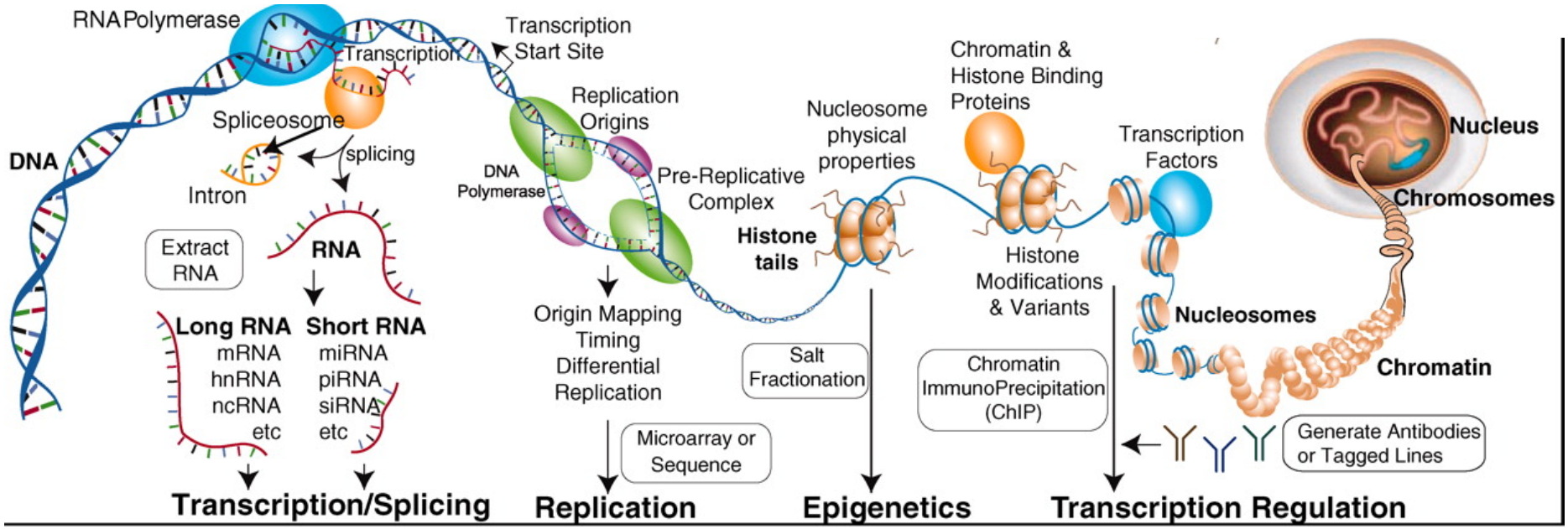
University of Zurich[UZH]

Institute of Molecular Life Sciences

Some loose ends: ChIP-seq + HMMs + segmentation + integrative/exploratory analyses

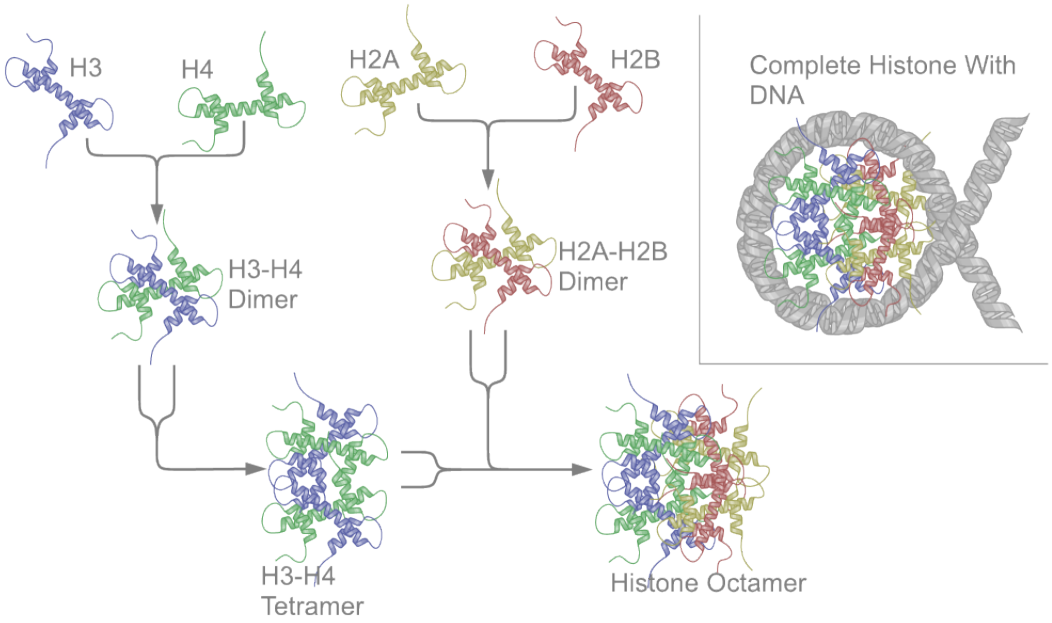
Mark D. Robinson, Institute of Molecular Life Sciences

Various other epigenetic (and regulatory) factors



Roy et al. *Science* 2010

Histone variants and post-translation modifications



Two of each of H2A, H2B, H3 and H4 form a "nucleosome", which 147bp of DNA can wrap around

Histone variants and post-translation modifications

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]				

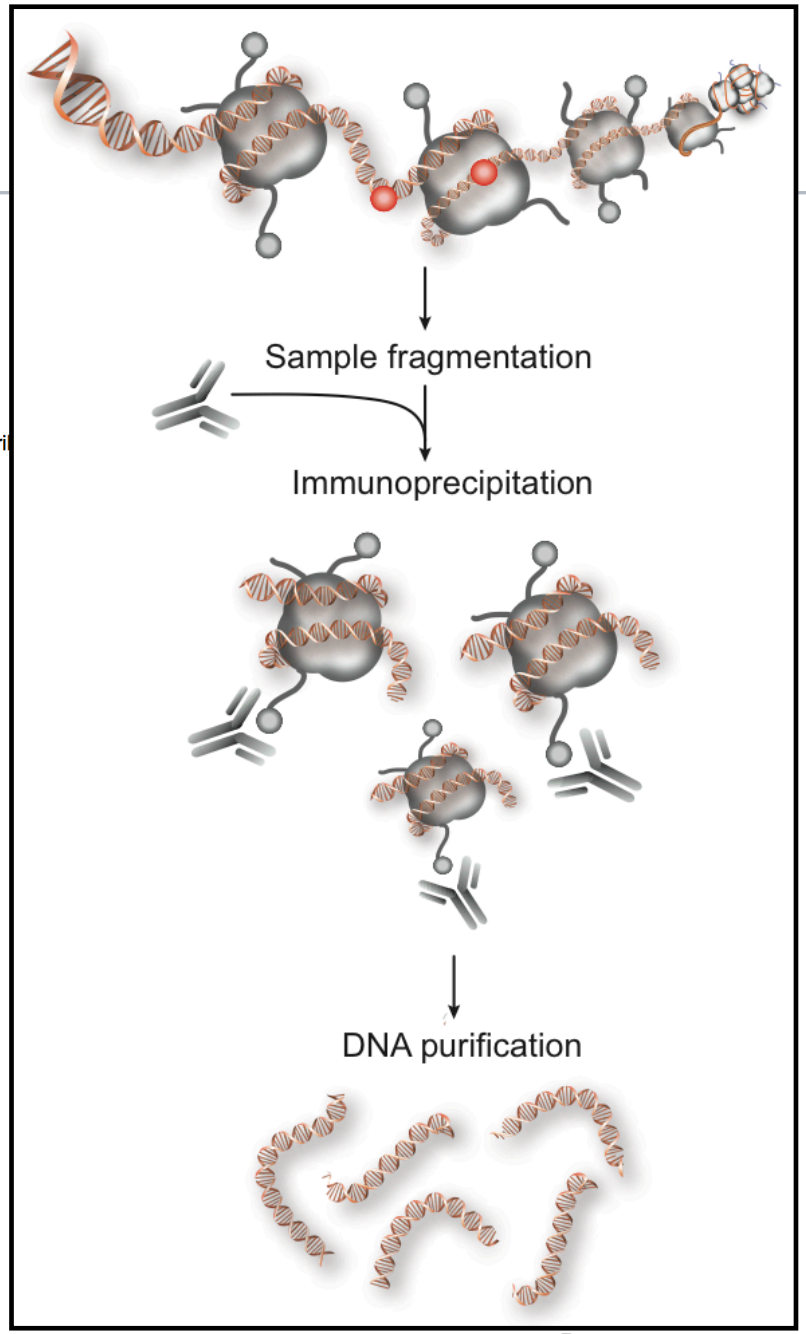
- H3K4me3 is found in actively transcribed promoters, particularly just after the transcription start site.
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively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

Chromatin immunoprecipitation for protein-DNA interactions

A very basic summary of the histone code for gene expression status is given below (histone nomenclature is descri

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.



Pipelines: sequencing reads to data analysis

Many sequencing experiments have some common initial preprocessing elements (e.g. read mapping); microarray experiments – normalization.

Downstream informatic analyses are catered to the scientific question.

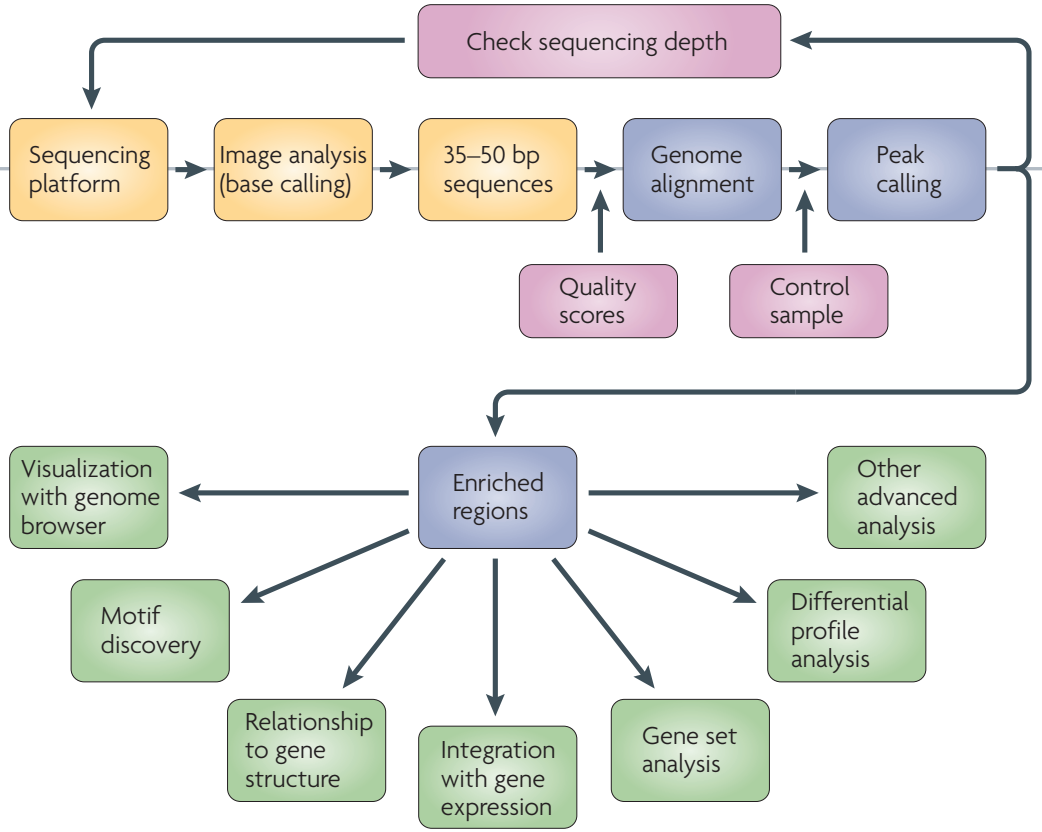


Figure 4 | **Overview of ChIP–seq analysis.** The raw data for chromatin immunoprecipitation followed by sequencing (ChIP–seq) analysis are images from the next-generation sequencing platform (top left). A base caller converts the image data to sequence tags, which are then aligned to the genome. On some platforms, they are aligned with the aid of quality scores that indicate the reliability of each base call. Peak calling, using data from the ChIP profile and a control profile (which is usually created from input DNA), generates a list of enriched regions that are ordered by false discovery rate as a statistical measure. Subsequently, the profiles of enriched regions are viewed with a browser and various advanced analyses are performed.

Nature Reviews Genetics 10, 669-680 (October 2009)

Region/peak finding depends on the epigenetic mark

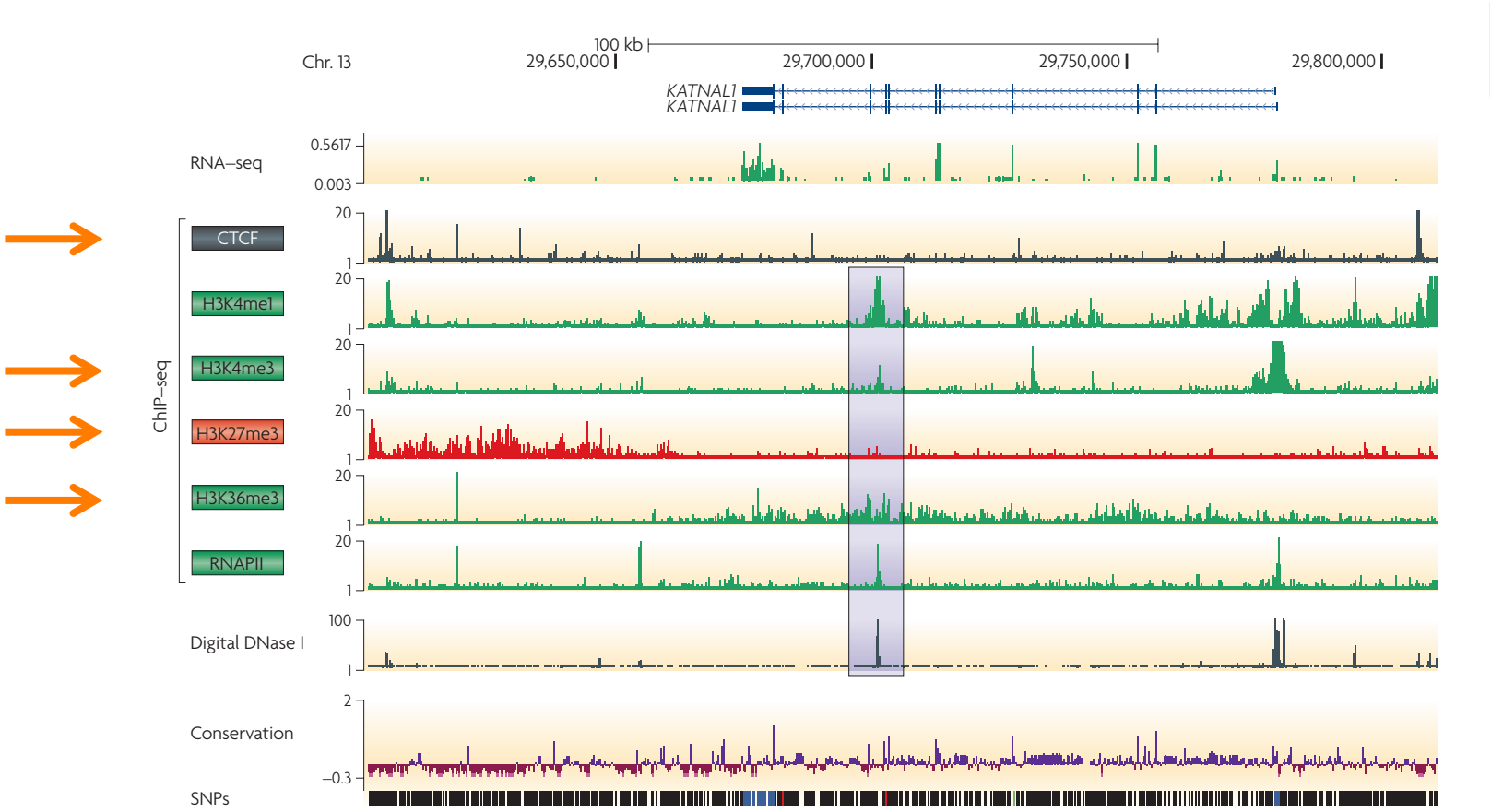
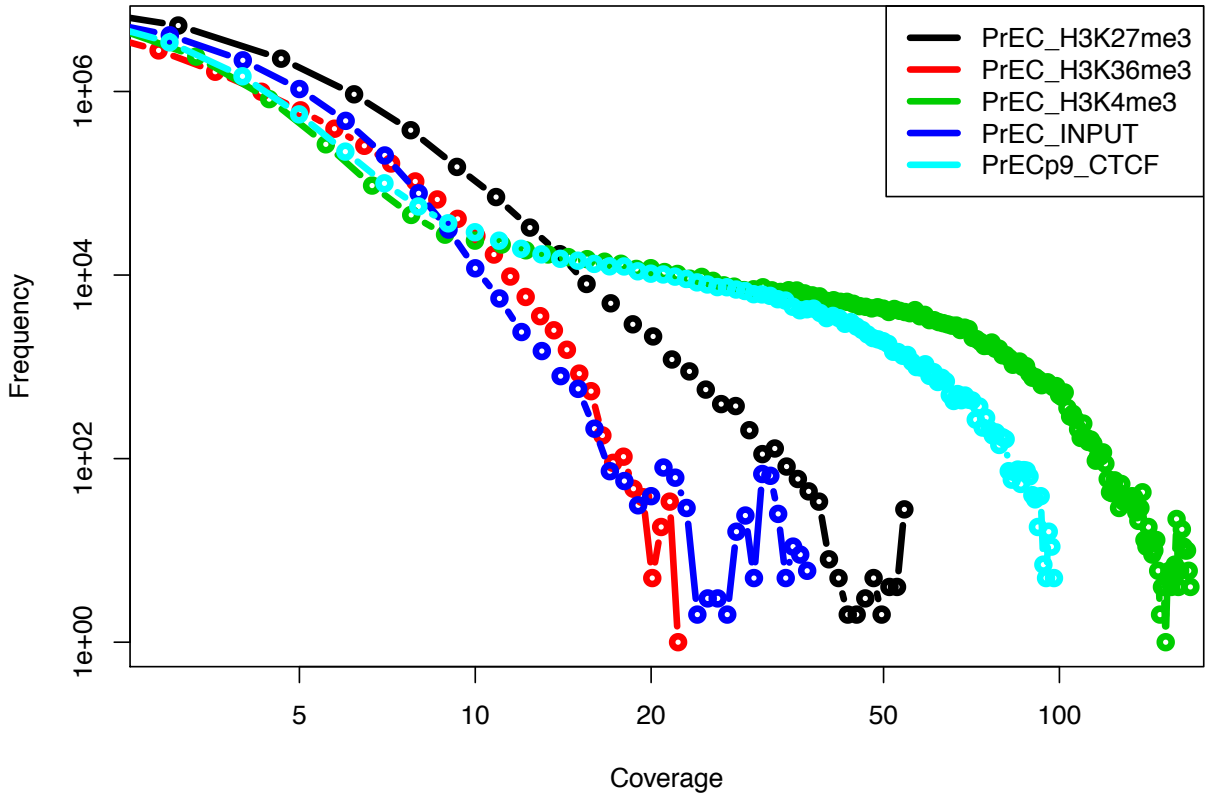


Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing

Quality checks – for ChIP-seq data

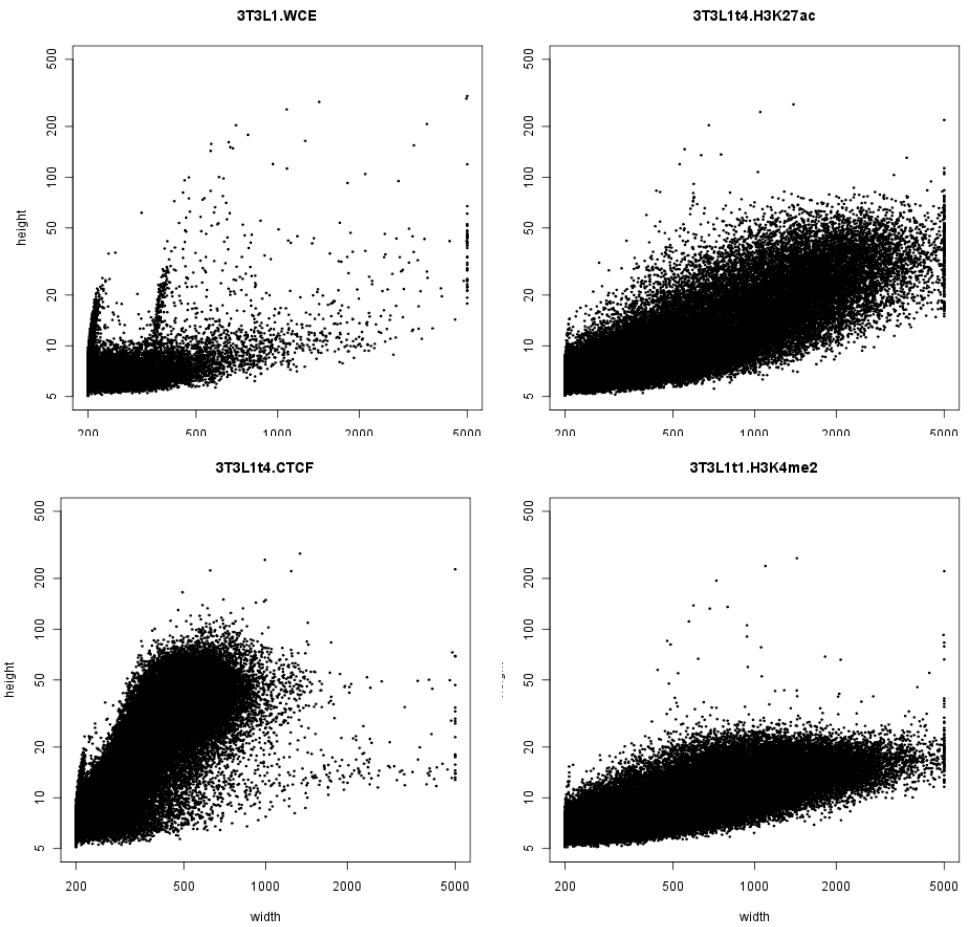


Heights/widths of CHERs (Hubert's lecture)

Heights and Widths of CHERs

The heights and widths of enriched regions strongly depends on the genomic mark that is tested.

Peaks with high coverage and lengths being a multiple of the fragment lengths represent artifacts.



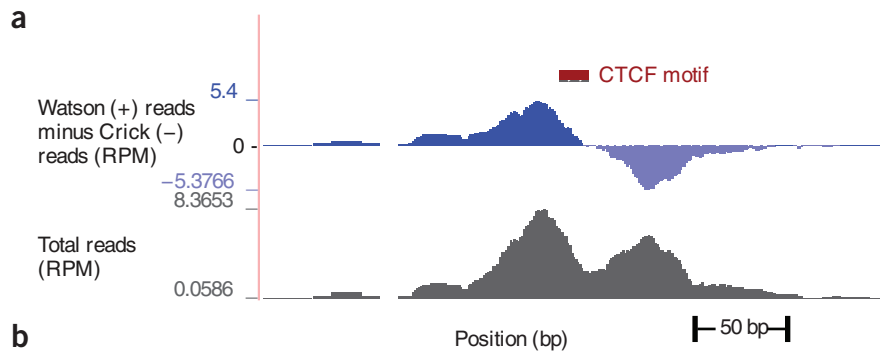
Each dot here is a "peak".

Requires a peak/ region finder to be run.

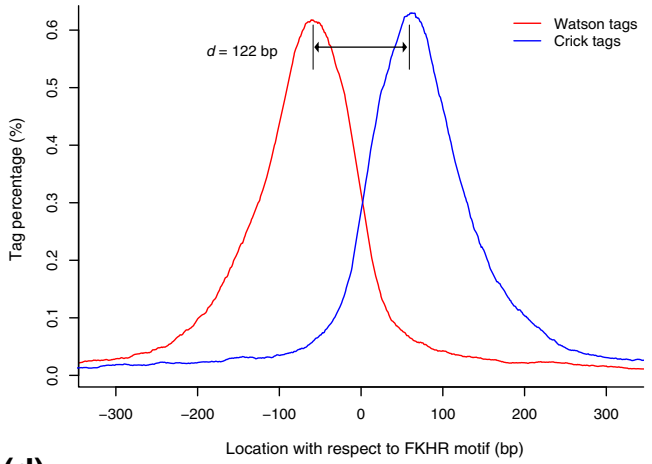
Peak/region detection for ChIP-seq data

MACS: model-based analysis of ChIP-seq data

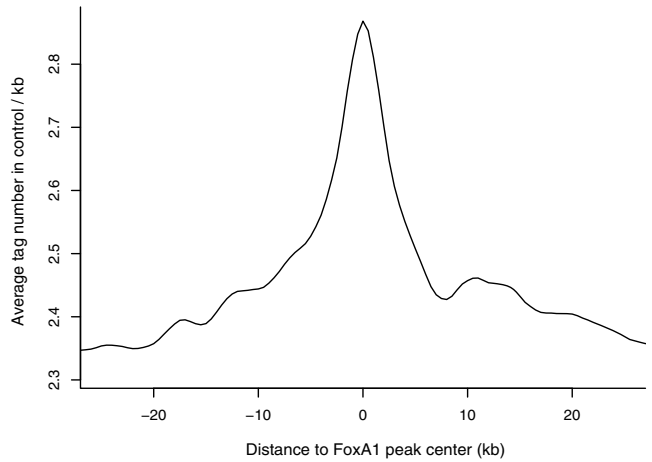
Accounting for strandedness of the reads



(b)



(d)

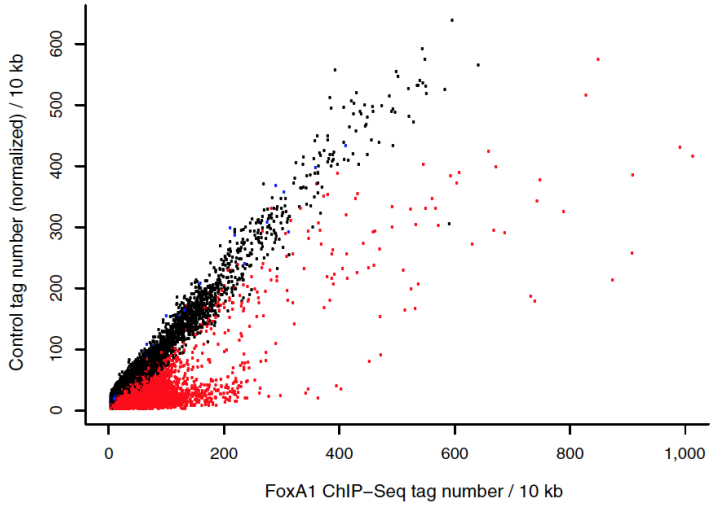


(f)

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



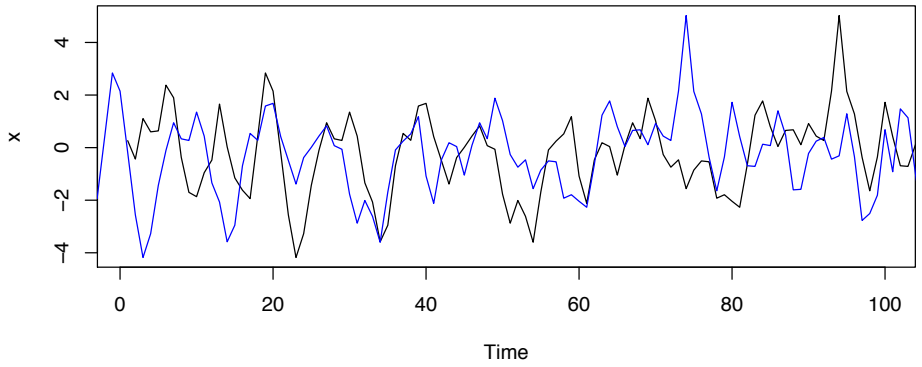
$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$$

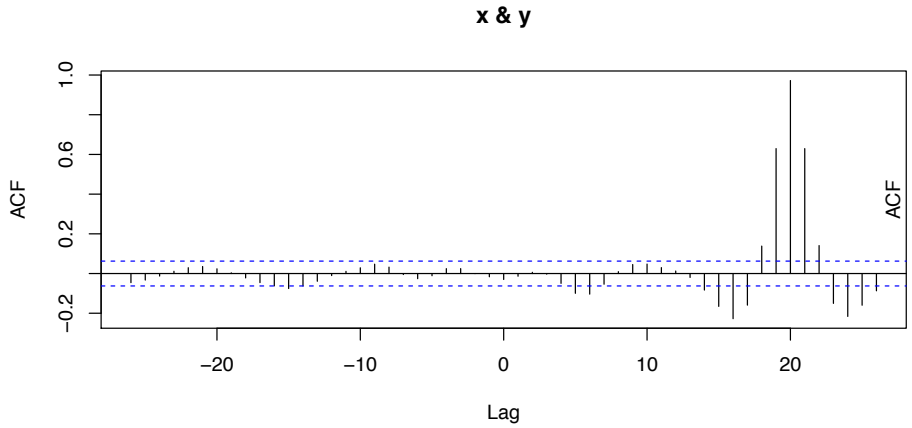
For a ChIP-Seq experiment with controls, MACS empirically estimates the false discovery rate (FDR) for each detected peak using the same procedure employed in the previous ChIP-chip peak finders MAT [13] and MA2C [14]. At each *p*-value, MACS uses the same parameters to find ChIP peaks over control and 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

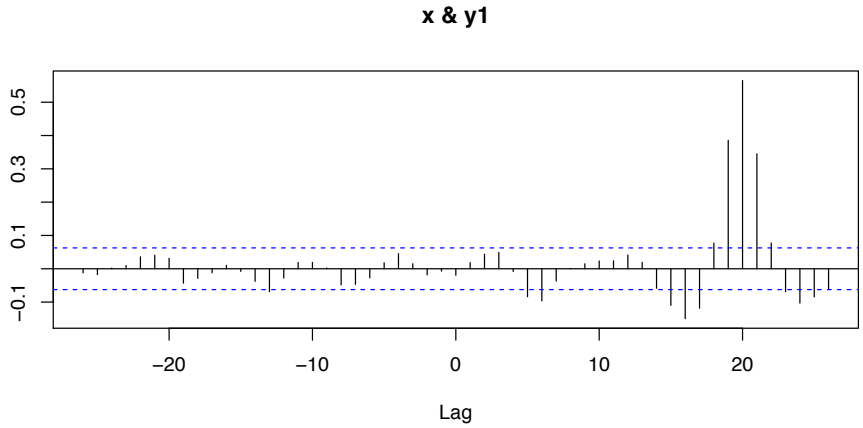
$$(f \star g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f^*[m]\ g[n+m].$$

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)
```

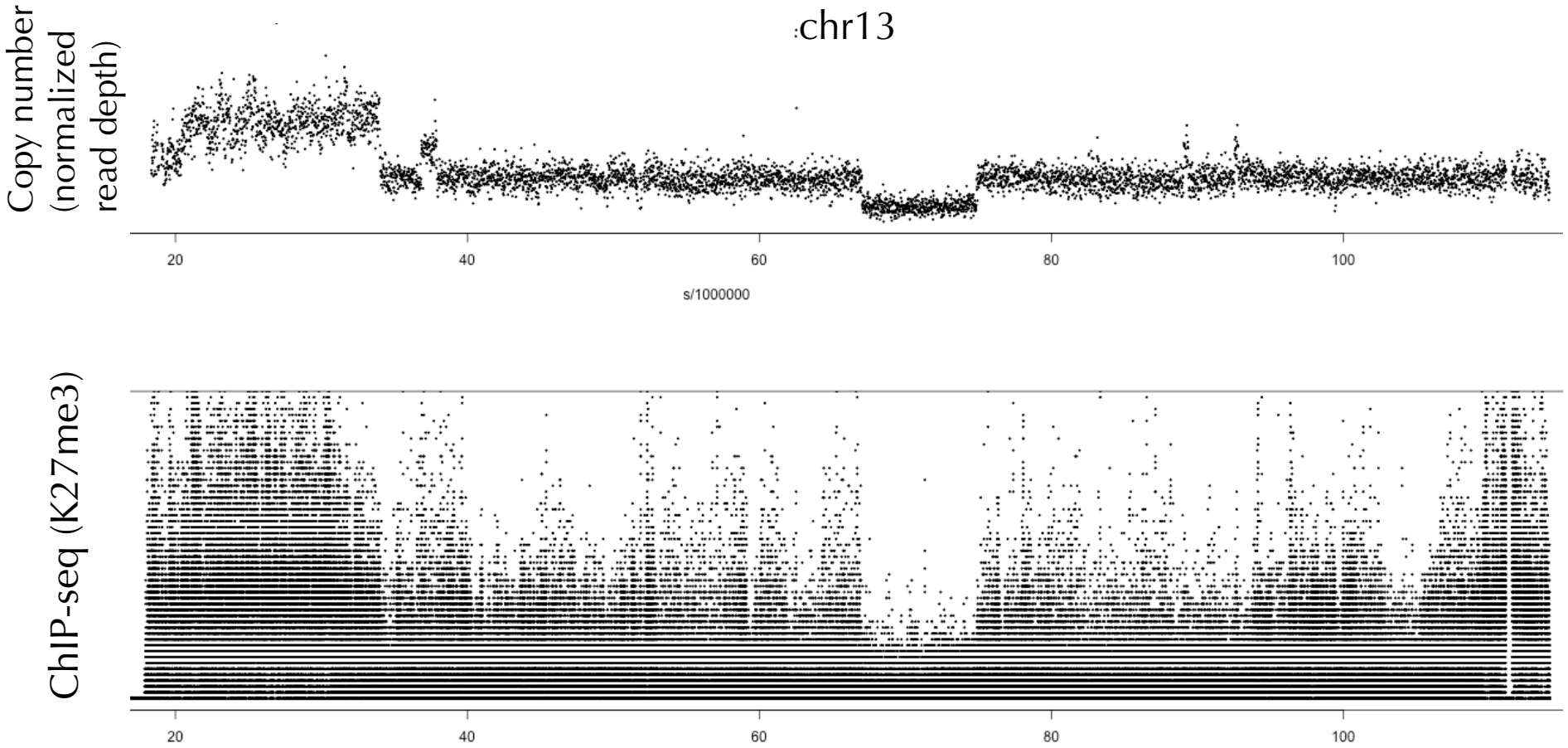




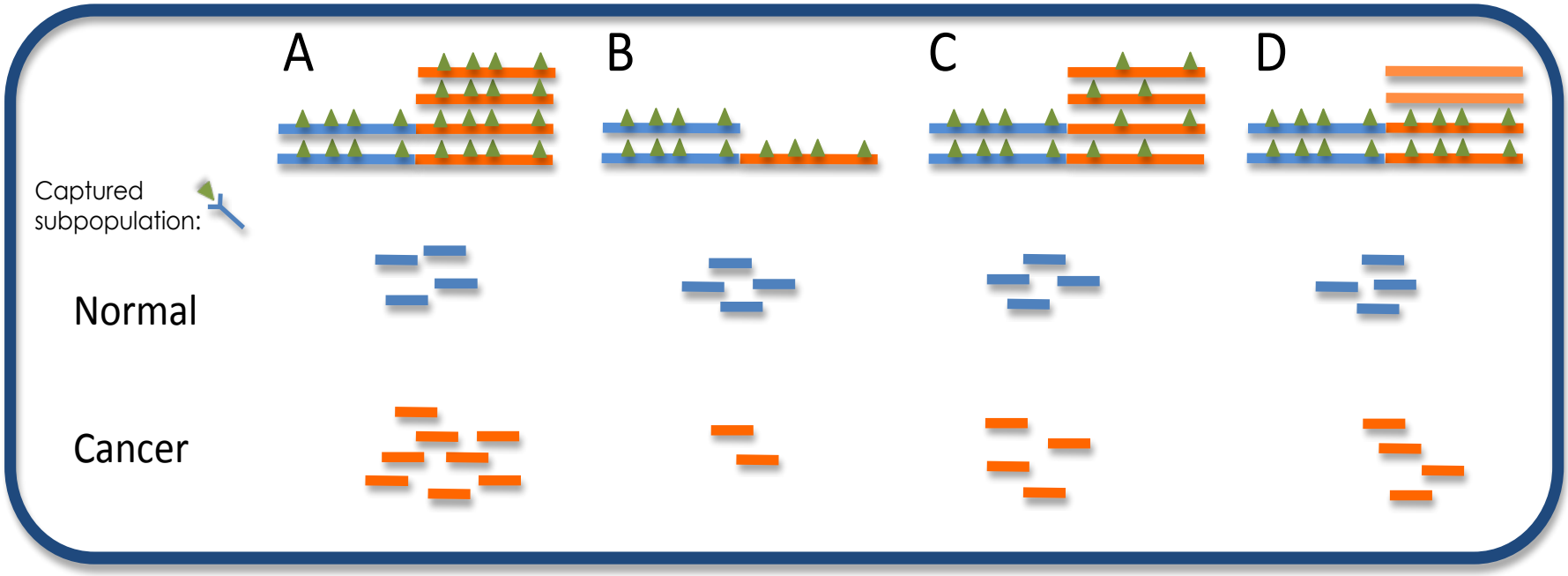


quantitative DNA-seq signal

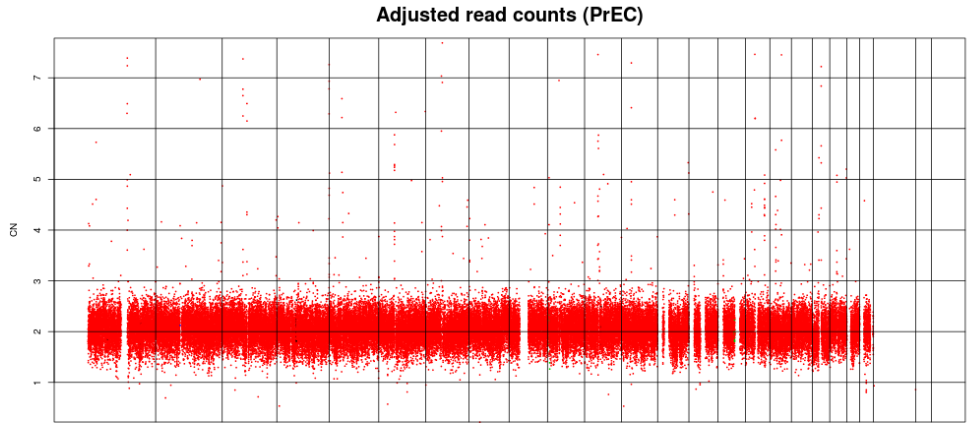
= biology (copy number, enrichment) + technical effects



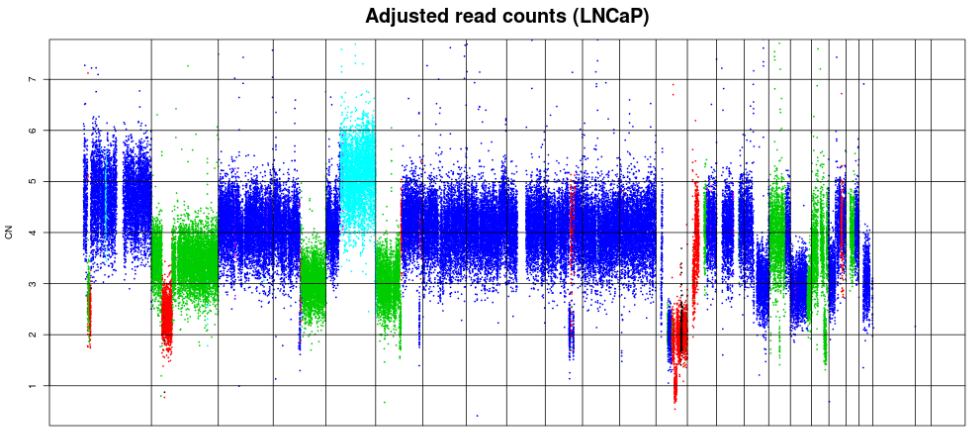
CNV affects differential comparisons: various scenarios



Case study: comparing epigenomes where differences in genomes exist.



PrEC (prostate epithelial cells) Normal copy number



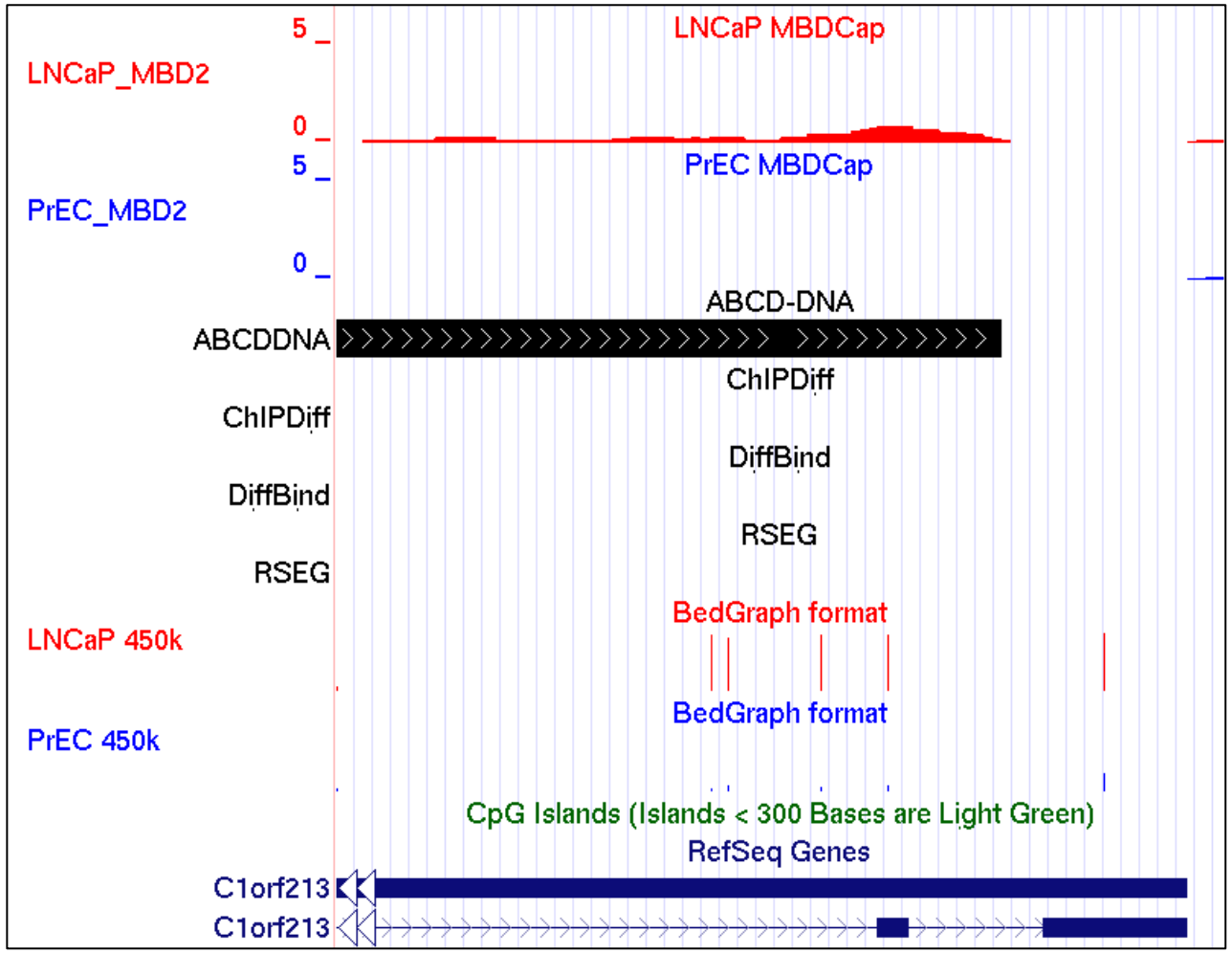
LNCaP (prostate cancer cells) primarily 4 copies

Read depths of low coverage sequencing, coloured by CNV calls by Affymetrix SNP 6.0

University of Zurich[UZH]

Institute of Molecular Life Sciences

This region has **2 copies** in normal PrEC cells and **2 copies** in prostate cancer LNCaP cells (We normalize LNCaP=4 to PrEC=2, so this is effectively a net ***loss*** of copy number)



Existing tools: False *negative* due to CN "loss"

"Truth" from Reference dataset

Robinson et al. 2012 Genome Research

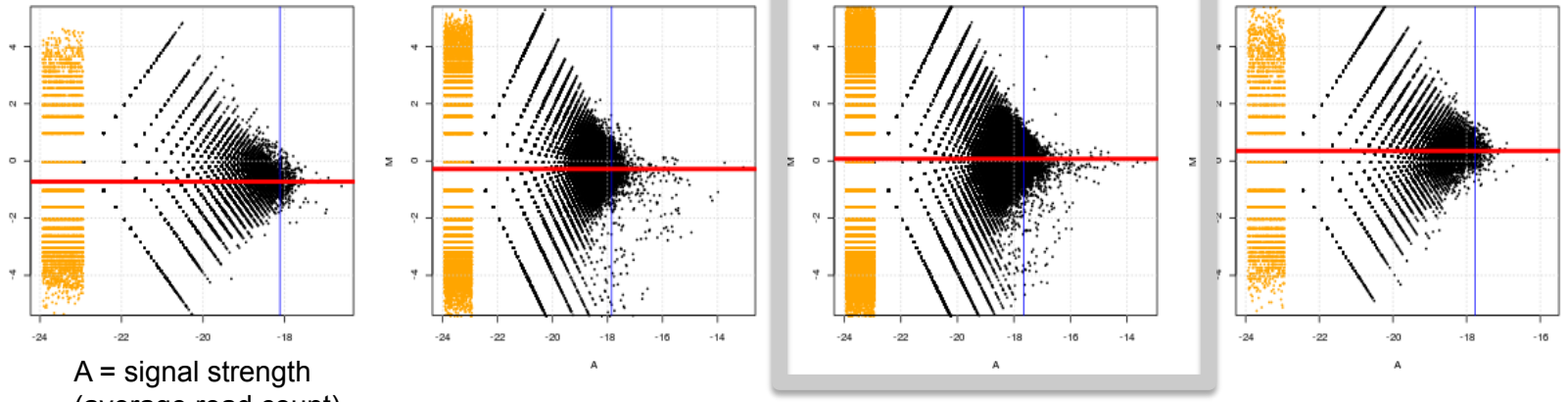
Copy number can be represented as a “Normalization” problem

Copies in cancer (2 in normal):

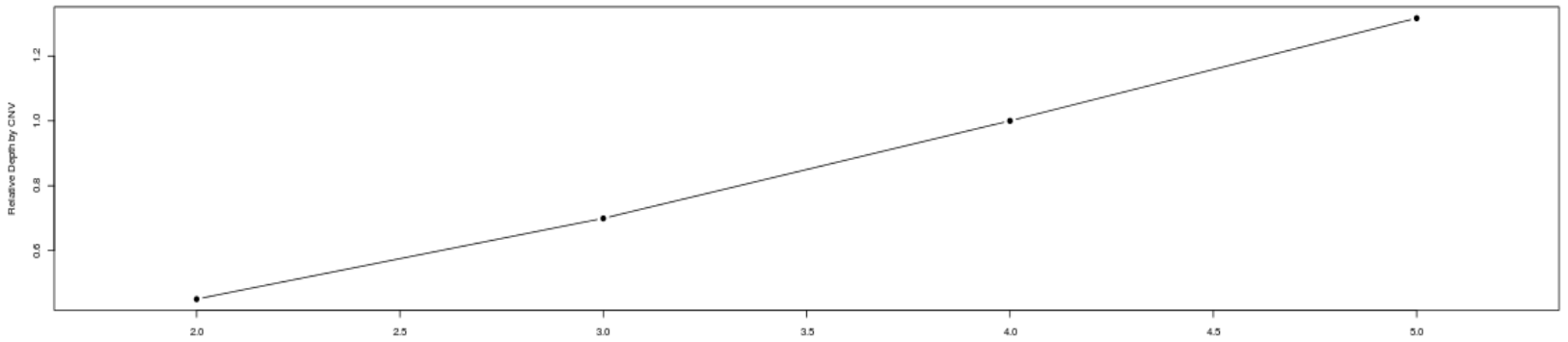
Neutral

2 3 4 5

M = log-fold-change **Cancer / Normal**



A = signal strength (average read count)



Statistical details of ABCD-DNA

We model read densities, $Y_{ij,}$ in a generalized linear model:

$$\log(E[Y_{ij}]) = O_{ij} + B_i X$$

O_{ij} is an r x n matrix of **offsets**

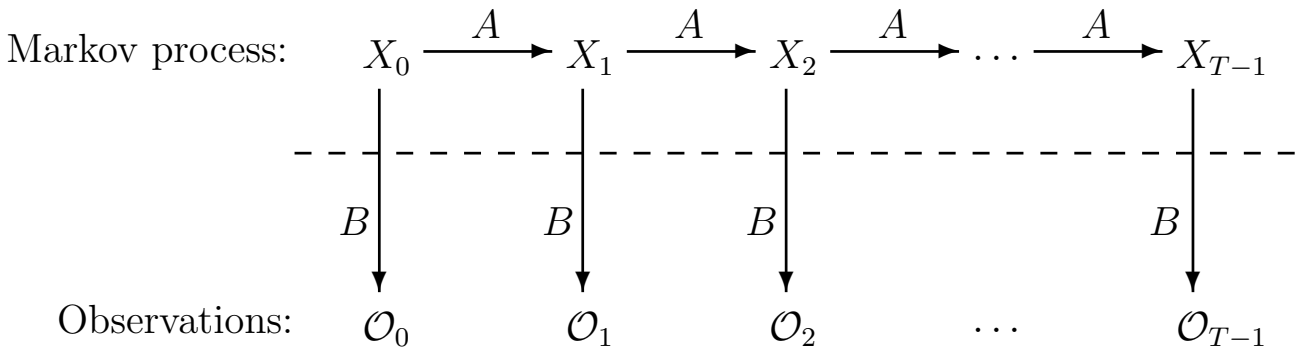
X is an k x n **design matrix**

B_i is a r x k matrix of region-specific **coefficients**

O_{ij} can be decomposed into $\log(CN_{ij}) + \log(1\ D_j)$

Using independent data to estimate offsets

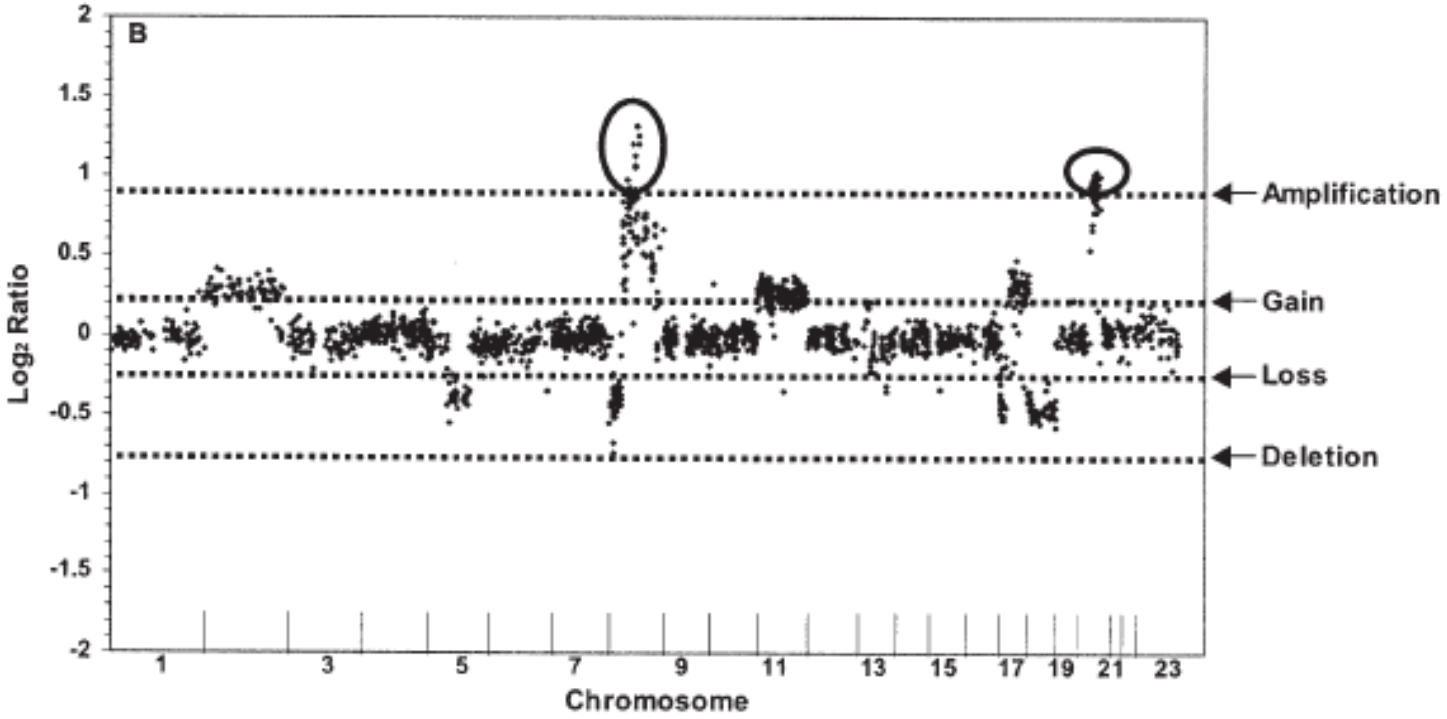
Introduction to Hidden Markov Models



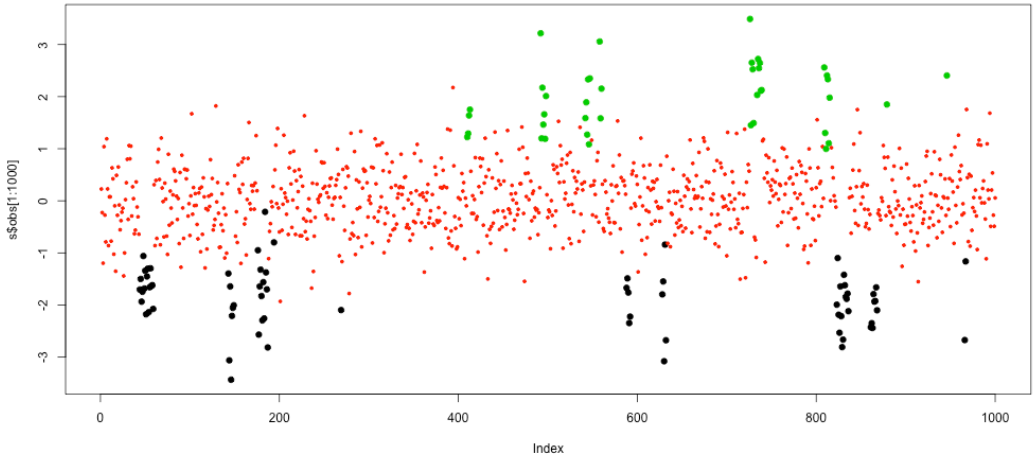
X_i – hidden ("latent", unobserved state)
O_i – "emitted" observation

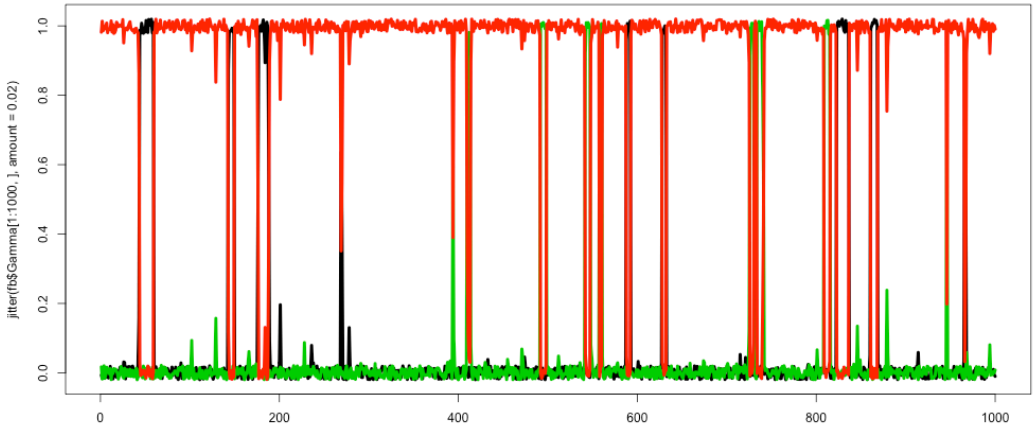
A – transition probabilities
B – emission probabilities/distributions

Examples: HMMs in genomics – copy number



A “vanilla” HMM: Normal emission distributions





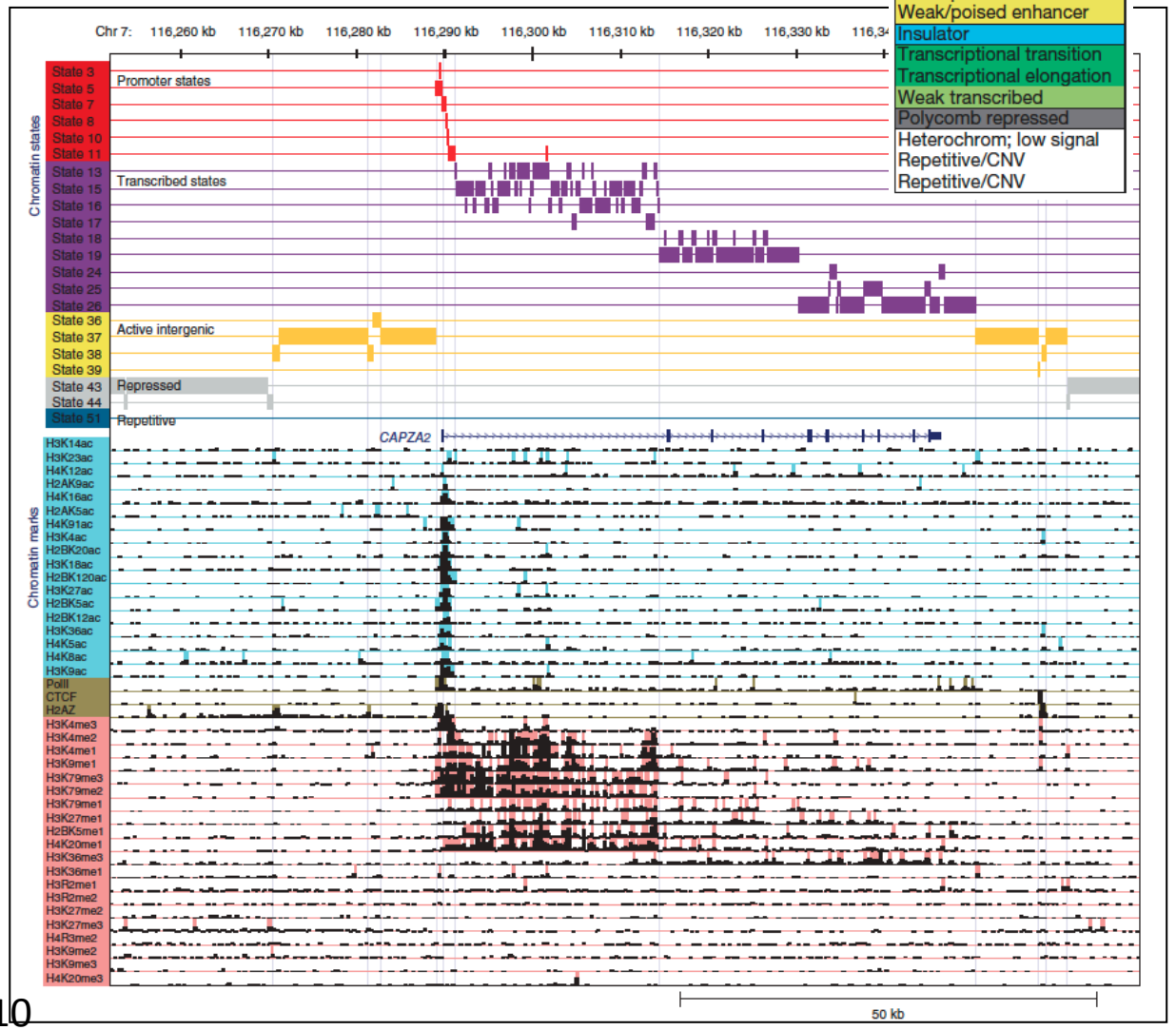

```
library(RHmm)
h <- distributionSet("NORMAL", mean=c(-2, 0, 2),
                                    var=c(.4, .4, .4))
ip <- c(0,1,0)
tr <- rbind(c(.8,.2,0), c(.01, 0.98, .01), c(0,.2,.8)

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

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

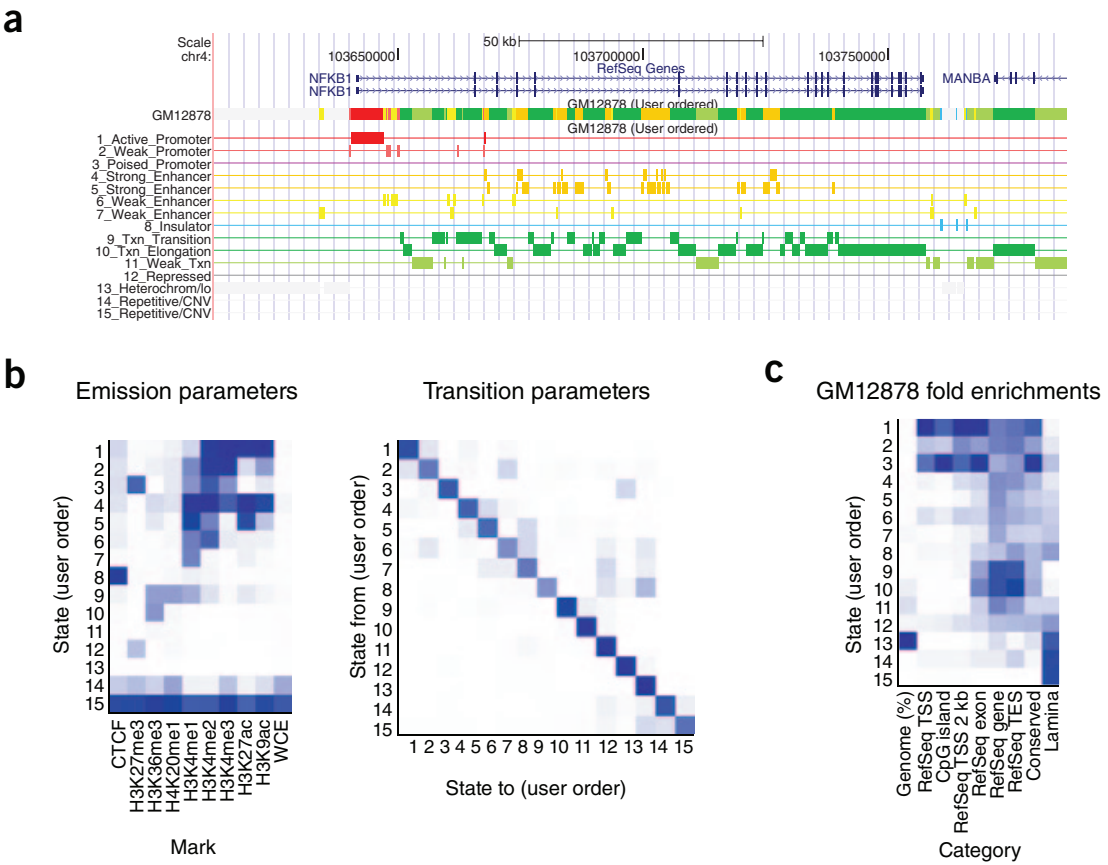
Multivariate HMM is trained; genome is partitioned into 15 states



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:

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

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

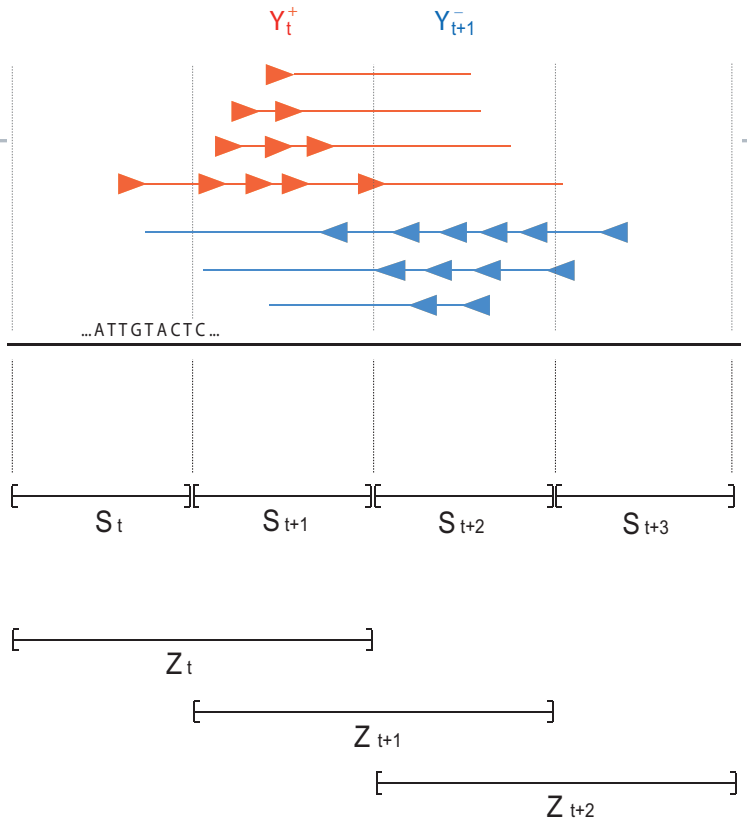


Figure 1

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

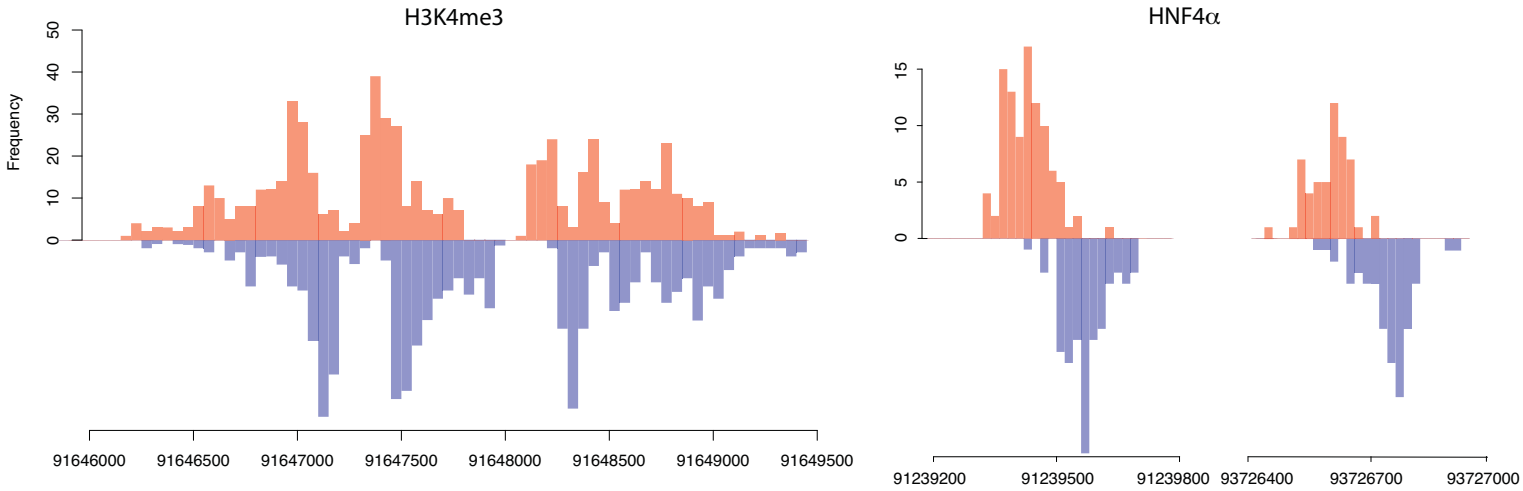
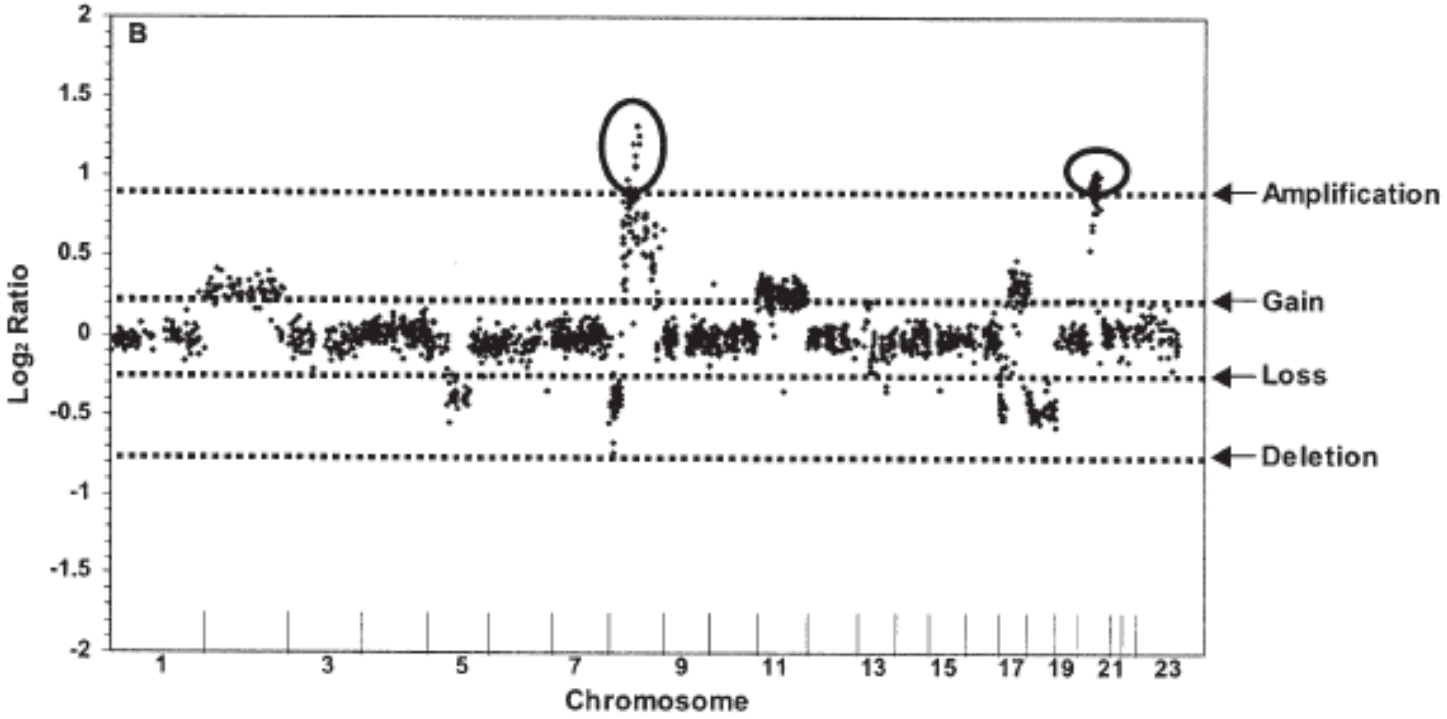


Figure 3
A closer view of some HeK4me3 and HNF4α peaks. These histograms present the counts of the 5' ends of the reads from the H3K4me3 and the HNF4α data, forming peaks on the forward (red) and reverse (blue) strand. The offset between them shows how the enclosed area corresponds to an enriched region. 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.

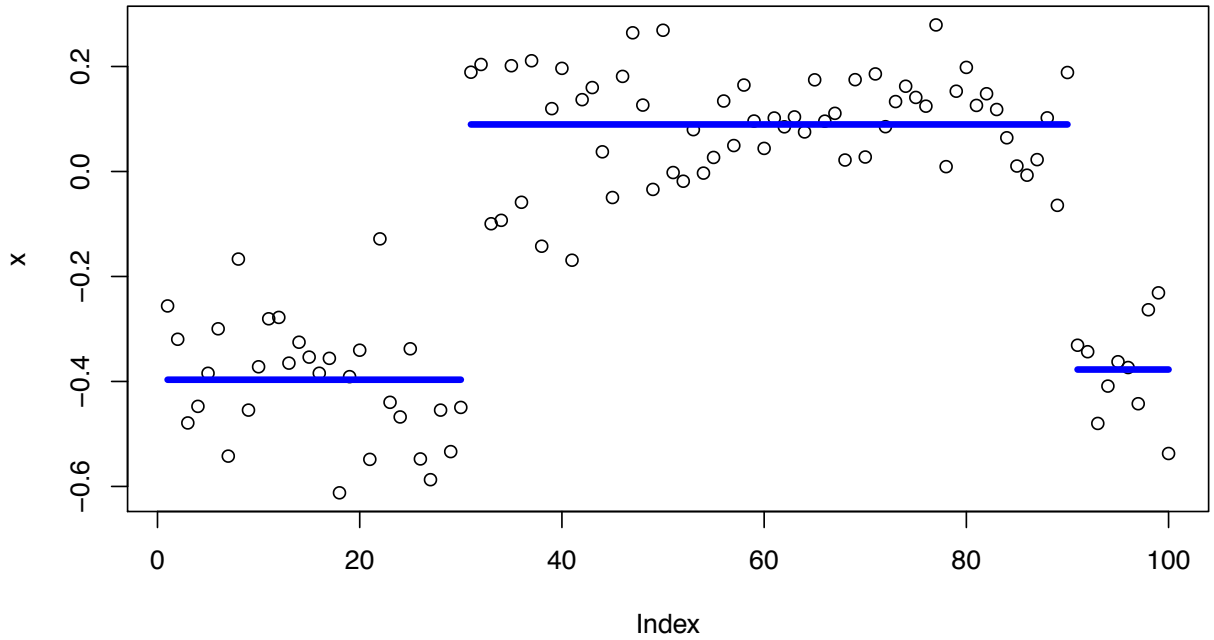
Copy number: HMM → segmentation



Example segmentation

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

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



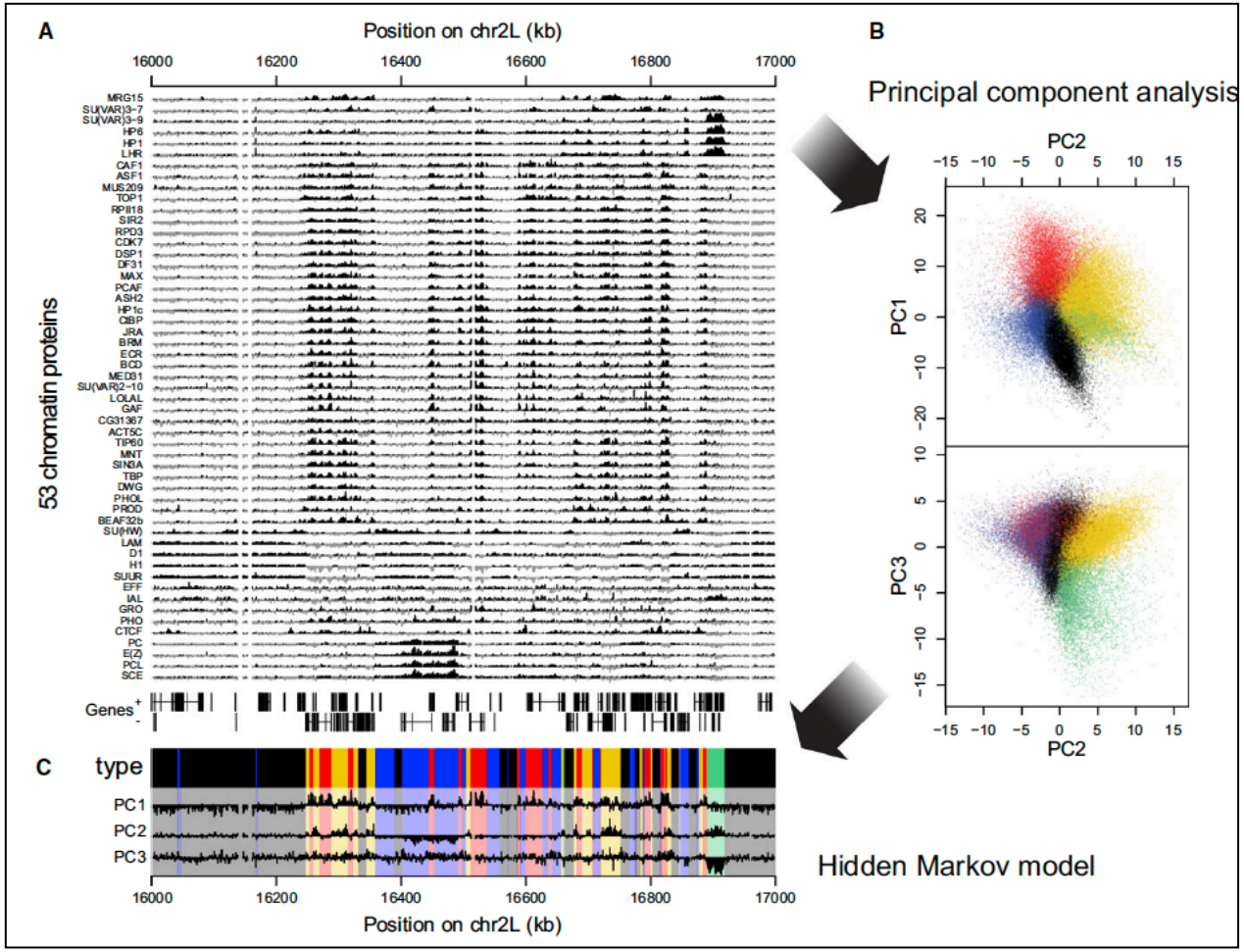
Exploratory analyses

53 chromatin factors (ChIP-seq)

Compression to 3 principal components

Learn HMM

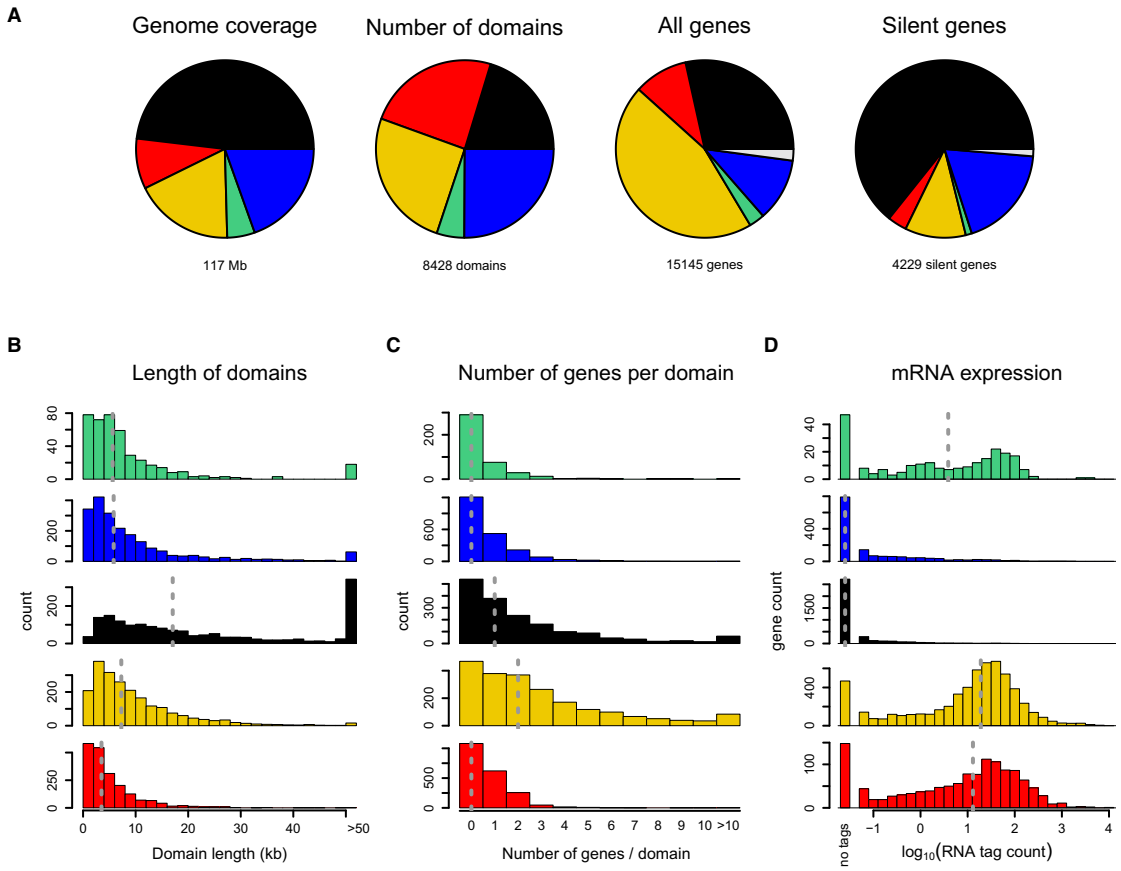
Every region of the genome partitioned into 5 “states” (here, assigned a colour)



Filion et al, Cell 2010

Exploratory analyses

"Colours" are reflective of various features



Filion et al, Cell 2010

Functional category analysis, gene set testing

OPEN ACCESS Freely available online

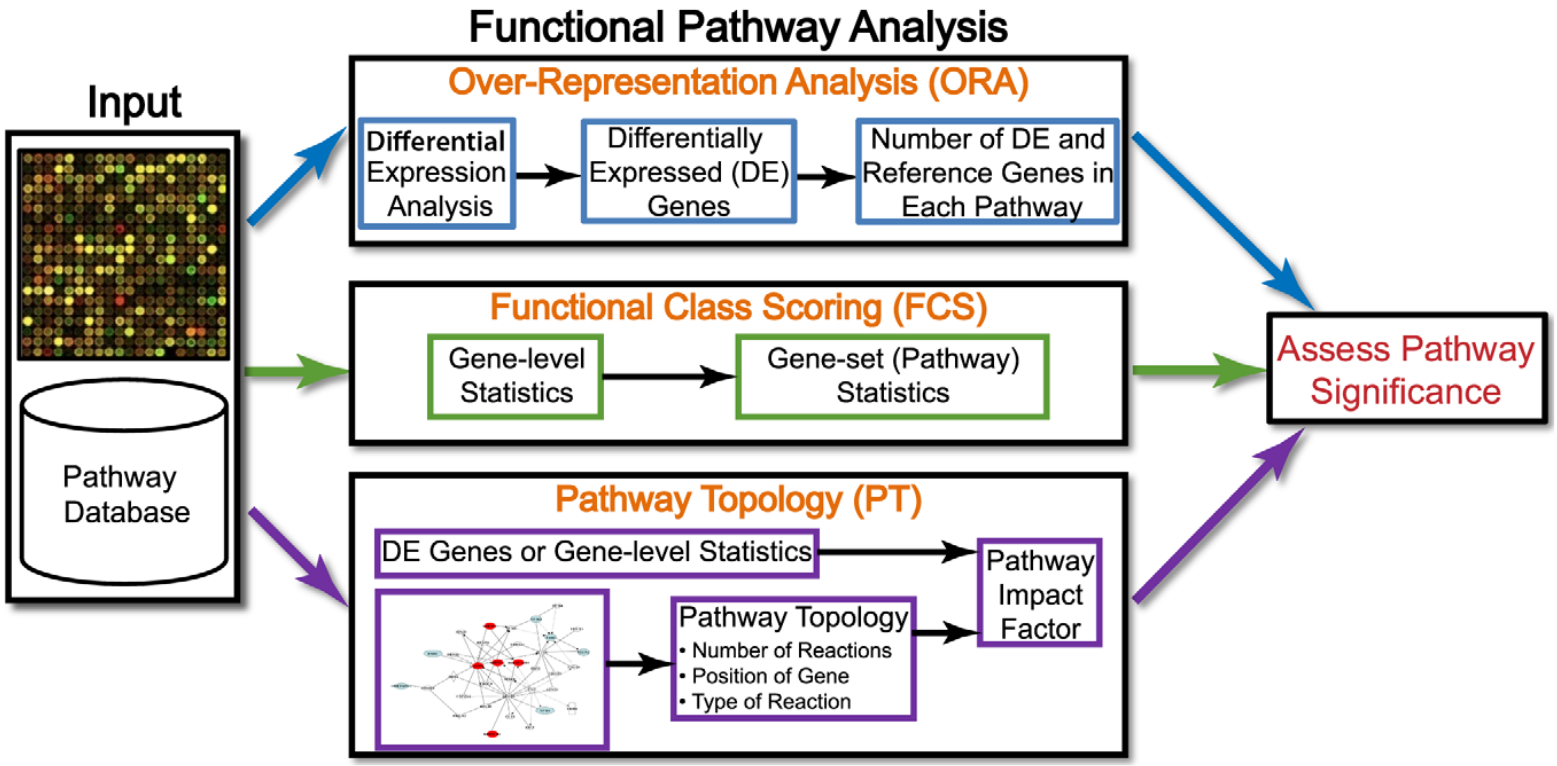
PLoS COMPUTATIONAL BIOLOGY

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, Stanford, California, United States of America, **2** Lucile Packard 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 significant, even if individual genes are not.

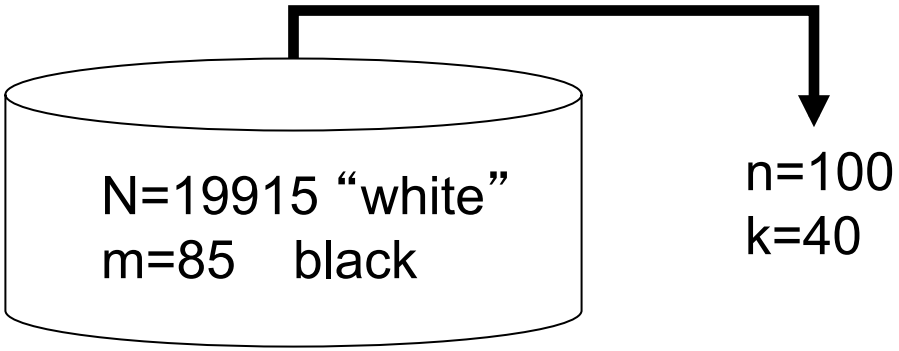
Starting point:	threshod, set of DE genes	gene-level statistics
Tool examples:	DAVID [C] **goseq** [C]	GSEA [S] **roast** [S] **CAMERA** [C]

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

Length bias in RNA-seq

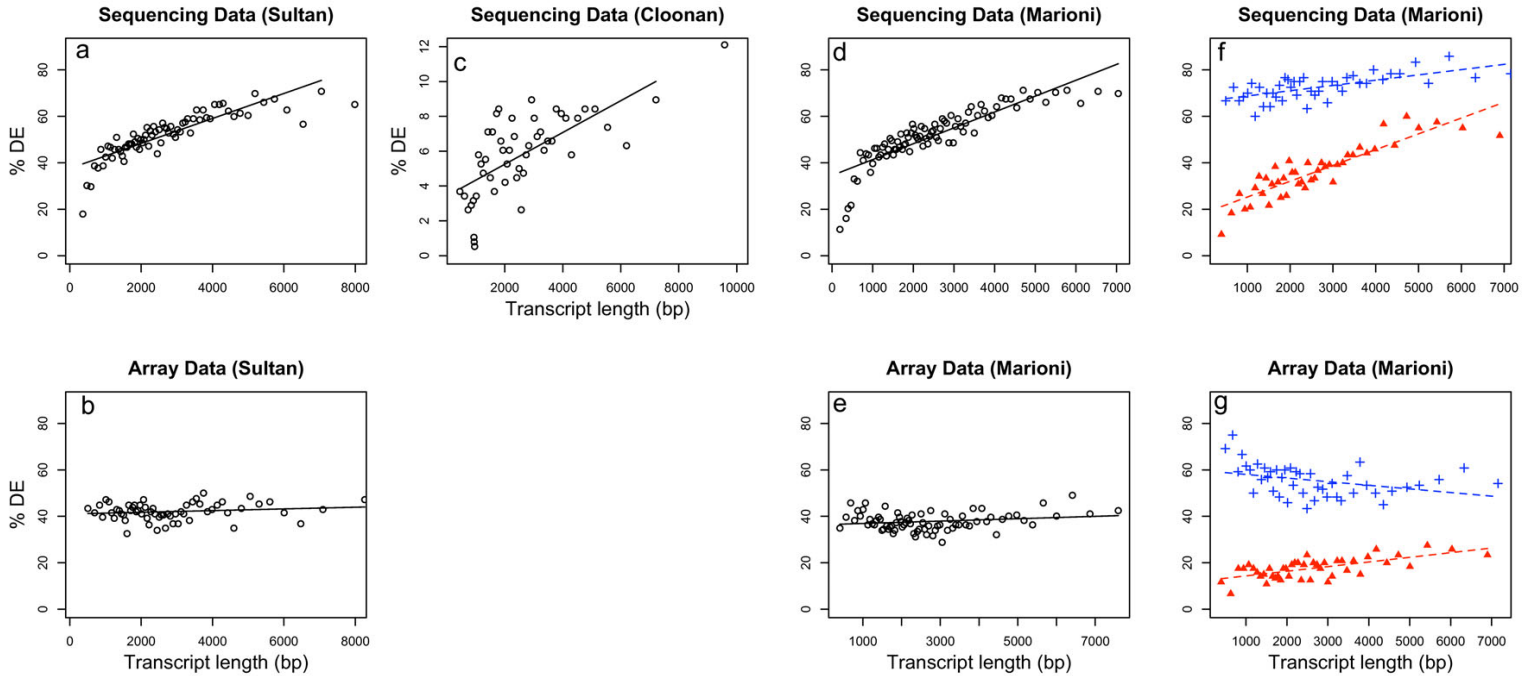
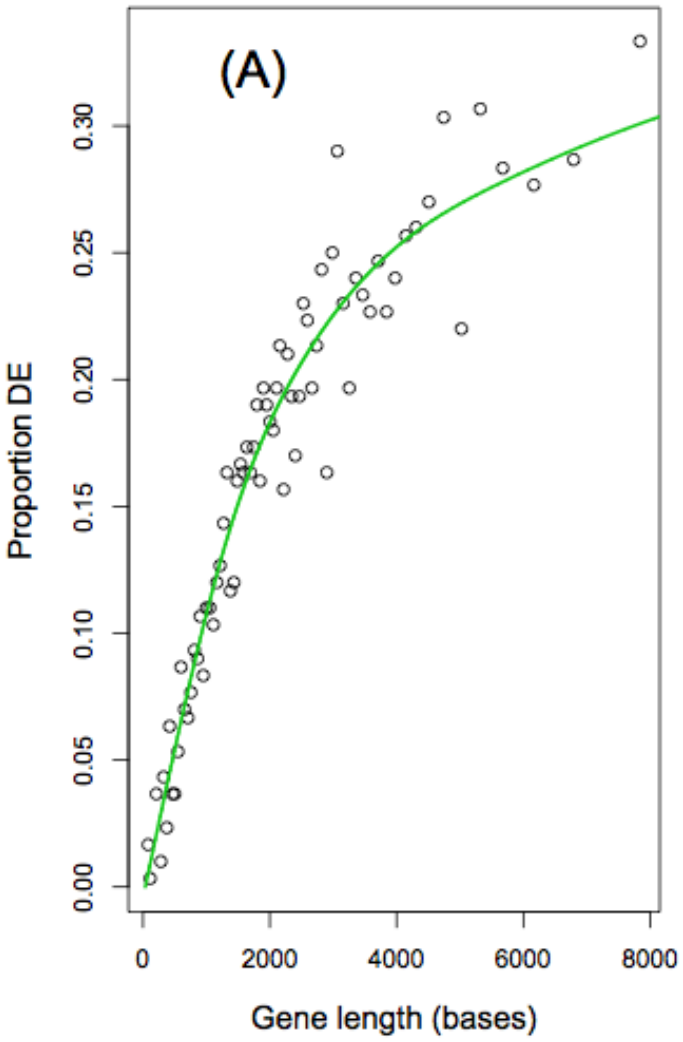
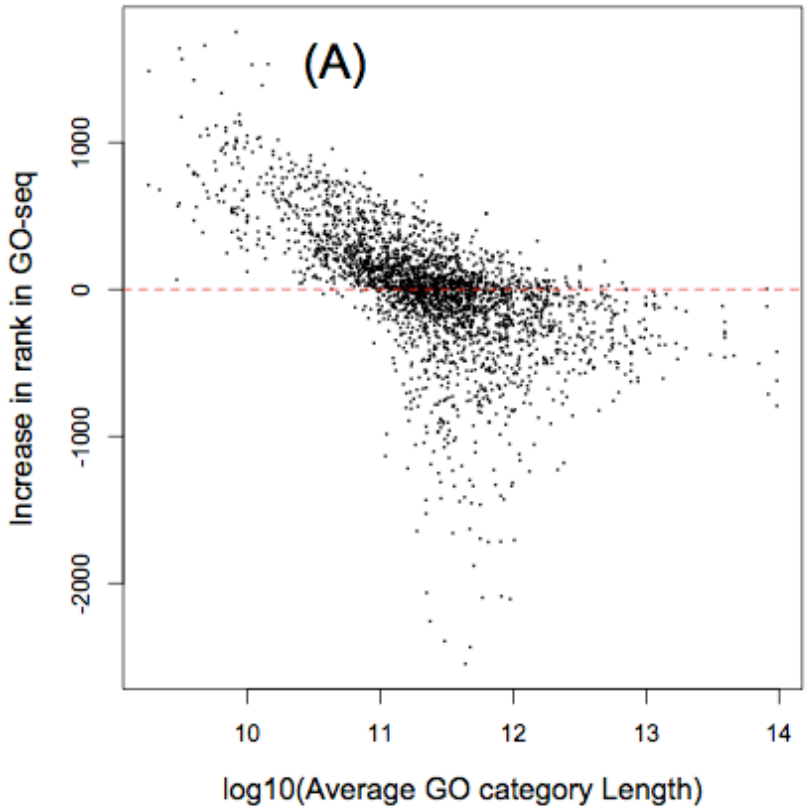


Figure 1
Differential expression as a function of transcript length. The data is binned according to transcript length and the percentage of transcripts called differentially expressed using a statistical cut-off is plotted (points). A linear regression is also plotted (lines). **a** – **e** use all the data from RNA-seq and the microarrays from studies [4-6] respectively. **f** and **g** plot 33% of genes with highest expression levels (blue crosses) and 33% of genes with low expression (red triangles) taken from the microarray data for genes which appear on both platforms in [6]. The regression gives a significant trend for the percent of differential expression with transcript length for **a**, **c**, **d** and **f** and the lowly expressed genes in **g**. Note that this figure illustrates common data features between disparate experiments and is not a comparison between platforms, methods or experiments.

University of Zurich[UZH]
Institute of Molecular Life S

Bioconductor package goseq: Adjusting for gene length bias



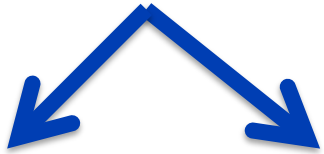


Categories with short genes get a higher rank in GOseq

Categories with long genes get a lower rank

Young, Wakefield, Smyth and Oshlack, Genome Biology, 2010

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

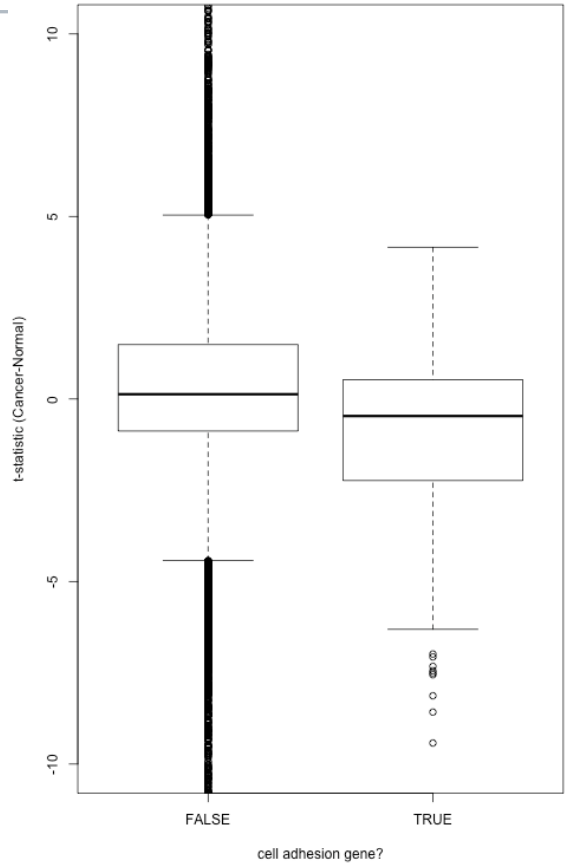


competitive	self-contained
Genes in the set tend to be more strongly DE than randomly chosen genes	*At least some genes in the set are truly DE*

Goeman & Bühlmann, *Bioinformatics* 2007

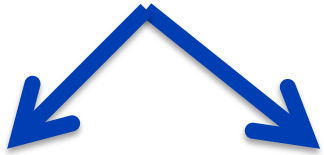
Viewing gene sets

Cell adhesion genes



Genes regulated by MYB

Gene set analysis. If permutation p-value, permute genes or permute samples?



competitive — gene

self-contained — sample

Goeman & Bühlmann, *Bioinformatics* 2007

Gene set enrichment analysis (GSEA)

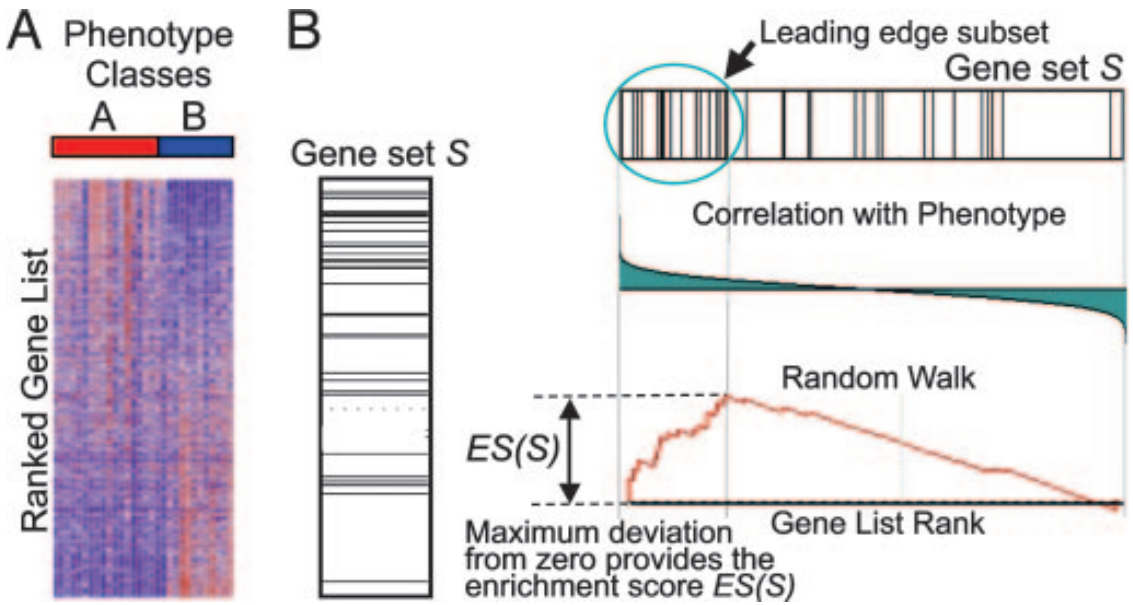


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the ''gene tags,'' i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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

Self-contained.

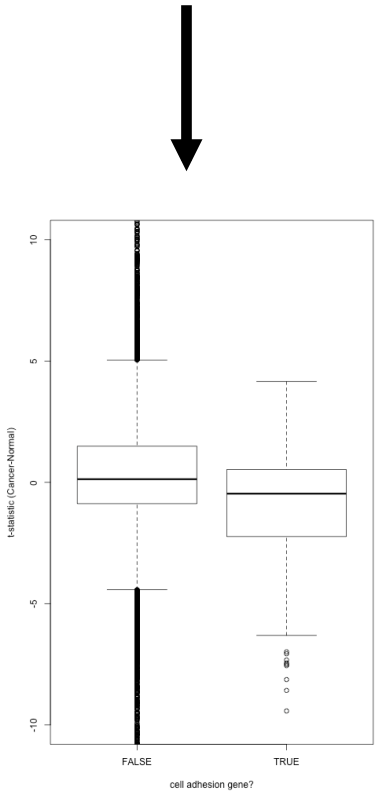
Permutation P-value: Sample permutation is done, which preserves gene correlation.

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

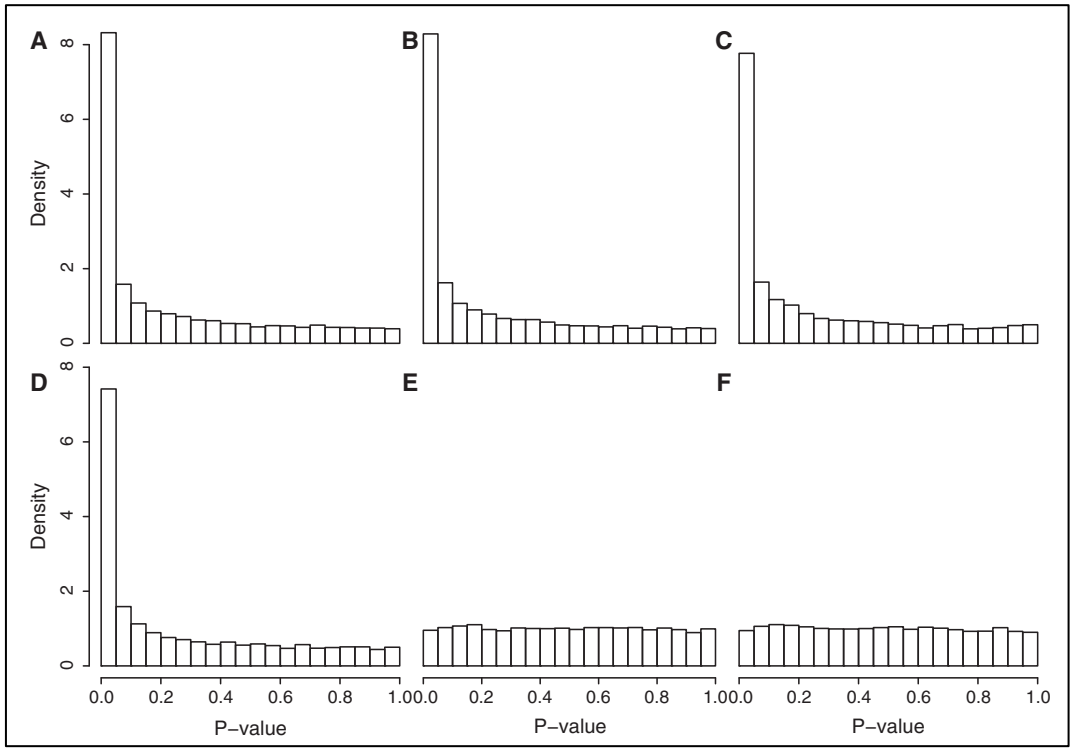
Now switches to a gene-based permutation (competitive) in small samples.

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

University of Zurich[UZH]

Institute of Molecular Life Sciences

GSEA, "Simpler EA"

Statistical Methods in Medical Research 2009; **18**: 565–575

Gene set enrichment analysis made simple

Rafael A Irizarry Department of Biostatistics, Johns Hopkins School of Public Health, 615 N. Wolfe St. E3620, Baltimore, MD 21205, USA, **Chi Wang** Statistics Department, University of California, Riverside, 900 University Avenue, 2626 Statistics Computer Bldg. Riverside, CA 92521-0122, USA, **Yun Zhou** Gilead Sciences, Inc. 333 Lakeside Dr., Foster City, CA 94404, USA and **Terence P Speed** Department of Statistics, 367 Evans Hall, #3860, University of California, Berkeley, CA 94720-3860, USA

Proposed a simple average-of-t-statistics measure.

Defends GSEA, talks about effect of correlations.

Gene Set Enrichment Analysis Made Right

Pablo Tamayo[1], George Steinhardt[2], Arthur Liberzon[1], and Jill P. Mesirov [1,2]

1. The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA.
2. Boston University Bioinformatics Program. Boston University, Boston, MA 02215. USA.

How much of this is storytelling?

MBE Advance Access published July 17, 2012

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*.