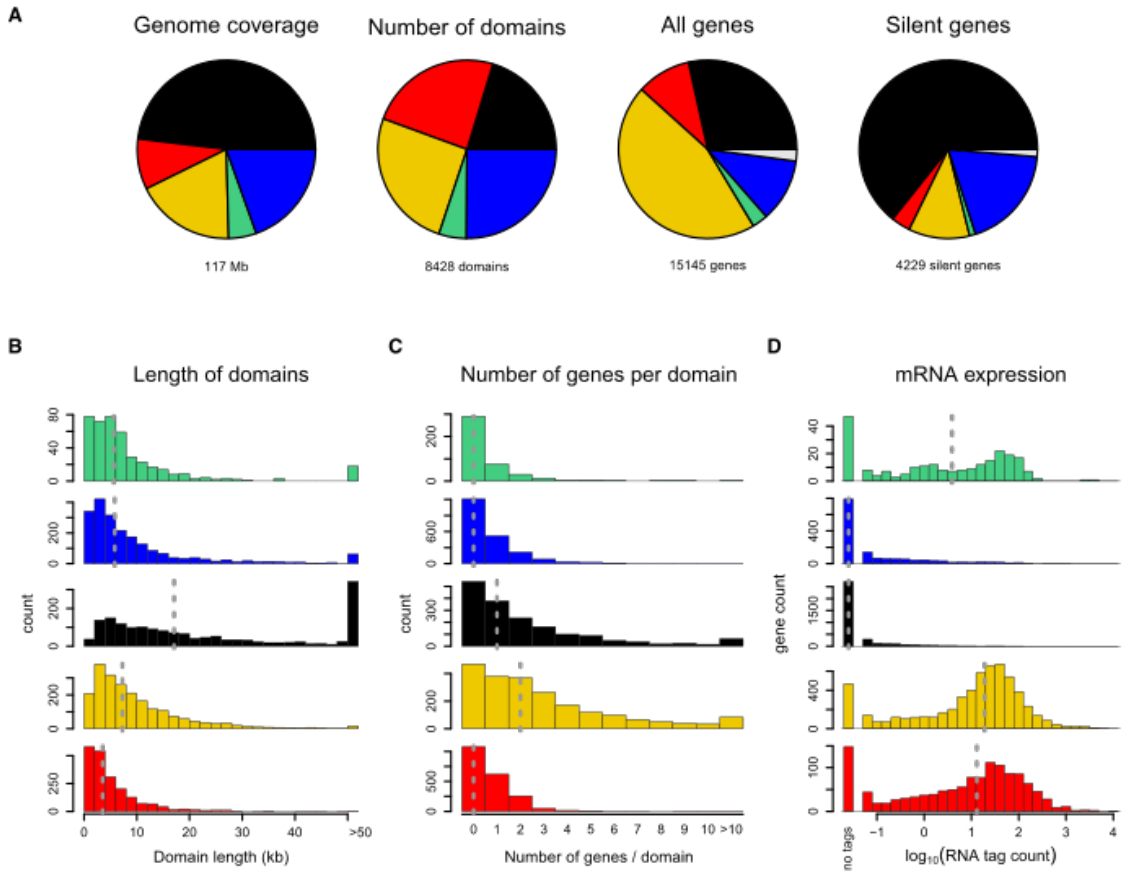
Every region of the genome partitioned into 5 “states” (here, assigned a colour)



Filion et al. Cell 2010

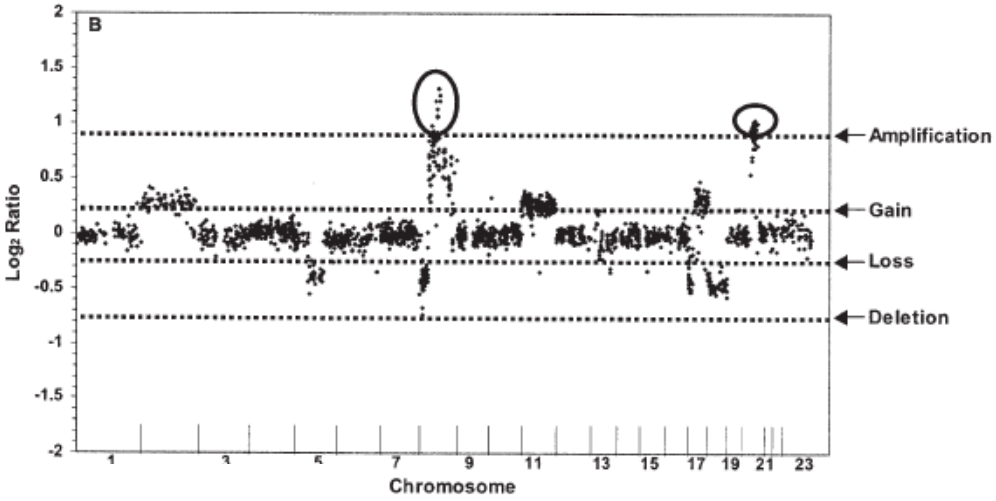
X

"Colours" are reflective of various features



Filion et al. Cell 2010

Alternative to HMM: Segmentation

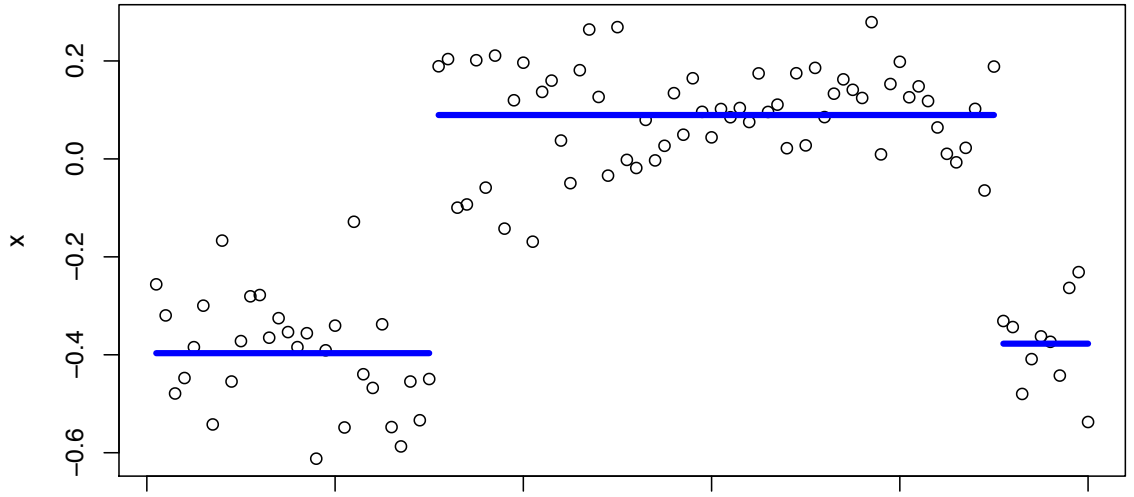


Partial sums: $S_i = X_1 + \cdots + X_i, 1 \leqslant i \leqslant n.$

Difference in partial sum: $Z_i = \{1/i + 1/(n-i)\}^{-1/2}\{S_i/i - (S_n - S_i)/(n-i)\}.$

$$Z_{ij} = \{1/(j-i) + 1/(n-j+i)\}^{-1/2}\{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)\}.$$

Our modification of the binary segmentation procedure, which we call *circular binary segmentation* (CBS), is based on the statistic $Z_C = \max_{1 \leqslant i < j \leqslant n} |Z_{ij}|$. Note that Z_C allows for both a single change



Olshen et al. 2014
Circular binary segmentation for the analysis of array-based DNA copy number data.

OPEN ACCESS Freely available online

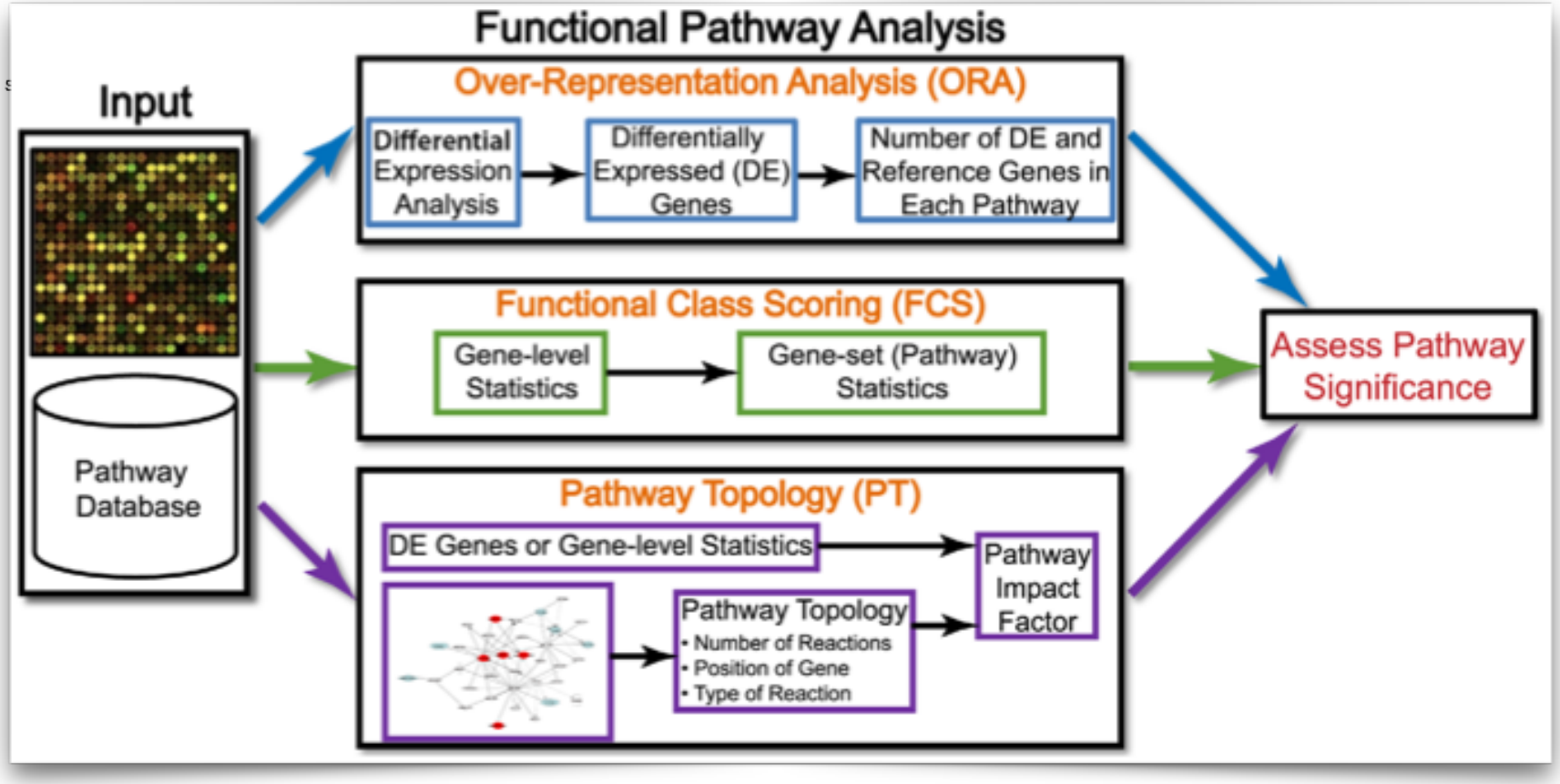
PLOS COMPUTATIONAL BIOLOGY

Review

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri[1,2*], Marina Sirota[1,2], Atul J. Butte[1,2*]

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, S
Children's Hospital, Palo Alto, California, United States of America



Casting differential expression onto biological knowledge: Functional category analysis versus gene set analysis

Motivation: DE genes might belong to a known pathway or might be the top genes from a related experiment; gene set as a whole might be altered, even if individual genes are not.

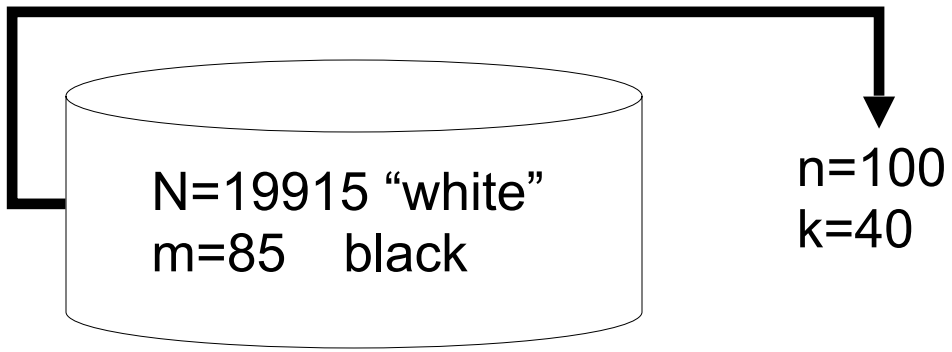
Starting point:	threshod, set of DE genes	gene-level statistics
Tool examples:	DAVID [C] **goseq** [C]	GSEA [S] **roast** [S] **CAMERA** [C]

S = self-contained
C = competitive

Functional category analysis: Overlap statistics

Question: Say you have a set of 85 genes (of a total 20000 genes) known to be associated with some function. Calculate the probability of randomly selecting 40 or more (overrepresented) of those genes in a list of 100 DE genes.

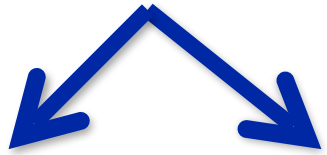
Answer: Hypergeometric (i.e. the "urn" problem).



$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

e.g. FunSpec (yeast) - Robinson et al. 2002 BMC Bfx; DAVID; topGO

Gene set analysis: what is the hypothesis (test)?

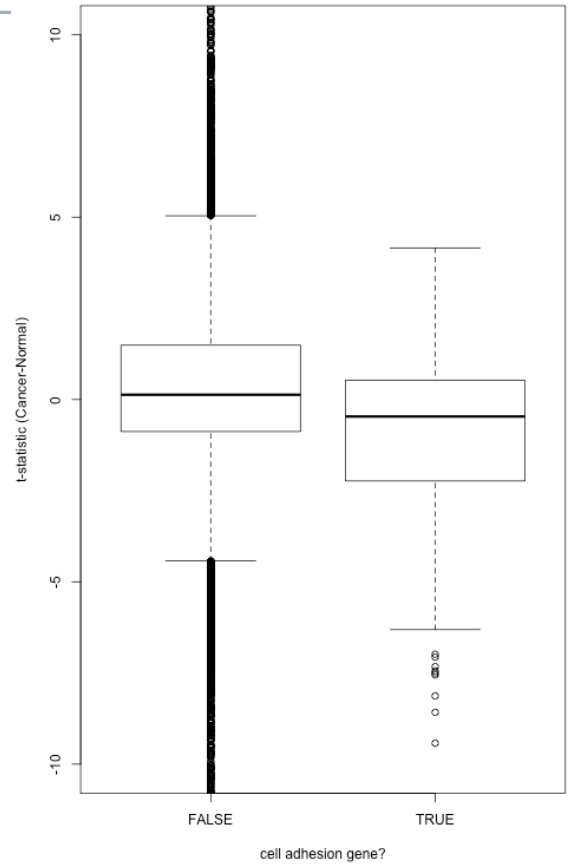


competitive	self-contained
Genes in the set tend to be more strongly DE than randomly chosen genes	*At least some genes in the set are truly DE*

Goeman & Bühlmann, *Bioinformatics* 2007

Viewing gene sets

Cell adhesion genes



Genes regulated by MYB

Gene set enrichment analysis (GSEA)

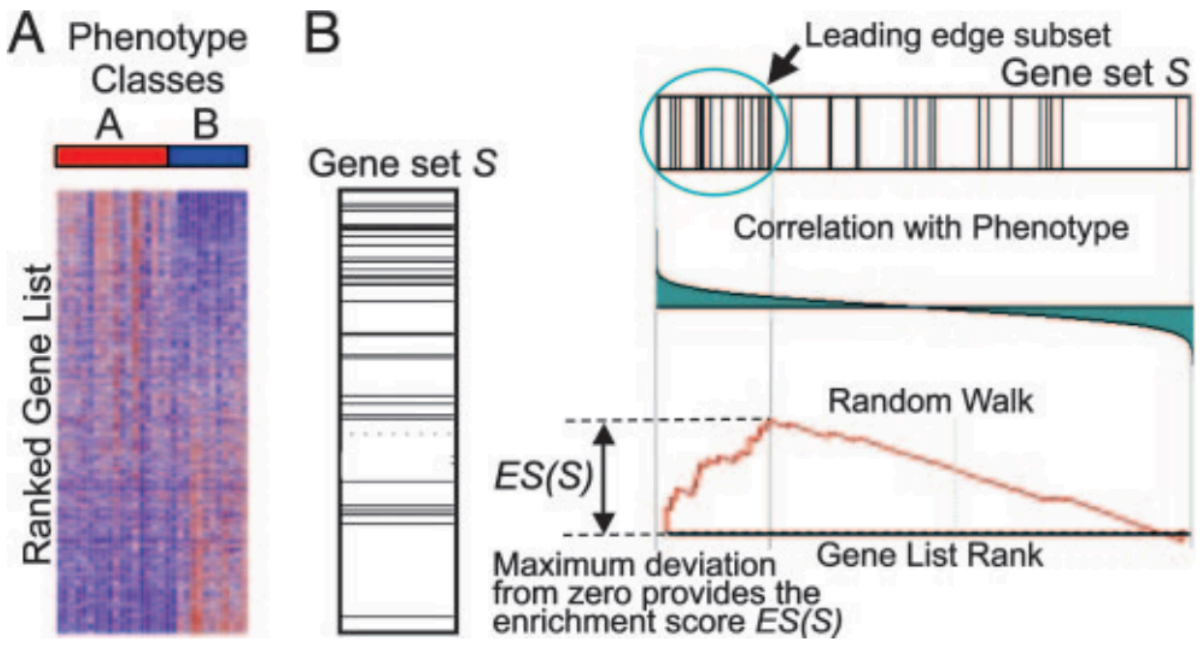


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Self-contained.

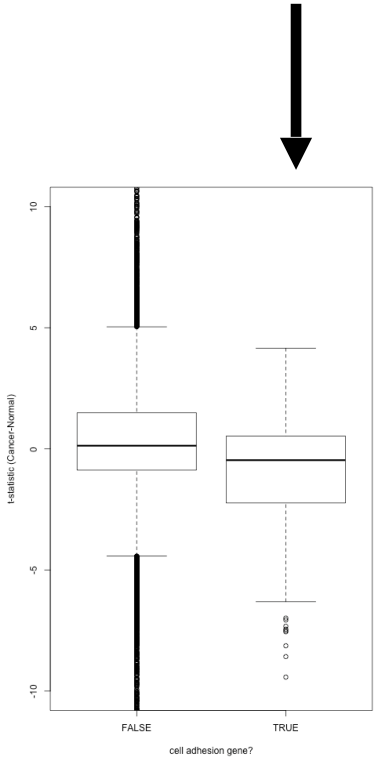
Permutation P-value: Sample permutation is done, which preserves gene correlation.

But, it has limited use in small samples (i.e. very few possible permutations).

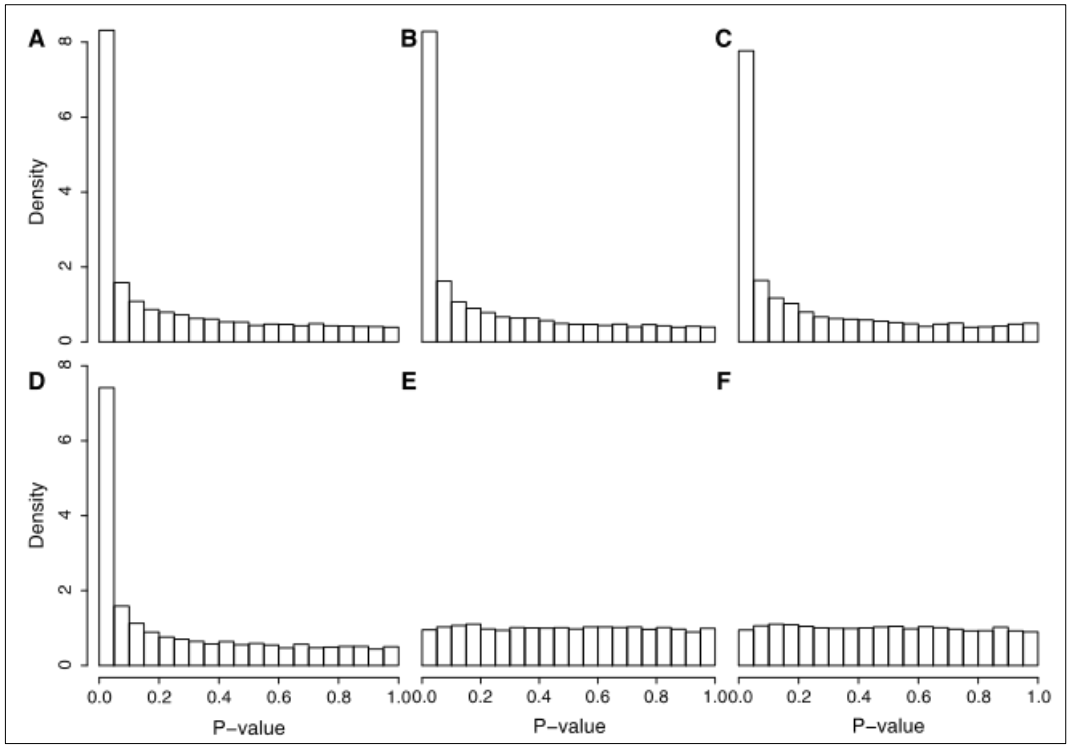
Now switches to a gene-based permutation (competitive) in small samples.

CAMERA (Correlation Adjusted MEan RAnk)

Cell adhesion genes



Main criticism of (naïve, gene-permutation) competitive tests is that the correlation structure is broken.



Distributions of p-value:

no differential expression

A geneSetTest
B geneSetTest [r]
C sigPathway
D PAGE
E CAMERA
F CAMERA [r]

Wu et al. NAR 2012

How much of this is storytelling?

A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans

Pavlos Pavlidis,*[1] Jeffrey D. Jensen,[2] Wolfgang Stephan,[3] and Alexandros Stamatakis[1]

[1]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

[2]Ecole Polytechnique Fédérale de Lausanne, School of Life Sciences, Lausanne, Switzerland

[3]Section of Evolutionary Biology, Biocenter, University of Munich, Planegg-Martinsried, Germany

***Corresponding author:** E-mail: pavlidisp@gmail.com.

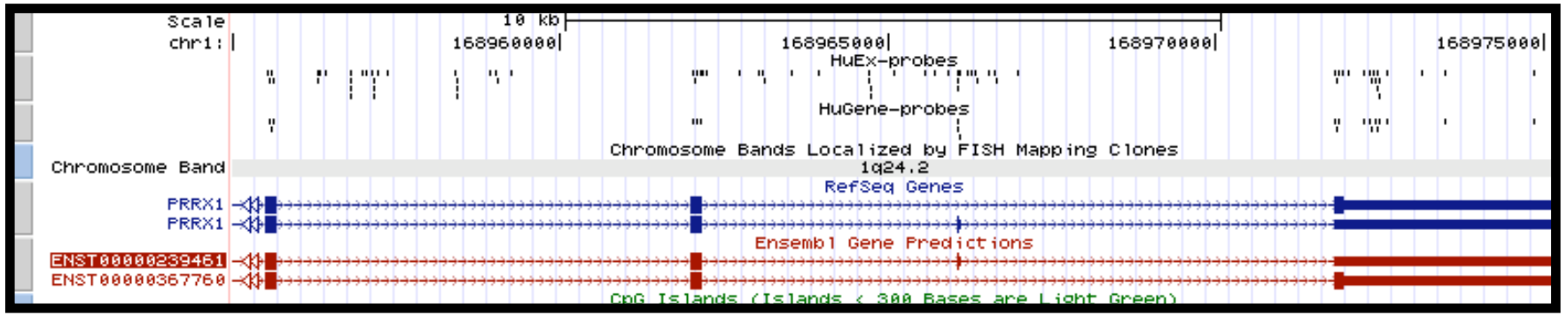
Associate editor: Arndt von Haeseler

Abstract

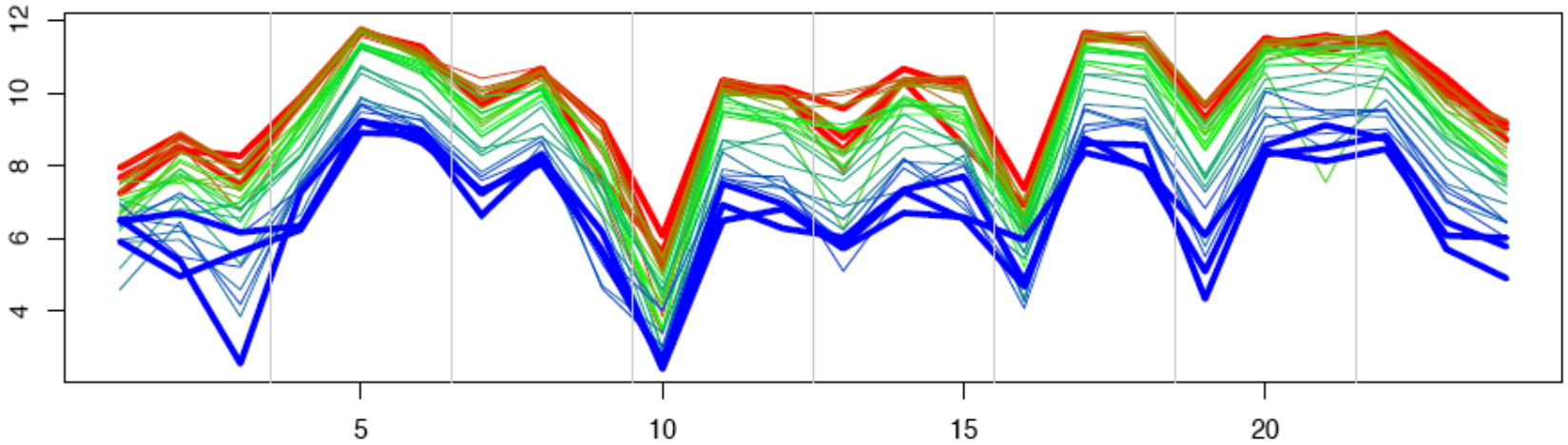
In the age of whole-genome population genetics, so-called genomic scan studies often conclude with a long list of putatively selected loci. These lists are then further scrutinized to annotate these regions by gene function, corresponding biological processes, expression levels, or gene networks. Such annotations are often used to assess and/or verify the validity of the genome scan and the statistical methods that have been used to perform the analyses. Furthermore, these results are frequently considered to validate "true-positives" if the identified regions make biological sense a posteriori. Here, we show that this approach can be potentially misleading. By simulating neutral evolutionary histories, we demonstrate that it is possible not only to obtain an extremely high false-positive rate but also to make biological sense out of the false-positives and construct a sensible biological narrative. Results are compared with a recent polymorphism data set from *Drosophila melanogaster*.

Key words: genome scanning, positive selection, gene ontology, validation, literature mining.

The nature of Affymetrix Probe Level Data

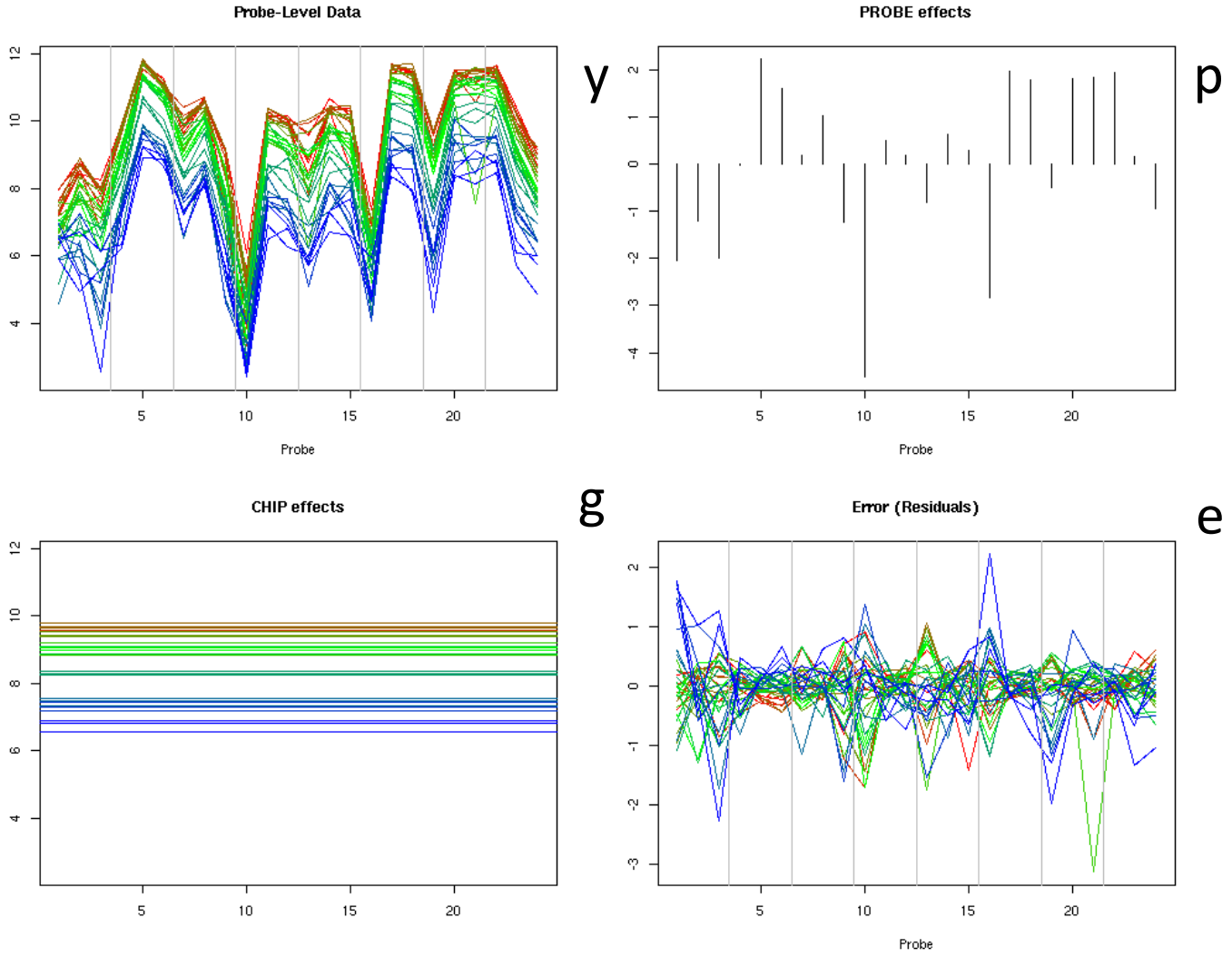


HuGene data [red–heart,blue–brain,mixtures] 10 ENSG00000116132



- Data for one gene that is differentially expressed between heart (red is 100% heart) and brain (blue is 100% brain).
- 11 mixtures x 3 replicates = 33 samples (33 lines)
- Note the parallelism: probes have different affinities

Linear model decomposes the probe-level data into PROBE effects and CHIP effects



Linear model:

$$y_{ik} = g_i + p_k + e_{ik}$$

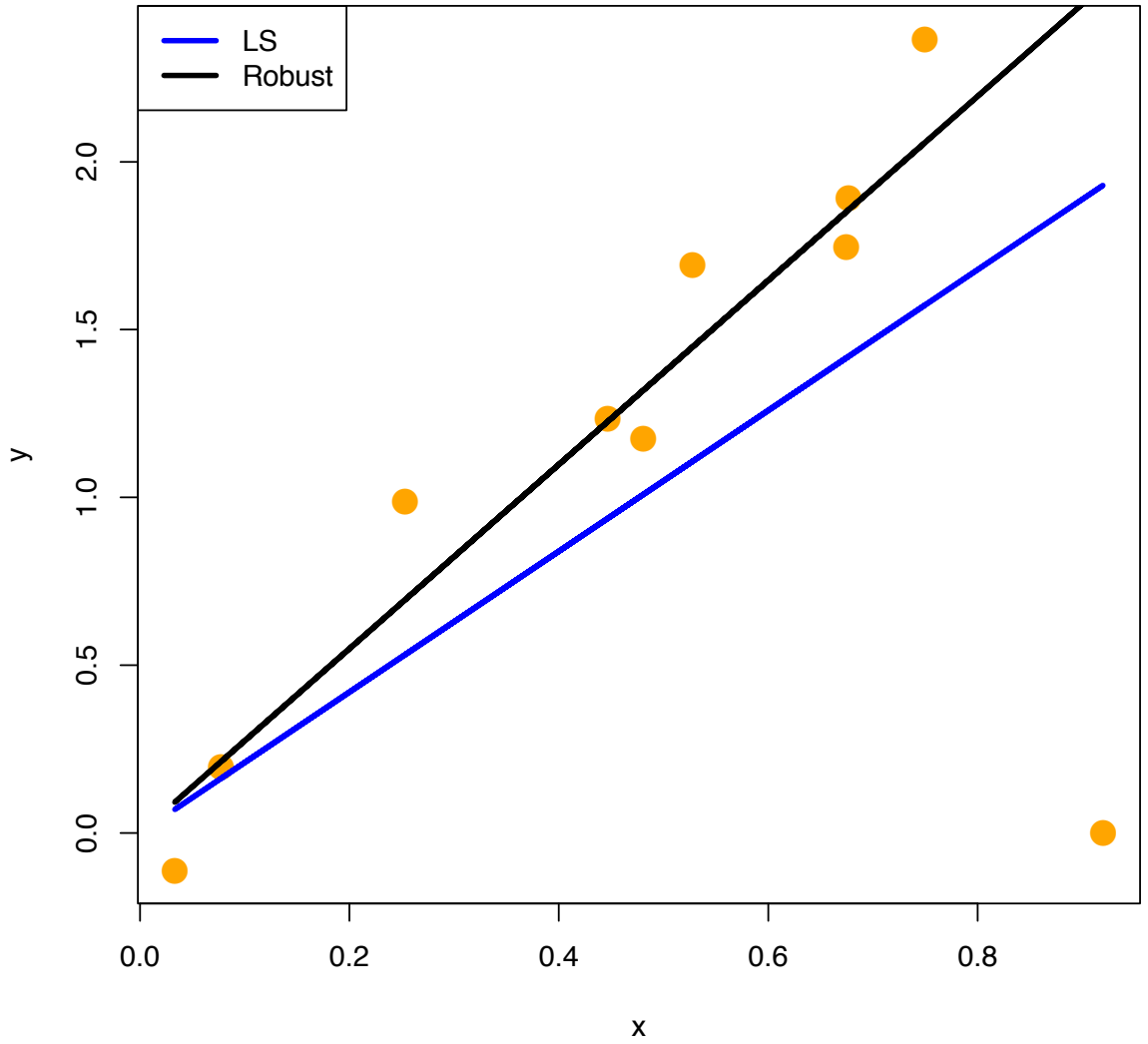
Robust Multichip Analysis (RMA) uses this model.
Irizarry et al. 2003, Biostatistics

Parameters are estimated robustly, meaning a small number of outliers have minimal effect

Tissue mixture dataset

http://en.wikipedia.org/wiki/Ordinary_least_squares

Robust regression – motivating example



OLS = ordinary least squares

The OLS estimator is … optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated … OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances.

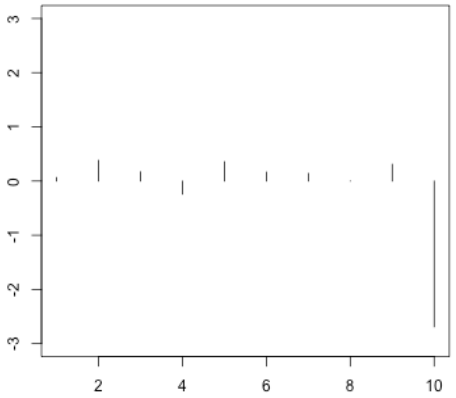
i.e., OLS has good properties, when the data is "nice" (approximately normally distributed).

http://en.wikipedia.org/wiki/Iteratively_reweighted_least_squares

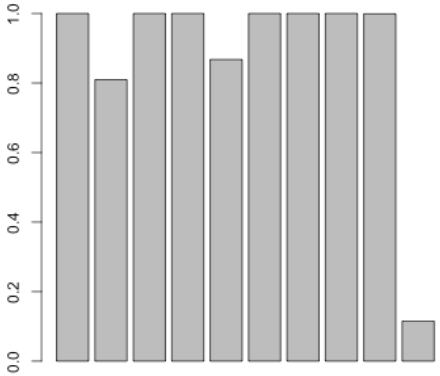
Replace:
$\arg\min_\beta \sum_{i=1}^{n} (y_i - f_i(\beta))^2$
with:
$\arg\min_\beta \sum_{i=1}^{n} w_i(\beta)(y_i - f_i(\beta))^2$

Robust regression – mechanics of iteratively reweighted least squares

Residuals



Weights

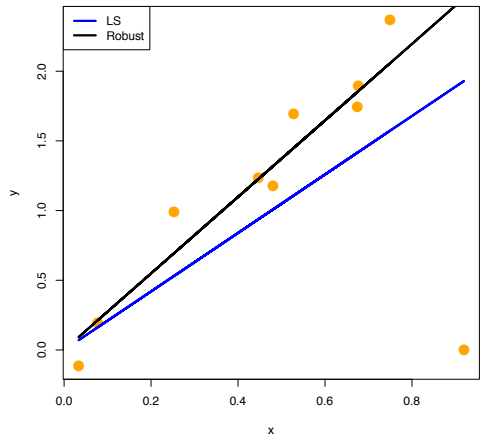


Sketch of IRLS:

Calculate initial estimates of parameters

Repeat the following until very little change:

- Calculate residuals
- From standardized residuals, calculate weight and downweight “extreme” observations
- Re-estimate parameters



More details – weight functions (as function of standardized residuals)