Overview

- Bioconductor Project
- Fold changes
- Tiling Arrays for ChIP-chip
- Tiling Arrays for Transcript Profiling

Wolfgang Huber
EMBL/EBI

- Biology is becoming a computational science
- Problems of data analysis and mathematical modeling require computational support and computational solutions
- We put a premium on code reuse
  - many of the tasks have already been solved
  - if we use those solutions we can put effort into new research
- Data complexity is dealt with using well designed, self-describing data structures
- Reproducible research requires open access to computational code

The S language

- **The S language has been developed since the late 1970s by John Chambers and his colleagues at Bell Labs.**
- **The language has been through a number of major changes but has been relatively stable since the mid 1990s**
- **The language combines ideas from a variety of sources (e.g. *Awk*, *Lisp*, *APL*...) and provides an environment for quantitative computations and visualization.**

Implementations

- S-Plus a commercialization of the Bell Labs code.
- R an independent open source version that was originally developed at the University of Auckland but which is now
- developed by a world wide group of developers.
- Each version has advantages and problems.

References

- ***The New S Language*, *Statistical models in S*, *Programming with Data*, by John Chambers and various co-authors**
- ***Modern Applied Statistics*, *S Programming* by W. N. Venables and B. D. Ripley**
- ***Introductory Statistics with R* by P. Dalgaard**
- ***Data Analysis and Graphics Using R* by J. Maindonald and J. Braun.**

Packages

- Packages are the main unit of software authoring, versioning and distribution
- CRAN is the major repository for R packages. It is hosted by TU Vienna and ETH Zürich, and has many mirrors world-wide
- Bioconductor is a repository for biology related packages. It is hosted at the Fred Hutchinson Cancer Research Centre.

- an open source and open development software project for the analysis of biomedical and genomic data.
- was started in the autumn of 2001 and includes core developers in the US, Europe, and Australia.
- R and the R package system are used to design and distribute software.
- A goal of the project is to develop software modules that are integrated and which make use of available web services to provide comprehensive software solutions to relevant problems.

Goals of the Bioconductor project

- Provide access to powerful statistical and graphical methods for the analysis of genomic data.
- Facilitate the integration of biological metadata (e.g. Entrez, Ensembl, GO(A), PubMed) in the analysis of experimental data.
- Allow the rapid development of extensible, interoperable, and scalable software.
- Promote high-quality documentation and reproducible research.
- Provide training in computational and statistical methods.

Why are we Open Source

- so that you can find out what algorithm is being used, and how it is being used
- so that you can modify these algorithms to try out new ideas or to accommodate local conditions or needs
- so that they can be used as components (potentially modified)

Component software

- most interesting problems will require the coordinated application of many different techniques
- thus we need integrated interoperable software
- web services are one tool
- well designed software modules are another
- you should design your piece to be a cog in a big machine

Data complexity

- Dimensionality.
- Dynamic/evolving data: e.g., gene annotation, sequence, literature.
- Multiple data sources and locations: in-house, WWW.
- Multiple data types: numeric, textual, graphical.
- No longer X_{nxp}!
  - We distinguish between biological metadata and experimental metadata.

Bioconductor packages

Release 1.9, Oct 2006

~200 Packages

General infrastructure

Biobase, BioStrings, graph,multtest

Annotation:

annotate, biomaRt, annaffy, AnnBuilder -- data packages.

Graphics:

geneplotter, hexbin

Pre-processing Affymetrix oligonucleotide chip data:

affy, gcrma, affycomp

Pre-processing other array types

limma, beadarray vsn, marray, arrayMagic

Differential gene expression:

limma, genefilter, GOstats,siggenes, Category

Graphs and networks:

graph, RBGL, Rgraphviz, GOstats.

Other data: prada, EBImage, DNAcopy, aCGH

Affymetrix preprocessing

- Traditional Affymetrix genechips: calculation of per-transcript expression level estimates from the hybridization intensities
  - Background correction, Between-chip normalization, Probe set summarization
- Manufacturer's original algorithm was highly unprecise + problematic
- Many academics started to develop their alternatives, leading to vast improvement in the data quality of the technology
- This was one of the first ”success stories” of Bioconductor

Affycomp: a benchmark for Affymetrix genechip expression measures

o Data:

Spike-in (Affymetrix)

16 genes, 14 concentrations, complex background

Dilution series (GeneLogic) 60 x HGU95Av2,

liver & CNS cRNA in different proportions and amounts

o Benchmark:

15 quality measures regarding

-reproducibility

-sensitivity

-specificity

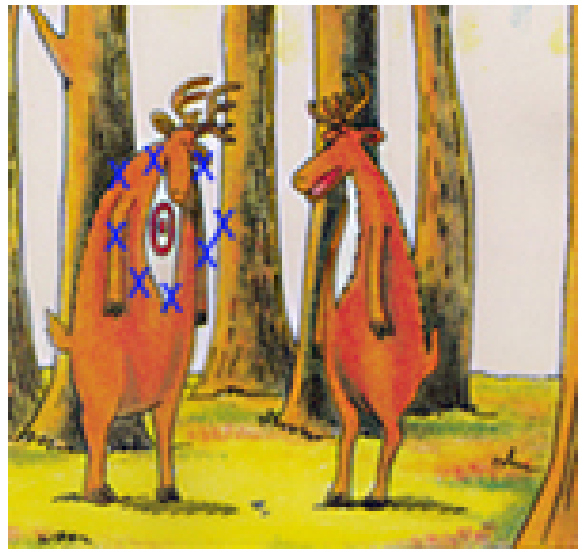
Put together by Rafael Irizarry (Johns Hopkins)

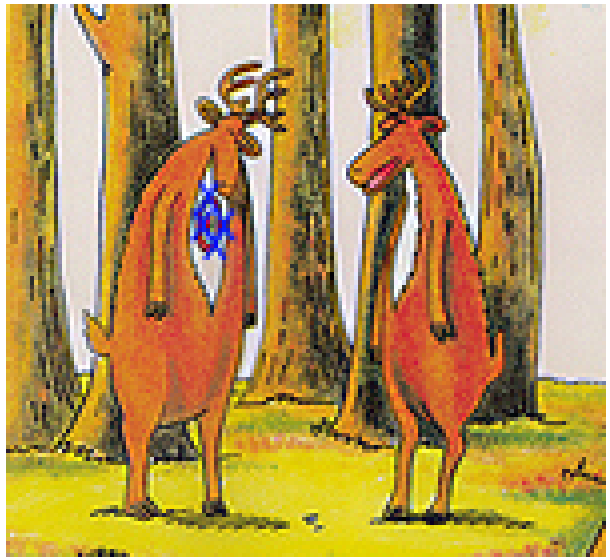
http://affycomp.biostat.jhsph.edu

Precision vs accuracy

←bias accuracy→

variance→ ← precision





ROC curves

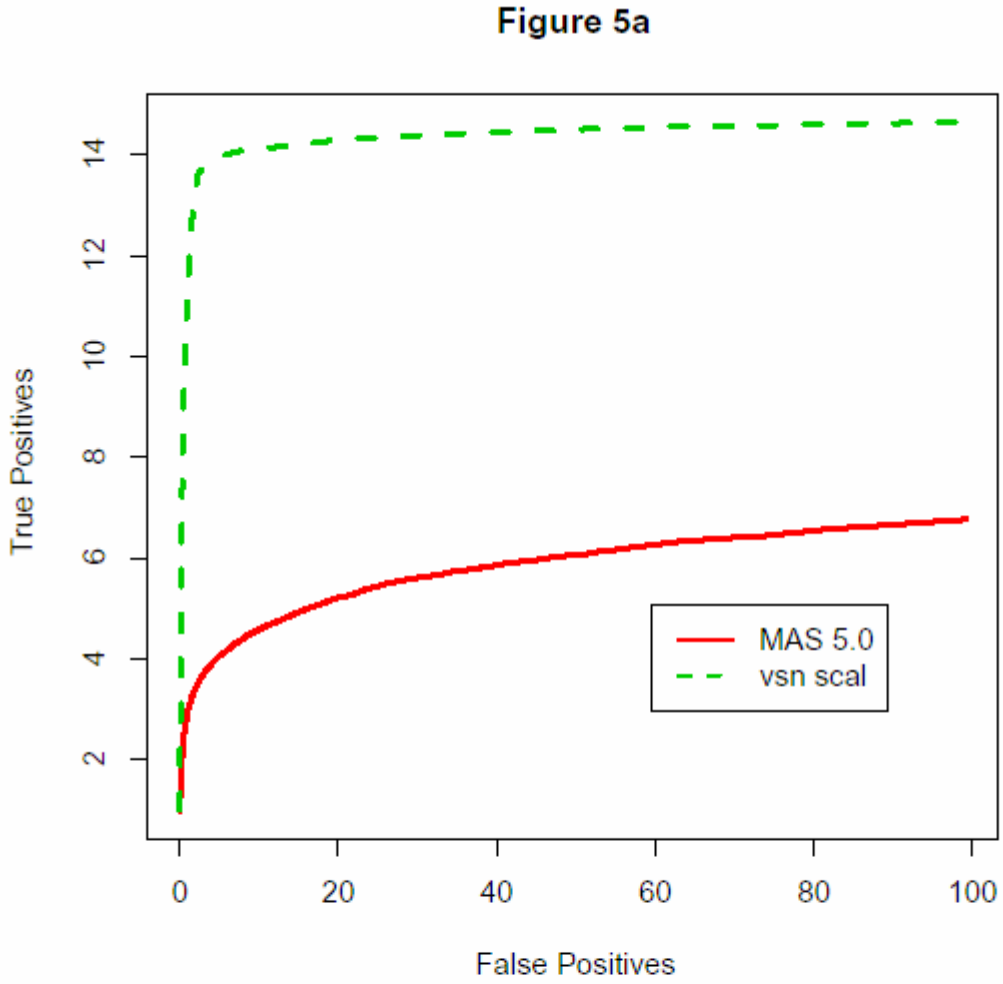
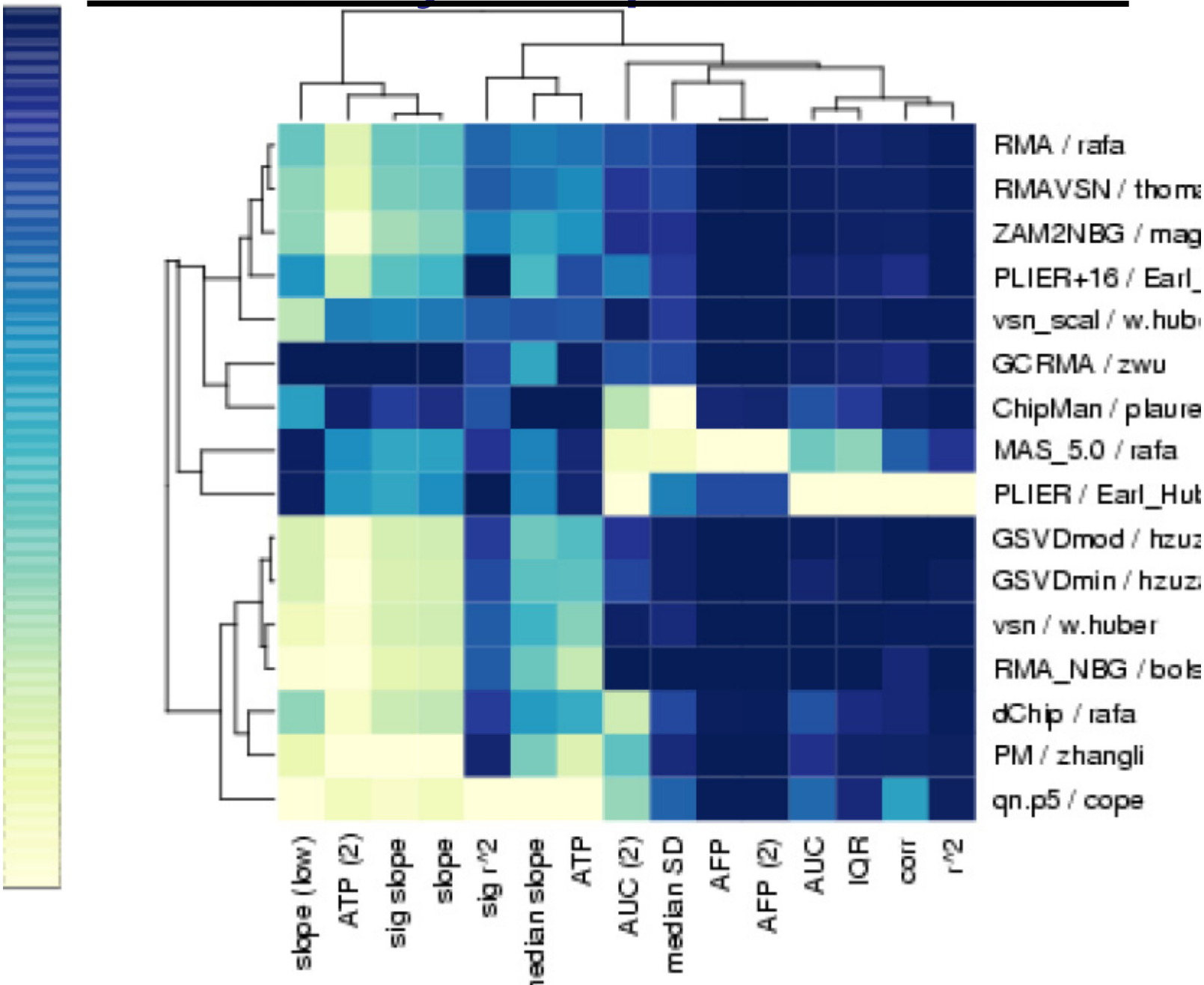


Figure 5a): A typical identification rule for differential expression filters genes with fold change exceeding a given threshold. This figure shows average ROC curves which offer a graphical representation of both specificity and sensitivity for such a detection rule.Average ROC curves based on comparisons with nominal fold changes ranging from 2 to 4096. b) As a) but with nominal fold changes equal to 2.

affycomp results



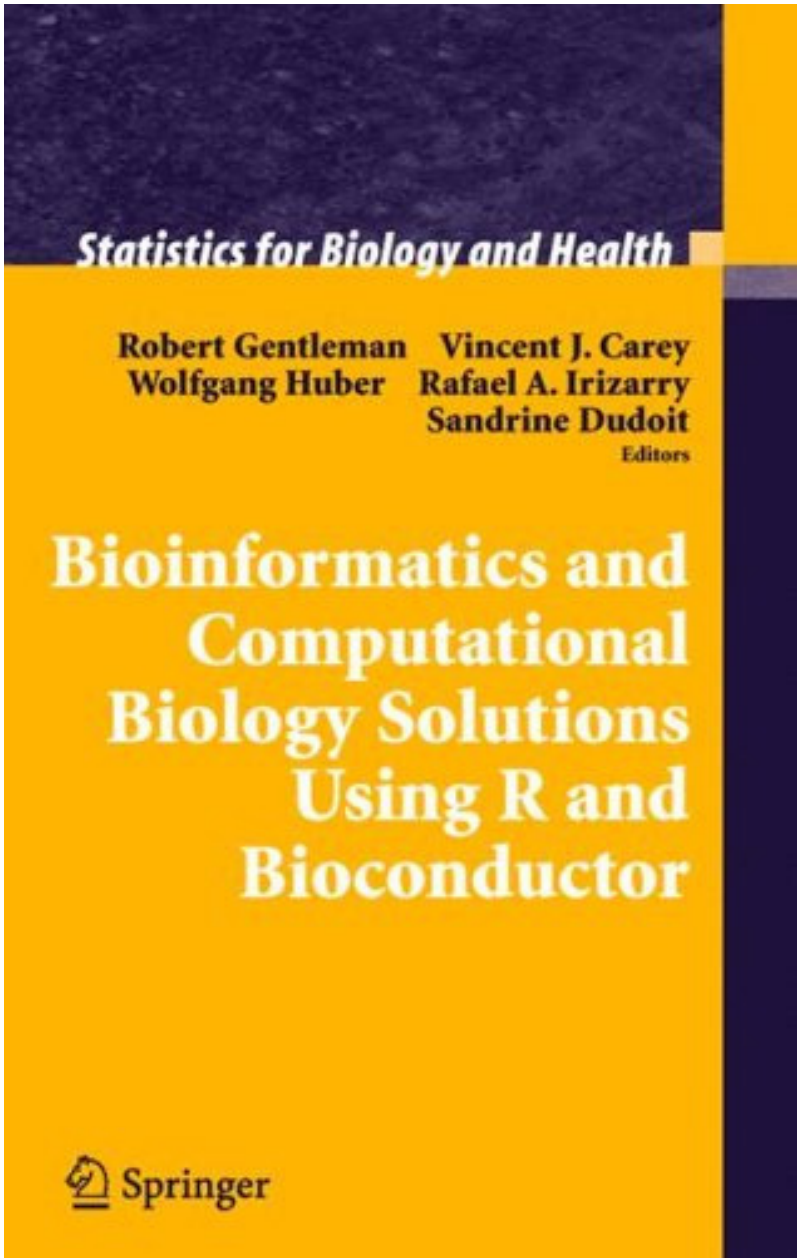
Conference

BioC2007, Fred Hutchison Research Center

August 2007

Developer's meeting and package demonstrations

Book



Preprocessing and normalization of microarray data, cell-based assays, mass spectrometry

Uni- and multivariate statistical analysis methods

Machine Learning

Harvesting and using metadata from biological databases

Visualization

Graphs and Networks in molecular biology

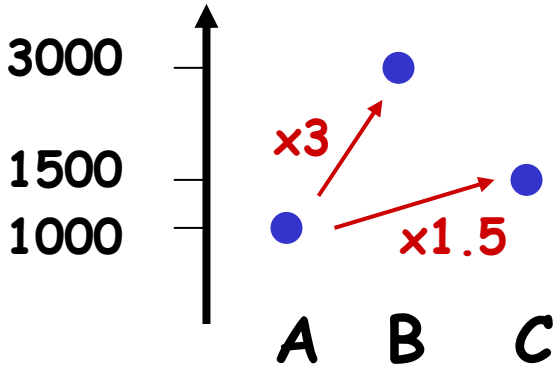
Including the software code (R) to reproduce all examples, figures etc.

Bioconductor

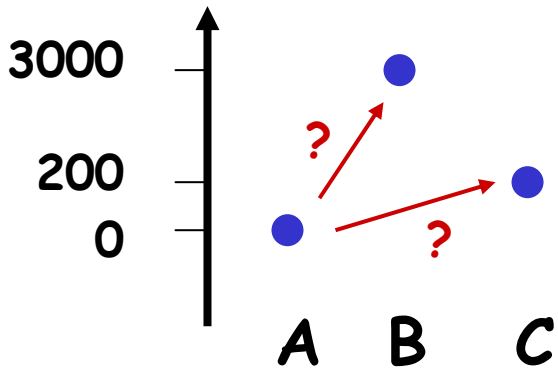
- **Bioconductor core team**
- Ben Bolstad, **UC Berkeley**
- Vince Carey**, Channing Laboratory, Harvard**
- Sandrine Dudoit**, Biostatistics, UC Berkeley**
- Seth Falcon**, FHCRC**
- Robert Gentleman, FHCRC
- Wolfgang Huber**, EMBL/EBI**
- Rafael Irizarry**, Biostatistics, Johns Hopkins**
- Nianhua Li**, FHCRC**
- Li Long**, ISB, Lausanne**
- Jim MacDonald**, U Michigan**
- Martin Morgan**, FHCRC**
- Herve Pages, **FHCRC**
- Gordon Smyth, **WEHI**
- Yee Hwa (Jean) Yang**, Sydney**

ratios and fold changes

Fold changes are useful to describe continuous changes in expression



But what if the gene is "off" (below detection limit) in one condition?



▶ fold change estimation and background correction

Many interesting genes will be off in some of the conditions of interest

Due to unspecific hybridization and optical noise, measured values are always > 0.

1. If you want expression measure to be an unbiased estimator of abundance

⇒ strong background correction, get many values ≤ 0

⇒ need something else than (log)ratio

2. If you let expression measure be biased (always>0)

⇒ weak background correction, then can keep ratios.

⇒ how do you choose the bias?

Sources of variation

amount of RNA in the sample
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-fluorescent detection

probe purity and length distribution
cross-/unspecific hybridization
stray signal

Systematic

- similar effect on many measurements
- corrections can be estimated from data

Stochastic

- too random to be explicitely accounted for
- remain as "noise"

Calibration

Error model

modeling ansatz

measured intensity = offset + gain × true abundance

$$y_{ik} = a_{ik} + b_{ik} x_k$$

$$a_{ik} = a_i + \varepsilon_{ik}$$

a_i per-sample offset

$\varepsilon_{ik} \sim N(0,\ b_i^2 s_1^2)$
"additive noise"

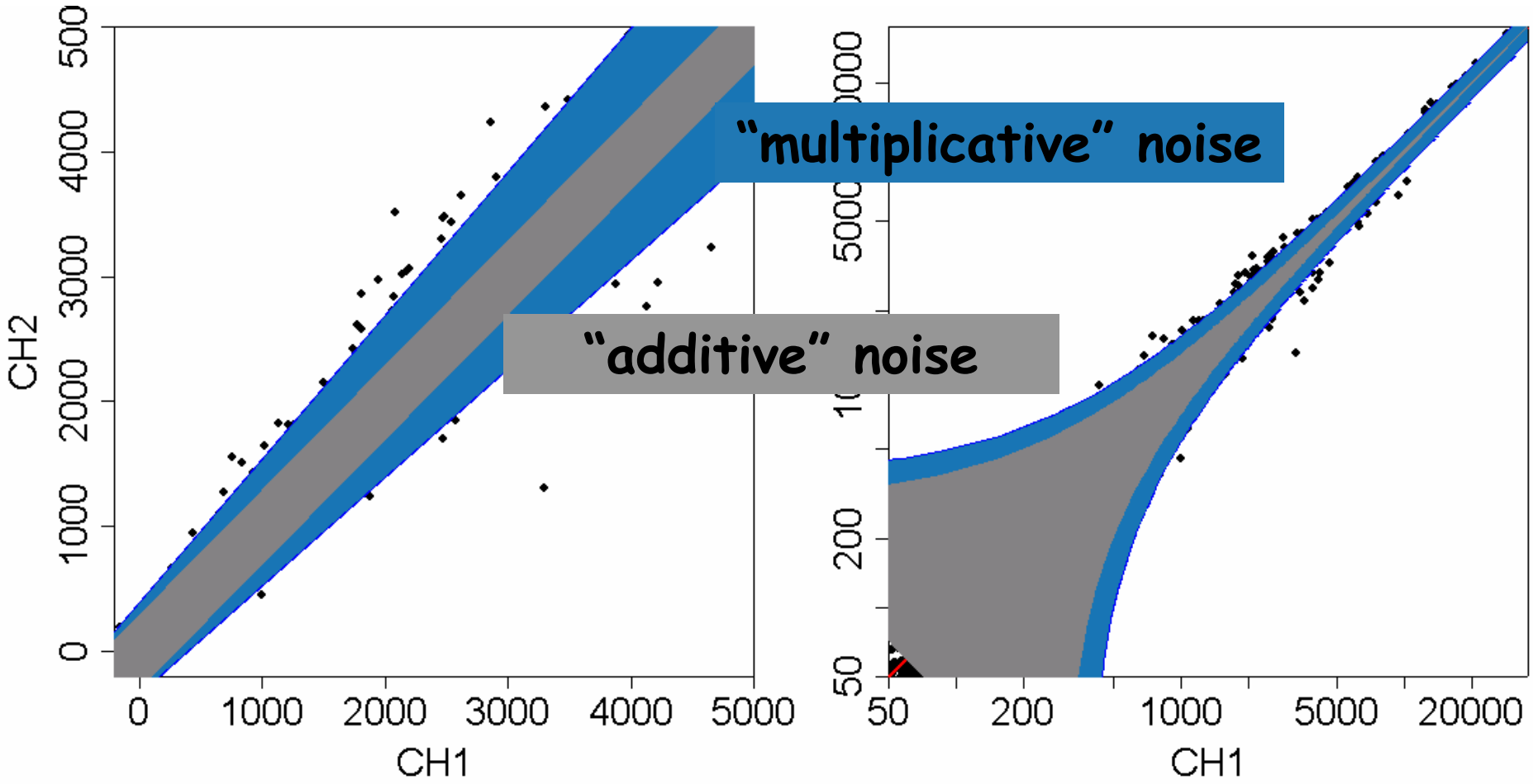
$$b_{ik} = b_i\ b_k\ \exp(\eta_{ik})$$

b_i per-sample normalization factor

b_k sequence-wise probe efficiency

$\eta_{ik} \sim N(0, s_2^2)$
"multiplicative noise"

The two-component model



raw scale **log scale**

B. Durbin, D. Rocke, JCB 2001

variance stabilizing transformations

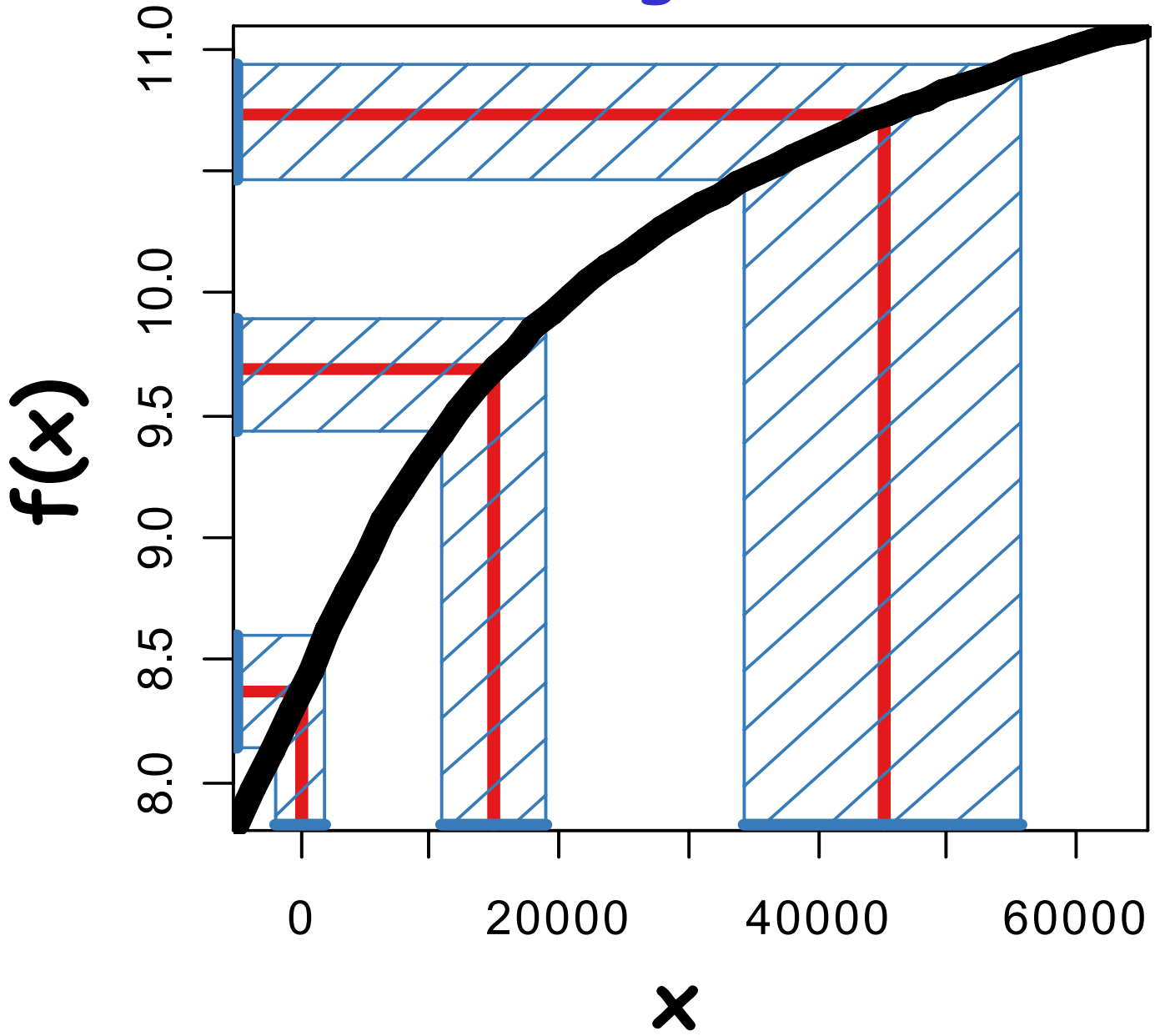
X_u a family of random variables with $EX_u = u$, $\mathrm{Var} X_u = v(u)$. Define

$$f(x) = \int^{x} \frac{1}{\sqrt{v(u)}}\, du$$

$\Rightarrow$ var $f(X_u) \approx$ independent of u

derivation: linear approximation

variance stabilizing transformations



variance stabilizing transformations

$$f(x) = \int^{x} \frac{1}{\sqrt{v(u)}}\, du$$

1.) constant variance ('additive') $v(u) = s^2 \quad \Rightarrow \quad f \propto u$

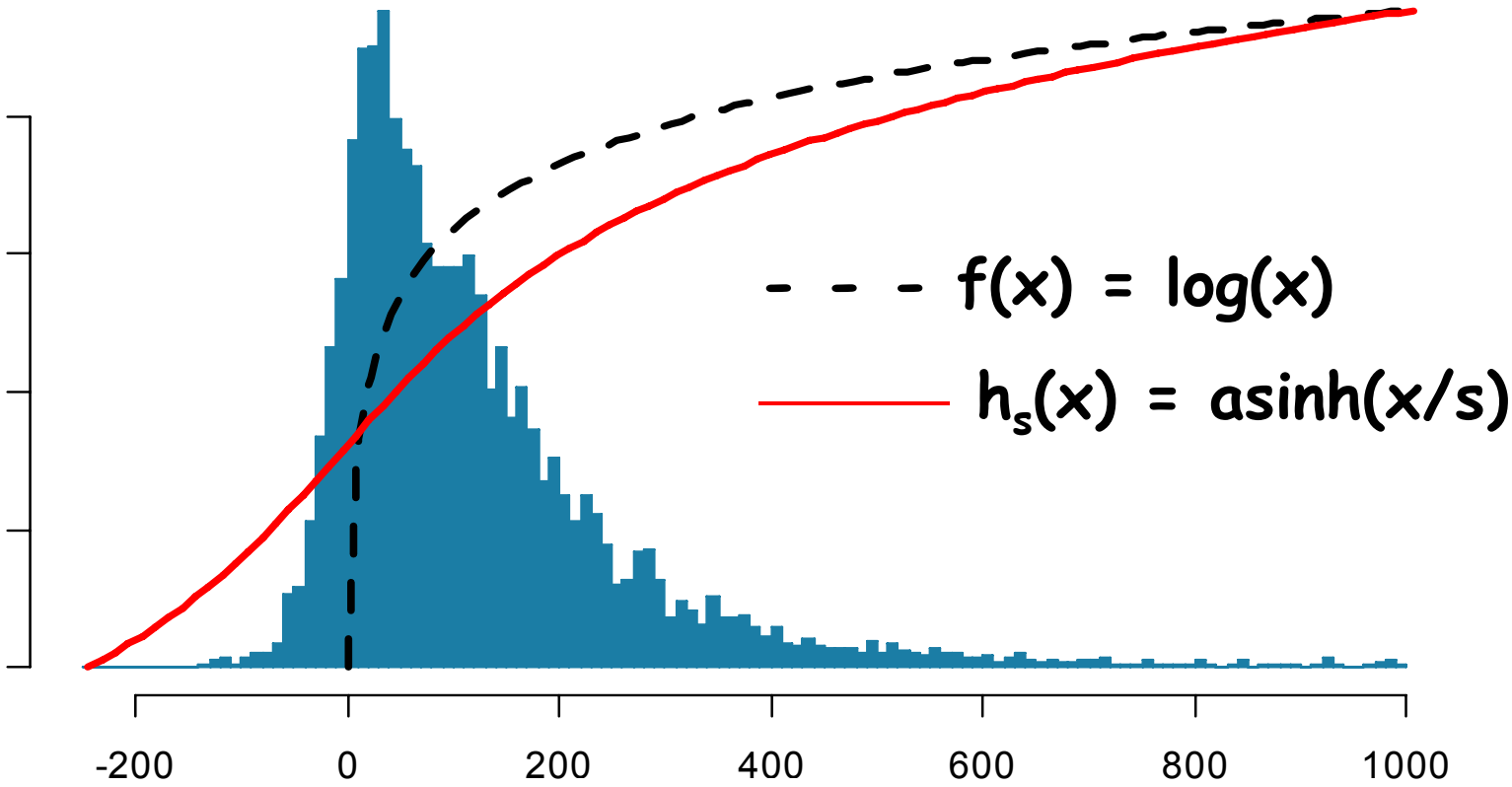
2.) constant CV ('multiplicative') $v(u) \propto u^2 \Rightarrow f \propto \log u$

3.) offset $v(u) \propto (u + u_0)^2 \quad \Rightarrow \quad f \propto \log(u + u_0)$

4.) additive and multiplicative

$$v(u) \propto (u + u_0)^2 + s^2 \Rightarrow f \propto \operatorname{arsinh} \frac{u + u_0}{s}$$

the "glog" transformation



$$\operatorname{arsinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right)$$

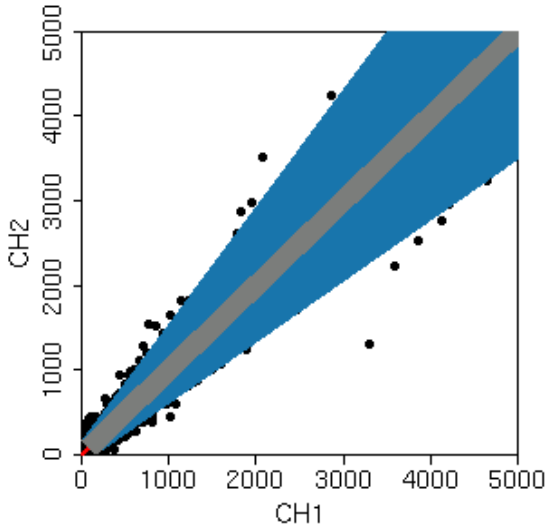
$$\lim_{x \to \infty} \left(\operatorname{arsinh} x - \log x - \log 2\right) = 0$$

P. Munson, 2001

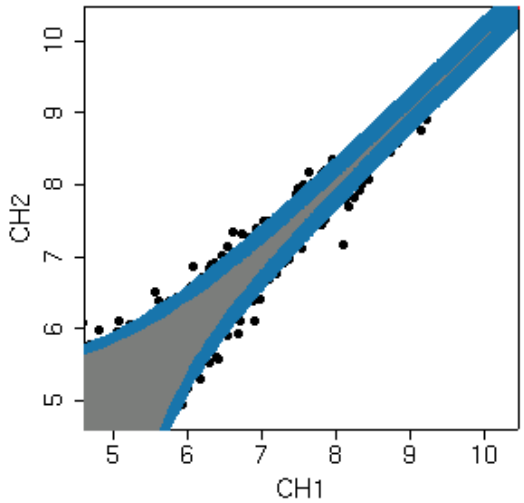
D. Rocke & B. Durbin, ISMB 2002

variance stabilization

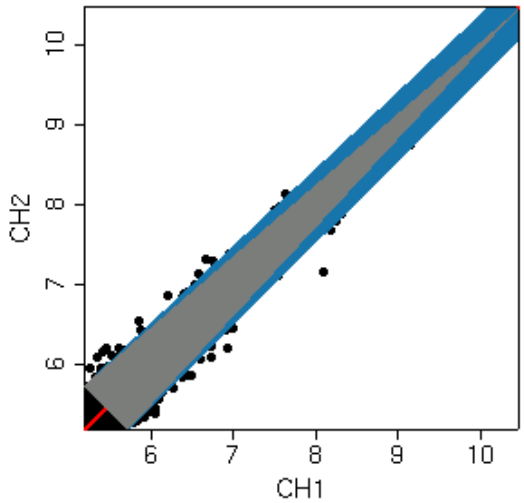
raw scale



log



glog



variance:

constant part

proportional part

parameter estimation

$$\operatorname{arsinh}\frac{y_{ki} - a_i}{b_i} = \mu_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim N(0, c^2)$$

- maximum lik… d
- but sensitiv…
- model holds … differentially …
- robust varic… Trimmed Sum…
- works well … differentially …

measured intensity = offset + gain * true abundance

$$y_{ik} = a_{ik} + b_{ik} x_{ik}$$

$a_{ik} = a_i + L_{ik} + \varepsilon_{ik}$

a_i per-sample offset

L_{ik} local background provided by image analysis

$\varepsilon_{ik} \sim N(0, b_i^2 s_1^2)$
"additive noise"

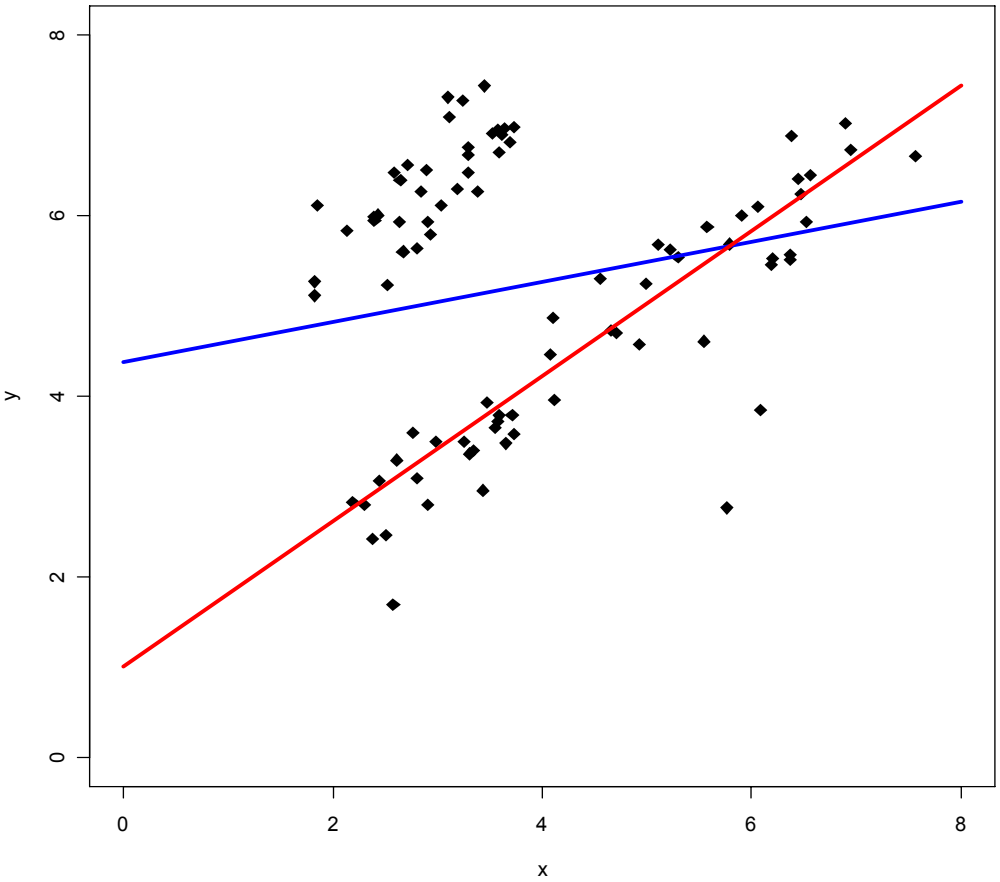
$b_{ik} = b_i \, b_k \exp(\eta_{ik})$

b_i per-sample normalization factor

b_k sequence-wise labeling efficiency

$\eta_{ik} \sim N(0, s_2^2)$
"multiplicative noise"

Least trimmed sum of squares regression



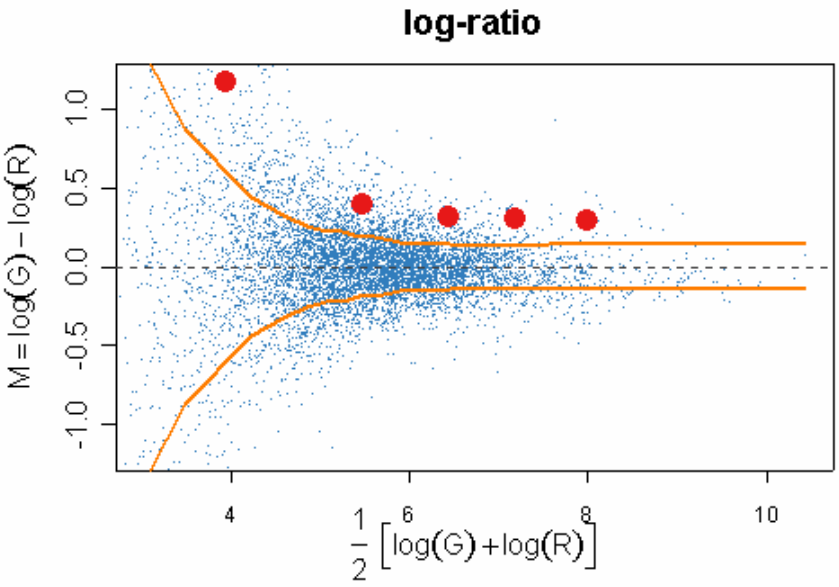
- least sum of squares
- least trimmed sum of squares

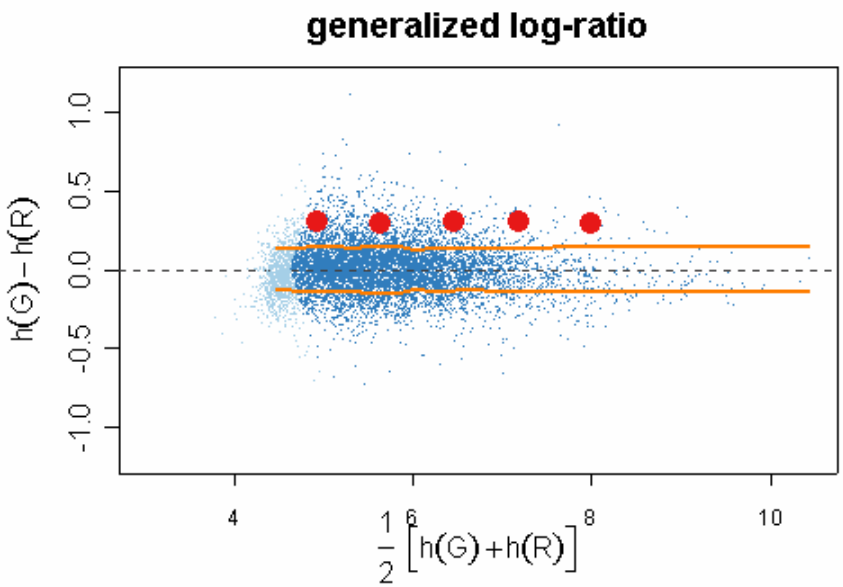
minimize

$$\sum_{i=1}^{n/2} \left(y_{(i)} - f(x_{(i)}) \right)^2$$

P. Rousseeuw, 1980s

glog





$$\log \frac{x_i}{x_j}$$

$$\log \frac{x_i + \sqrt{x_i^2 + c_i^2}}{x_j + \sqrt{x_j^2 + c_j^2}}$$

- For Affymetrix data, it turns out that the weak background correction method of RMA and the glog(-ratio) of vsn result in very similar results
- vsn also useful for other array platforms (e.g. spotted two-color)
- Don't be afraid of the "glog", it is equivalent to weak (=biased) background correction and normal log!
- vsn package (see vignette)
- Ref.: Huber, von Heydebreck et al., Bioinformatics 2002

Reproducible Research and Compendia

There is a tendency to accept seemingly realistic computational results, as presented by figures and tables, without any proof of correctness.

F. Leisch, T. Rossini, Chance 16 (2003)

We re-analyzed the breast cancer data from van't Veer et al. (2002). ... Even with some help of the authors, we were unable to exactly re-produce this analysis.

R. Tibshirani, B. Efron, SAGMB (2002)

Re-analysis of a breast cancer outcome study

E. Huang et al., Gene expression predictors of breast cancer outcome, The Lancet 361 (9369): 1590-6 (2003)

89 primary breast tumors on Affymetrix Chips (HG-U95av2), among them: 52 with 1-3 positive lymph nodes, 18 led to recurrence within 3 years, 34 did not.

Goal: predict recurrence

Claim: 5 misclassification errors, 1 unclear (leave-one-out cross-validation)

Method: Bayesian binary prediction trees (at the time, unpublished)

http://www.cagp.duke.edu

…we tried to reproduce these results, starting from the published μarray raw data (CEL files)

But couldn't.

The paper (and supplements) didn't contain the necessary details to re-implement their algorithm.

Authors didn't provide comparisons to simple well-known methods.

In our hands, all other methods resulted in worse misclassification results.

Is their new Bayesian tree method miles better than everything else?

Or was their analysis over-optimistic? (over-fitting, selection bias)

A general pattern

New publications often present a new microarray data set, and a new classification method.

Merits of the methods, and merits of the data are entangled.

Is it necessary to develop an ideosyncratic method?

Which result could be achieved with standard approaches? (accuracy vs. interpretability)

Is there a big difference and what are the reasons for it ? (errors happen … in implementation /validation)

Compendia

Interactive documents that contain:

- **Primary data**
- **Processing methods (computer code)**
- **Derived data, figures, tables and other output**
- **Text: research report (result, materials and methods, conclusions)**

Package compHuang: reanalysis of Huang et al. data, using different classification and preprocessing methods and a correct cross-validation procedure for estimating the prediction error

Based on R/Bioconductor's package and vignette technologies

M. Ruschhaupt, W. Huber, A. Poustka, U. Mansmann, Statistical Applications in Genetics and Molecular Biology (2004)

PLoS Medicine Feb 2005.

source markup (here: latex & R)

Sweave

processed document (here: PDF)

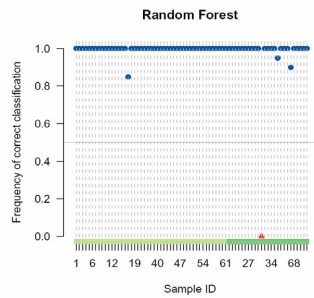
```
<<MCRestimate call,eval=FALSE,echo=TRUE>>=
r.forest <- MCRestimate(eset,class.label,
      class.function="RF.wrap",
      select.fun=red.fct,cross.outer=10,
      cross.inner=5, cross.repeat=20)
@
<<rf.save,echo=FALSE,results=hide>>=
savepdf(plot(r.forest, main="Random Forest"),"image-
RF.pdf")
@
<<result>>=
r.forest
@
The final document includes results of the
calculation, graphical outputs, tables, and optionally
parts of the R-Code which has been used. Also the
description of the experiment, the interpretation of
the results, and the conclusion can be integrated. In
this example we applied our compendium to T. Golubs
ALL/AML data~\cite{Golub.1999}.
\begin{figure}[h]
 \begin{center}
   \includegraphics[width=0.4\textwidth]{image-RF}
 \end{center}
 \end{figure}
\smallskip
<<summary,echo=FALSE>>=
method.list <- list(r.forest,r.pam,r.logReg,r.svm)
name.list   <- c("RF","PAM","PLR","SVM")
conf.table  <- MCRconfusion(method.list,
col.names=name.list)
@
<<writinglatex1,echo=FALSE, results=tex>>=
xtable(conf.table,"Overall number of
misclassifications",
label="conf.table",display=rep("d",6))
@
%\input{samples.1}
%\input{conf.table}
\begin{thebibliography}{1}
\bibitem[Golub et al. ,1999]{Golub.1999} Golub TR,
Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov
JP, Coller H, Loh ML, Downing JR, Caligiuri MA,
Bloomfield CD, Lander ES.
\newblock Molecular classification of cancer: class
discovery and class prediction by gene expression
monitoring
\newblock\textit{Science} 286(5439): 531-7 (1999).
\end{thebibliography}
```

	RF	PAM	PLR	SVM	Group size
ALL	0	0	0	0	47
AML	1	2	1	1	25
All	1	2	1	1	72

Table 1: Overall number of misclassifications

```
> r.forest <- MCRestimate(eset, class.label, class.function = "RF.wrap",
+     select.fun = red.fct, cross.outer = 10, cross.inner = 5, cross.repeat = 20)

> r.forest

  Result of MCRestimate with 20 repetitions of 10-fold cross-validation

Selection function     : g.red.highest.var.2000
Cluster function       : identity
Classification function: RF.wrap

The confusion table:
    ALL AML class error
ALL  47   0        0.00
AML   1  24        0.04
```

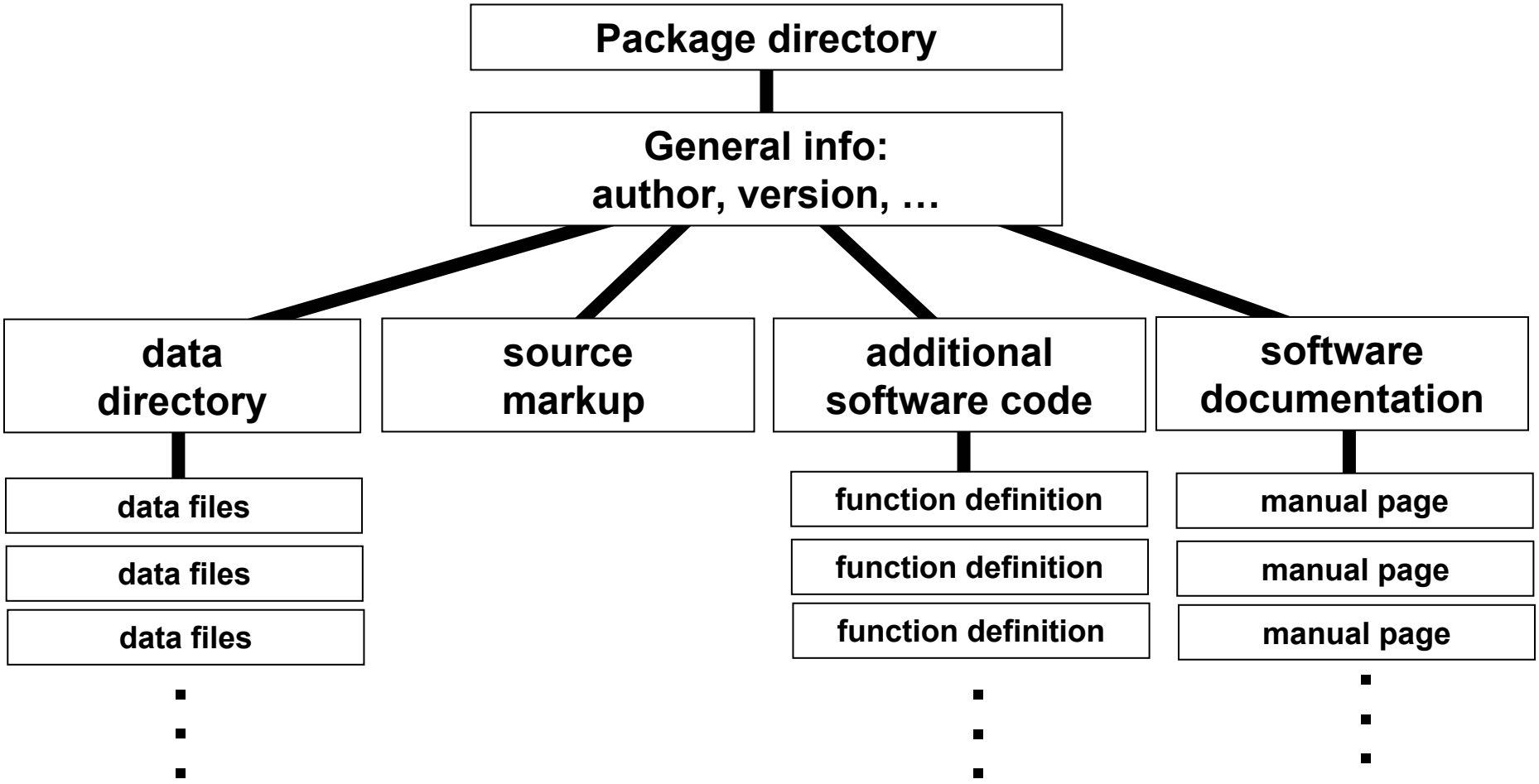
The final document includes results of the calculation, graphical outputs, tables, and optionally parts of the R-Code which has been used. Also the description of the experiment, the interpretation of the results, and the conclusion can be integrated. In this example we applied our compendium to T. Golubs ALL/AML data [Golub et al. ,1999].



References

[Golub et al. ,1999] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring *Science* 286(5439): 531-7 (1999).

Structure of a compendium



Compendia

See also the work by

Donald Knuth
HP Wolf
Günther Sawitzki
Friedrich Leisch
Robert Gentleman
Duncan Temple Lang

EBI

Oleg Sklyar
Ligia Bras
Elin Axelsson
Richard Bourgon
Jörn Tödling
Paul McGettigan
Tineke Casneuf
Matt Ritchie

EMBL

Lars Steinmetz
Fabiana Perocchi
Eugenio Mancera
Zhenyu Xu

DKFZ

Florian Fuchs
Viola Gesellchen
Thomas Horn
Dierk Ingelfinger
David Kuttenkeuler
Sandra Steinbrink
Michael Boutros

Florian Hahne
Dorit Arlt
Stefan Wiemann
Annemarie Poustka

Bioconductor

Robert Gentleman
Seth Falcon
Rafael Irizarry
Vince Carey

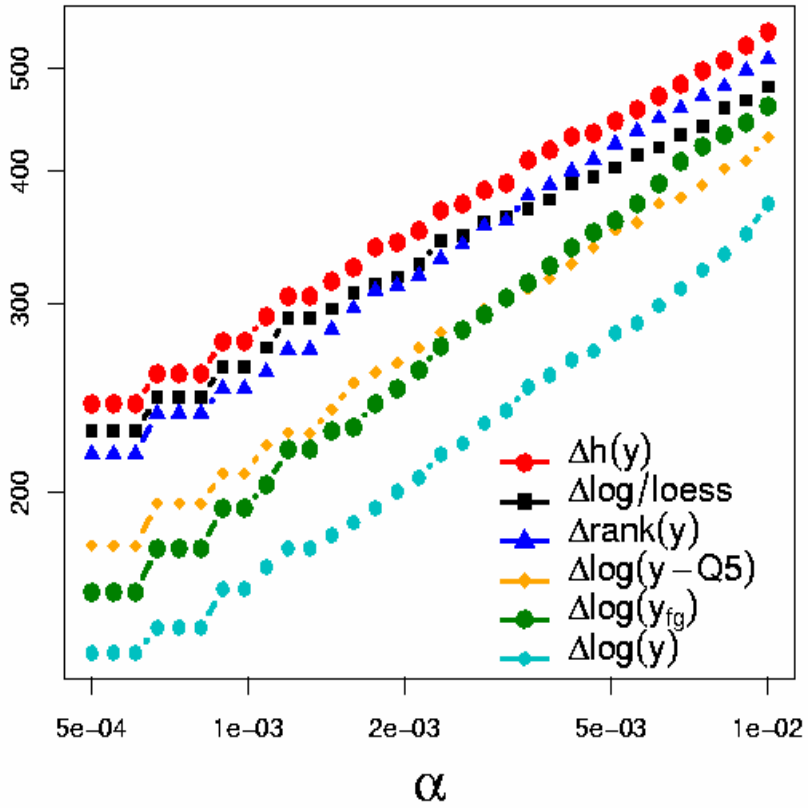
Amy Kiger

evaluation: sensitivity / specificity in detecting differential abundance

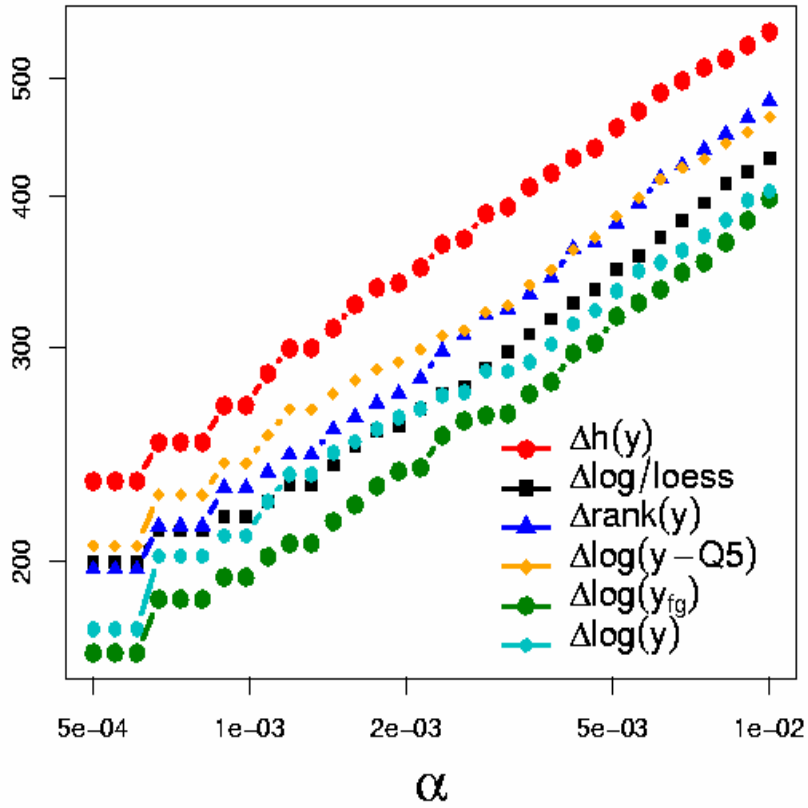
- **Data:** paired tumor/normal tissue from 19 kidney cancers, in color flip duplicates on 38 cDNA slides à 4000 genes.
- 6 different strategies for normalization and quantification of differential abundance
- Calculate for each gene & each method: *t*-statistics, permutation-*p*
- For threshold α, compare the number of genes the different methods find, $\#\{p_i \mid p_i \leq \alpha\}$

evaluation: comparison of methods

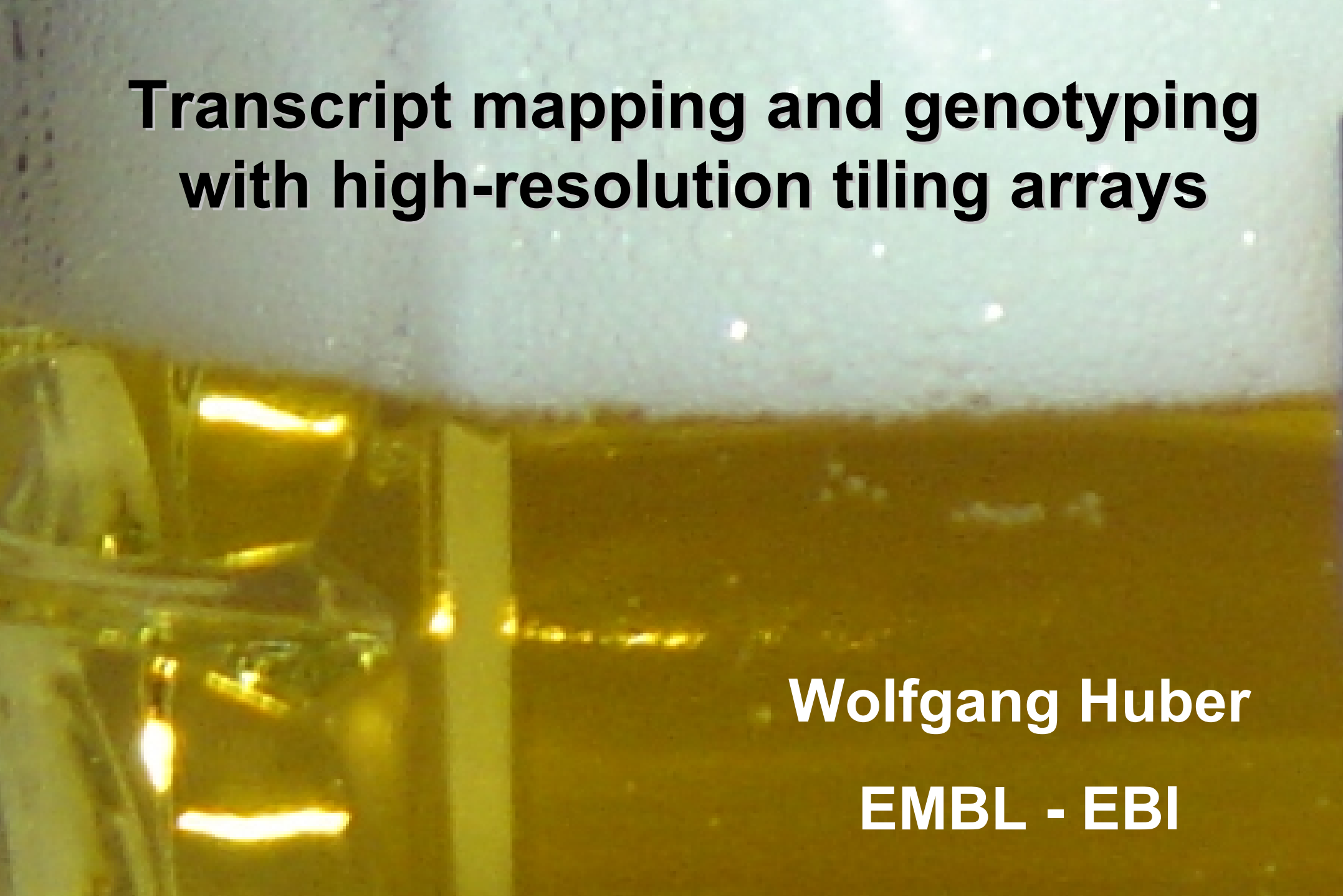
one-sided test for up



one-sided test for down



more accurate quantification of differential expression ⇒ higher sensitivity / specificity

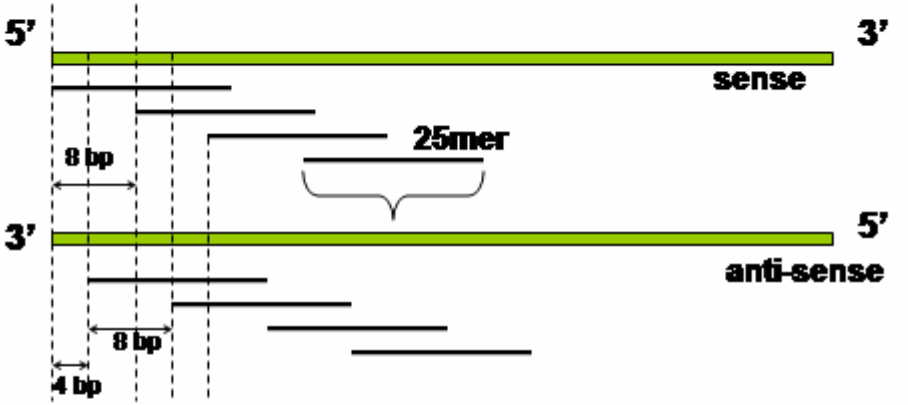


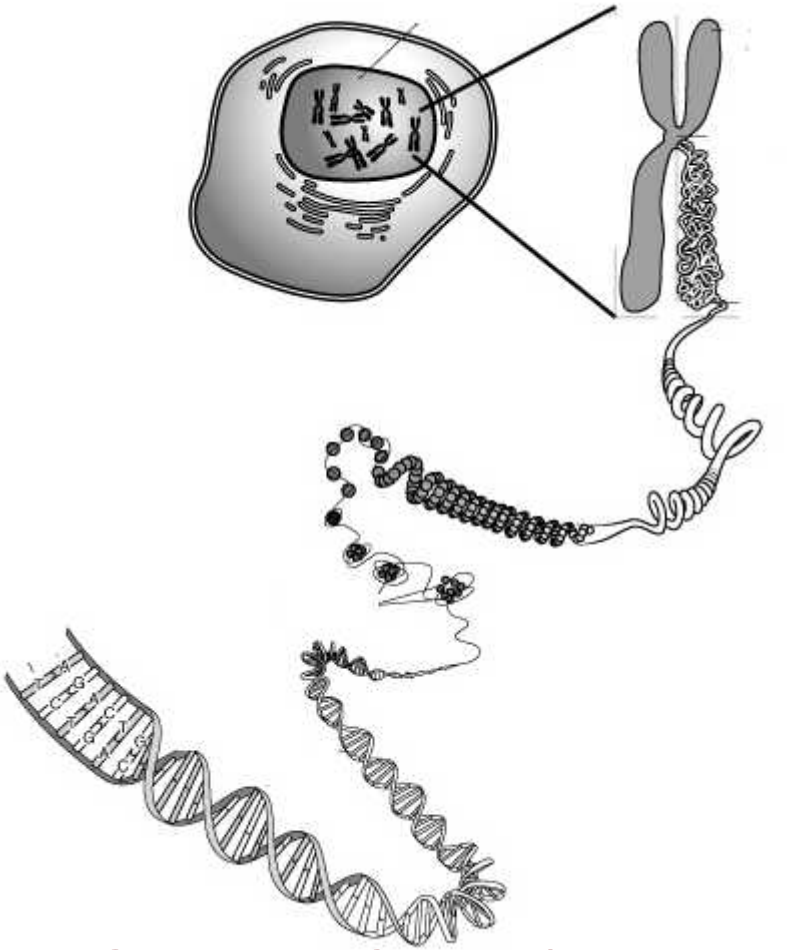
Transcript mapping and genotyping with high-resolution tiling arrays

Wolfgang Huber

EMBL - EBI

Genechip *S. cerevisiae* Tiling Array

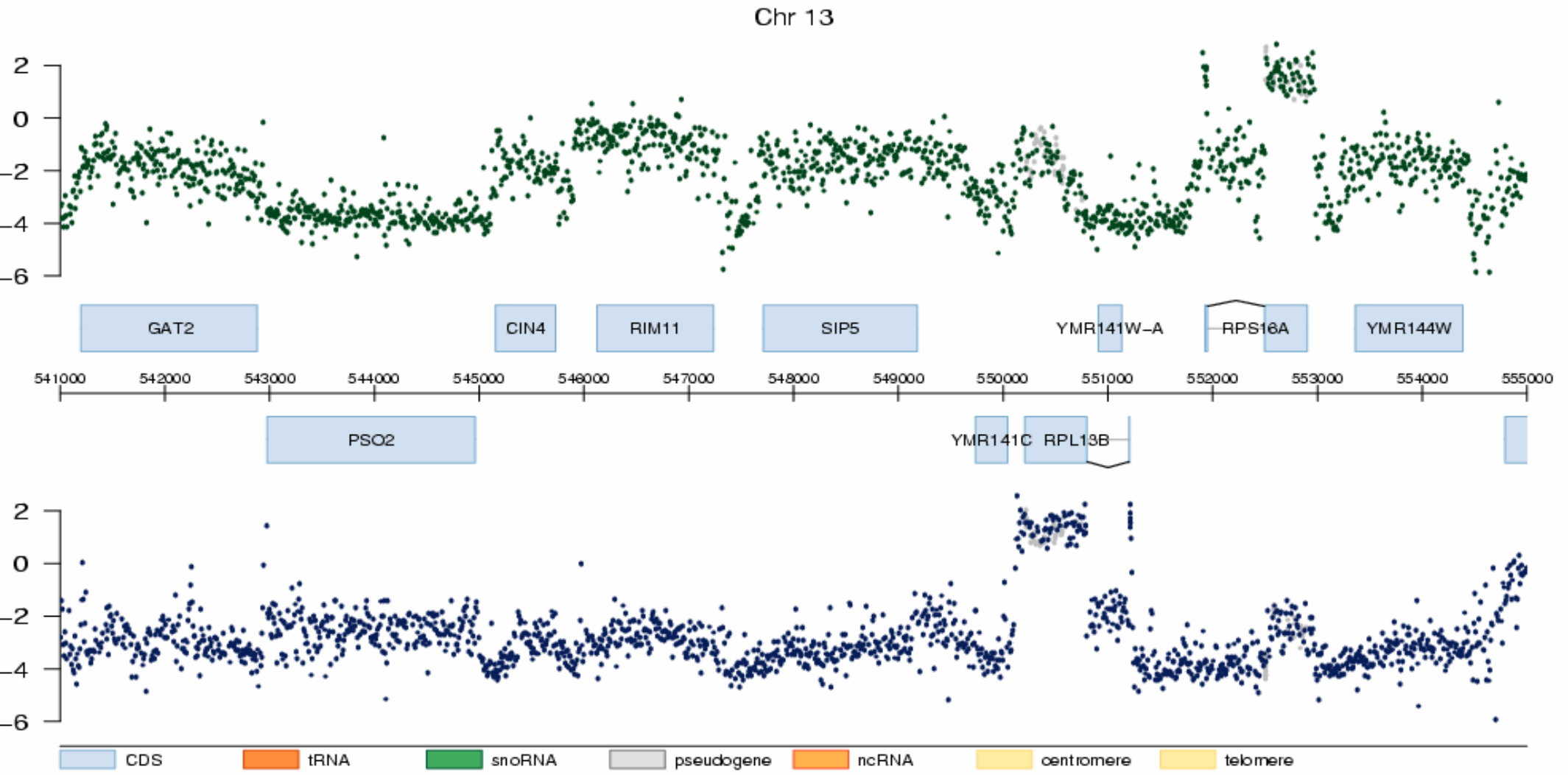




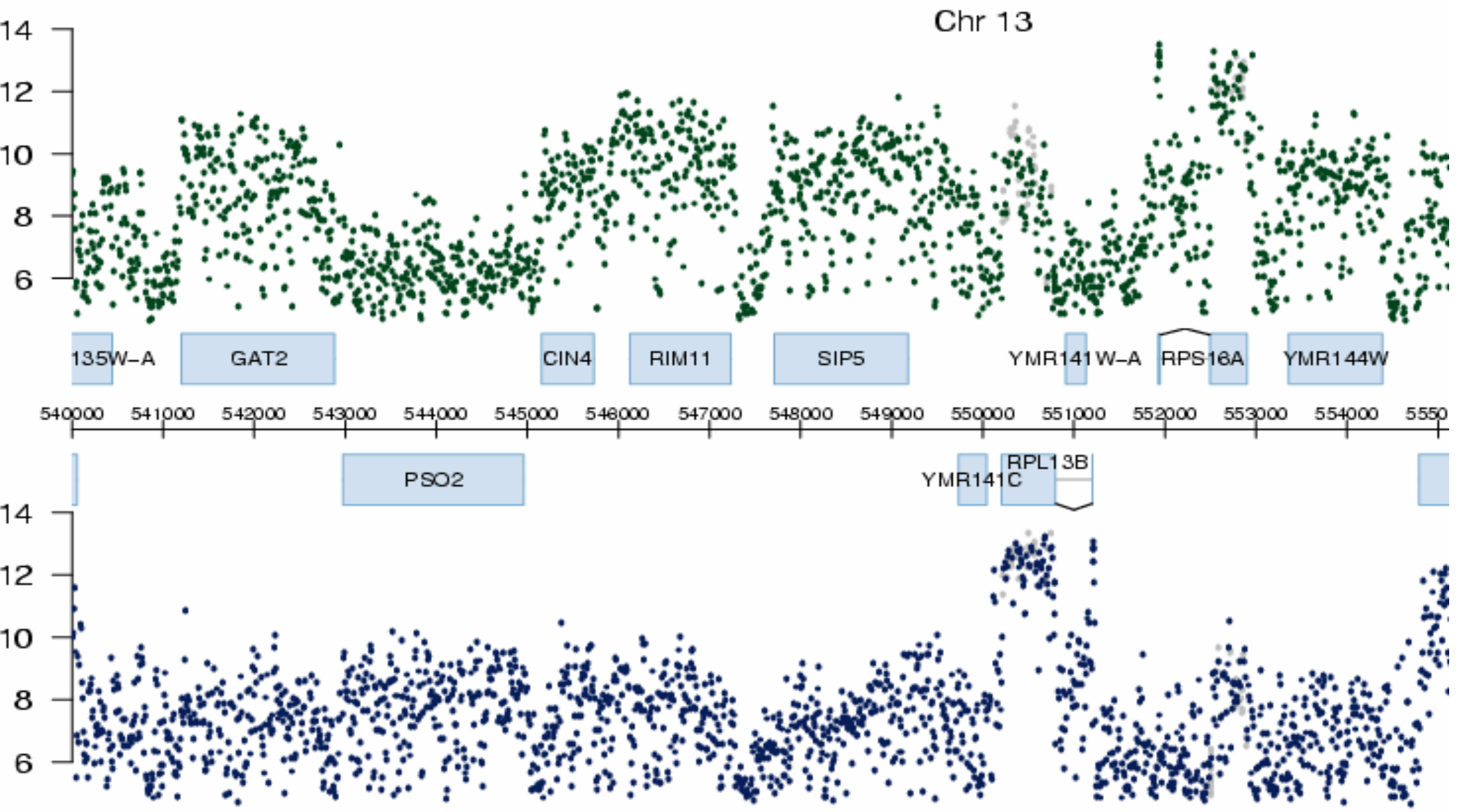
4 bp tiling path over complete genome
(12 M basepairs, 16 chromosomes)
Sense and Antisense strands
6.5 Mio oligonucleotides
5 μm feature size

manufactured by Affymetrix
designed by Lars Steinmetz (EMBL & Stanford Genome Center)

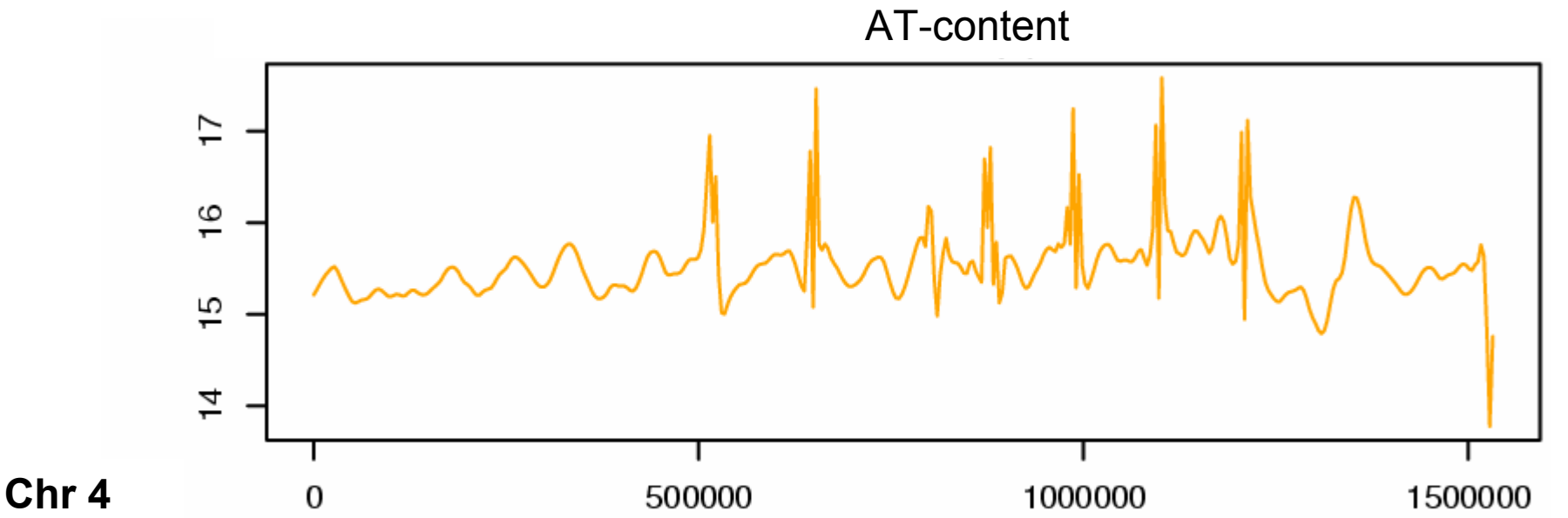
RNA Hybridization



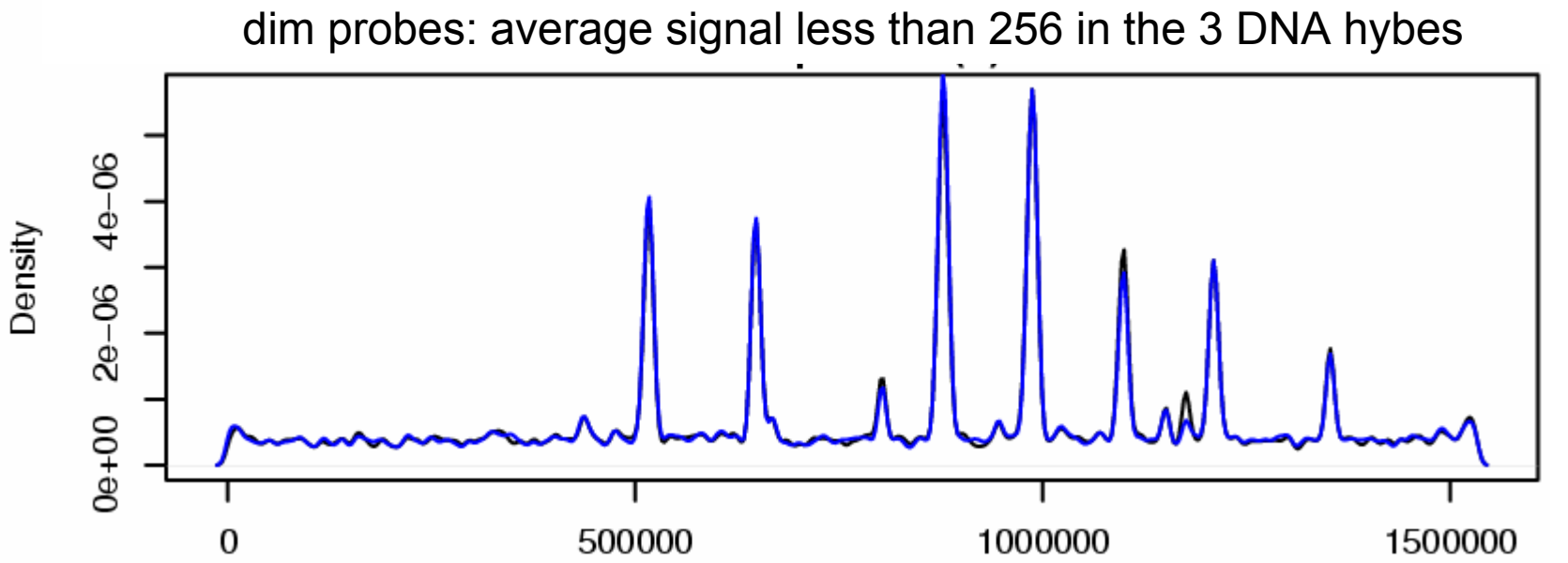
Before probe-specific normalization



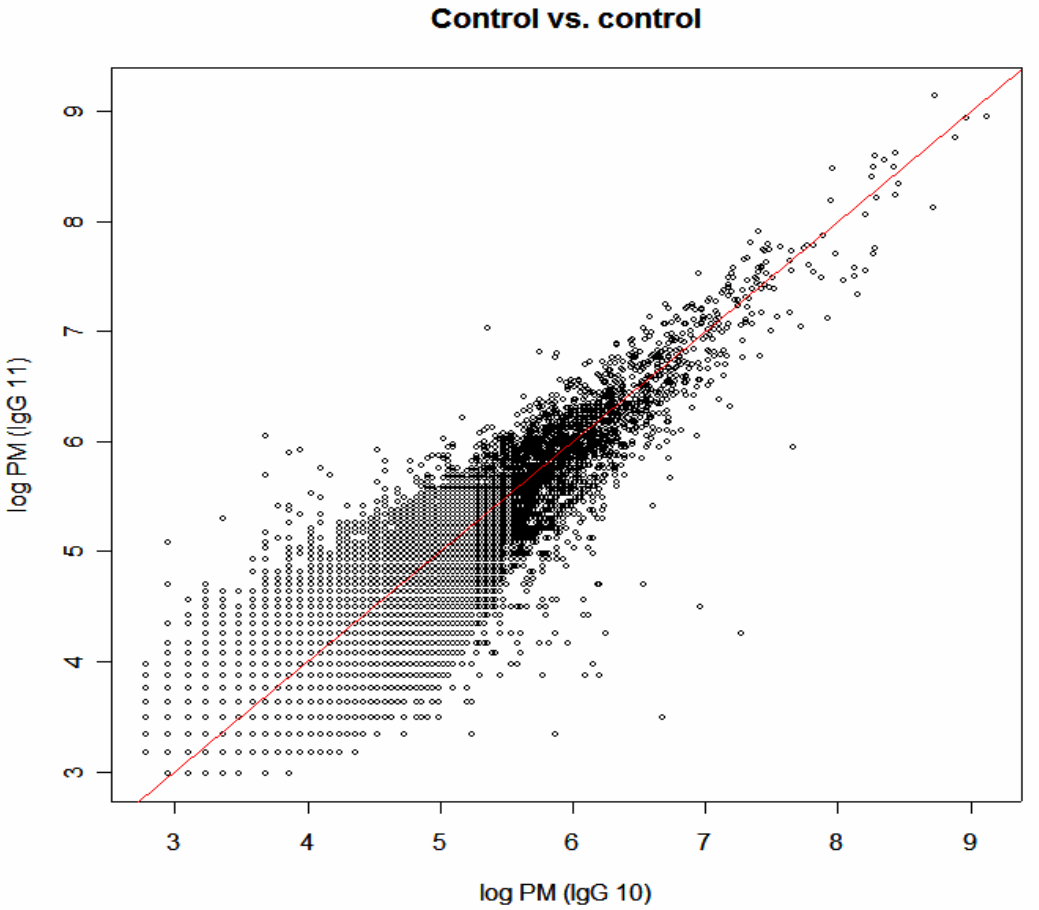
AT content and (weak) probe response

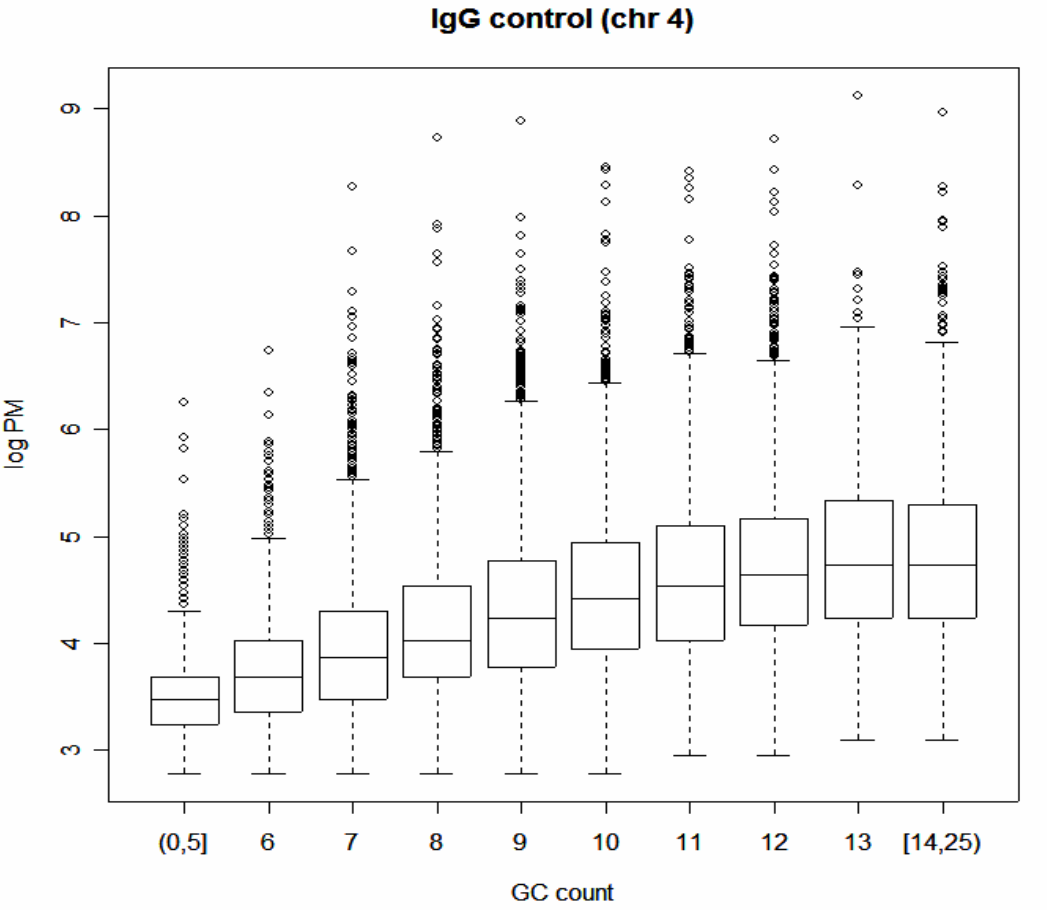


Chr 4



Probe response





Probe specific response normali-zation

$$\log_2 s_i$$

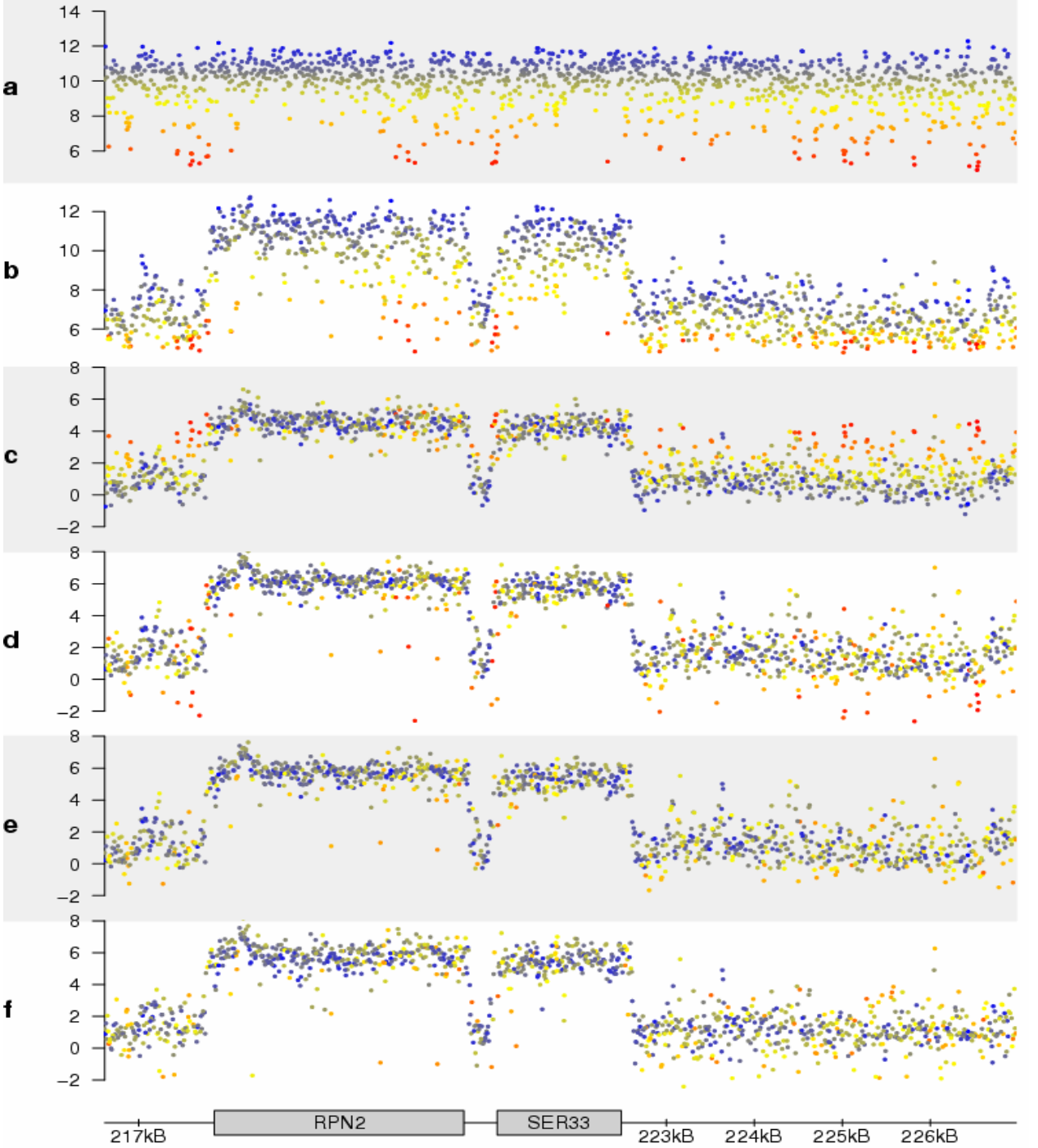
$$\log_2 y_i$$

$$q_i = \log_2 \frac{y_i}{s_i}$$

$$q_i = \mathrm{glog}_2 \frac{y_i - b(s_i)}{s_i}$$

remove 'dead' probes

$$q_i = \mathrm{glog}_2 \frac{PM_i - MM_i}{s_i}$$



S/N

3.22

3.47

4.04

4.58

4.36

Probe-specific response normalization

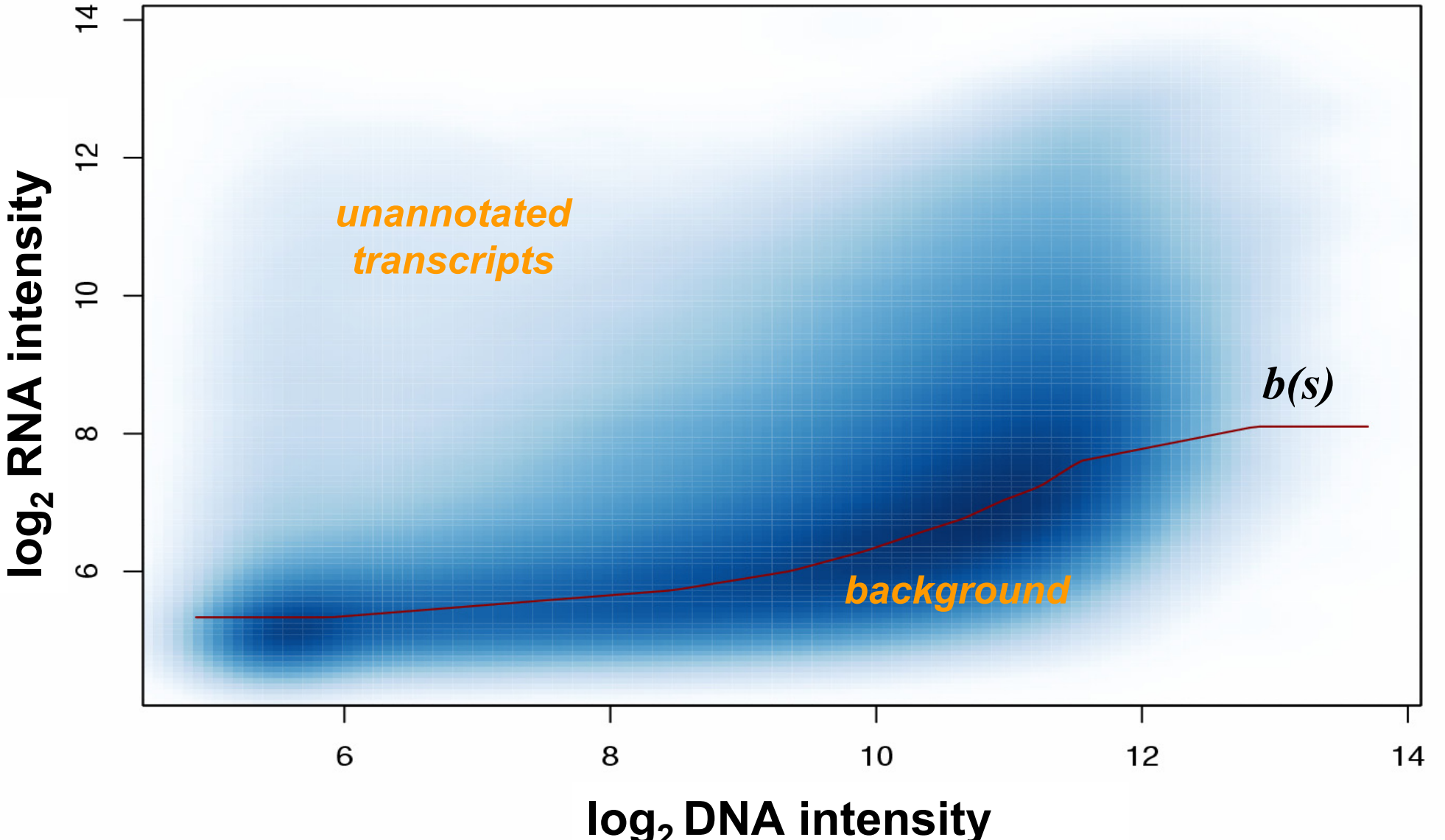
$$q_i = \mathrm{glog}_2 \frac{y_i - b(s_i)}{s_i}$$

s_i probe specific response factor.
Estimate directly from DNA hybridization data

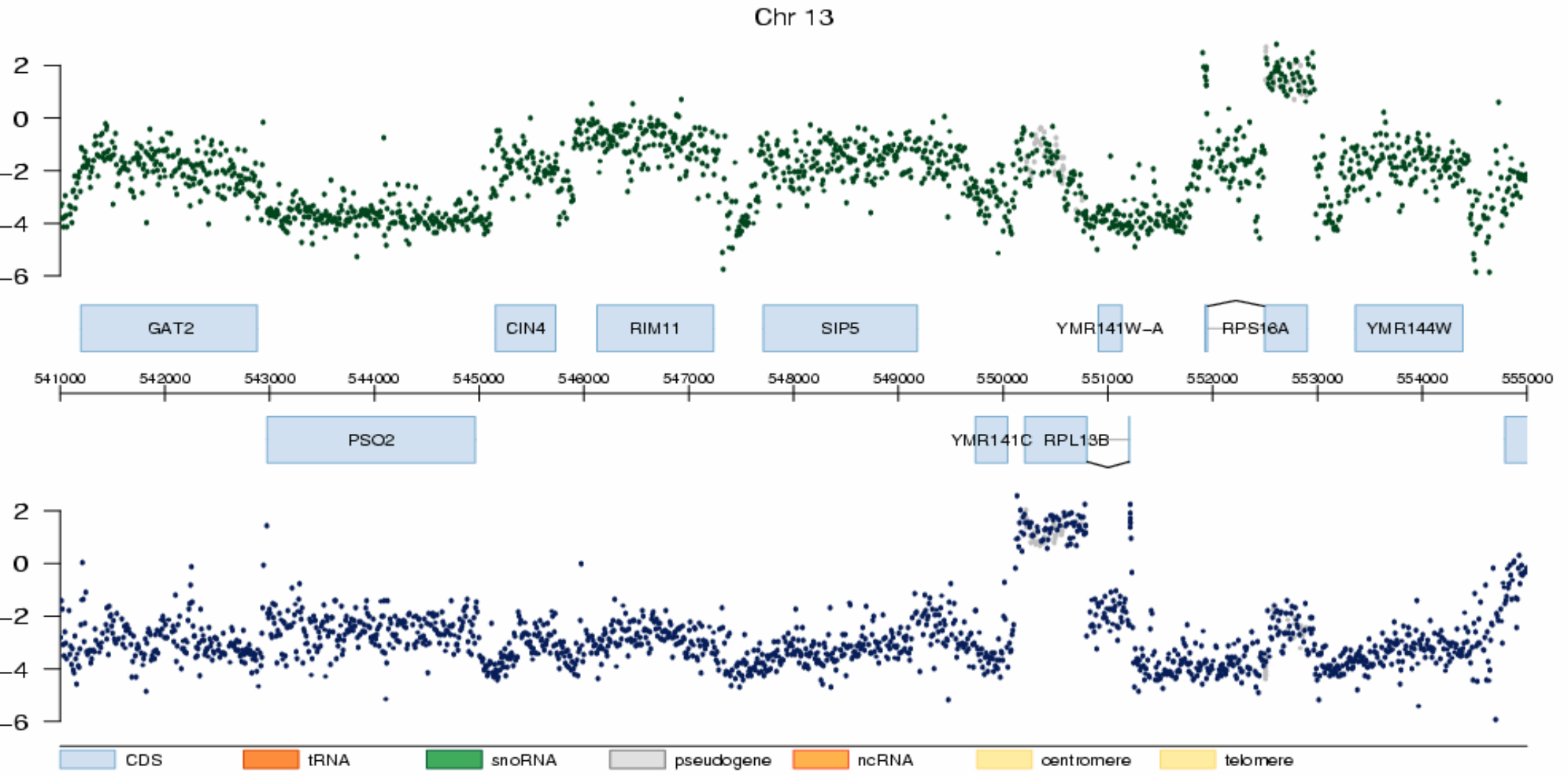
$b_i = b(s_i)$ probe specific background term. Estimate from signal of intergenic probes, interpolating to others by assuming that probes with similar s_i have similar b

Estimation of background *b*

from intergenic PM probes



After normalization



Segmentation

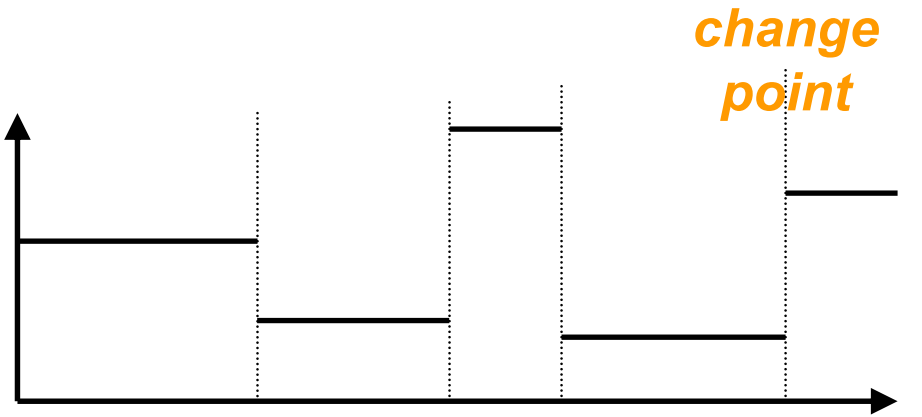
Two obvious options:

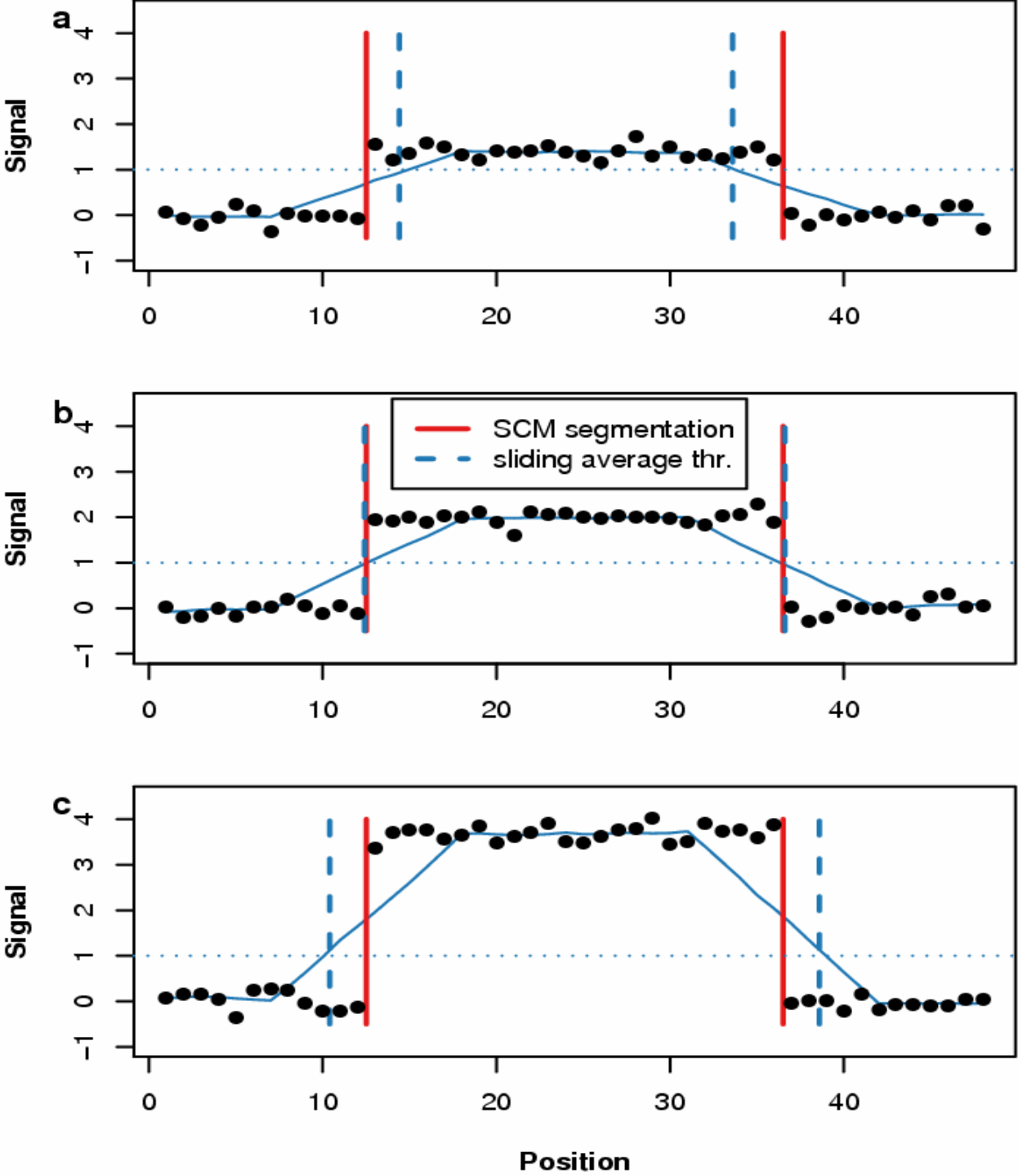
Smoothing **and thresholding: simple, but estimates of transcript boundaries will be *biased* and depend on expression level**

Hidden Markov Model **(HMM): but our "states" come from a continuum, unclear how to discretize**

Our solution:

Fit a piecewise constant function




a
b
c
Signal
Position
SCM segmentation
sliding average thr.
−1
0
1
2
3
4
0
10
20
30
40

Structural change model (SCM): piecewise constant functions

$$\forall x \in \left| t_{k-1}, t_k \right| :$$

$$Y(x) = \mu_k + \varepsilon(x)$$

$t_1, \ldots, t_S$: change points
Y: normalized intensities
x: genomic coordinates

μ_k: level of k-th segment

Model fitting

Minimize

$$G(t_1,\ldots,t_S)=\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{i>=t_s}^{i<t_{s+1}}\left(y_{ij}-\bar{y}_{sj}\right)^2$$

t_1,…, t_S: **change points**

J: **number of replicate arrays**

Optimization

Naïve optimization has complexity n^s, where $n \approx 10^5$ and $s \approx 10^3$.

Fortunately, there is a dynamic programming algorithm with complexity $O(n^2)$, and good heuristic $O(n)$:

$$k = 0, \quad \forall 0 \leq i < j \leq n \quad \hat{J}_1(i,j) = \sum_{x=i+1}^{j} \left\{ \log(2\pi \times \hat{\sigma}_1^2) + \left[\frac{y(x) - \hat{\mu}_1}{\hat{\sigma}_1} \right]^2 \right\}$$

$$\forall k \in [1, K_{max}] \quad \hat{J}_{k+1}(1,j) = \min_h \left\{ \hat{J}_k(1,h) + \hat{J}_1(h+1,j) \right\}$$

F. Picard, S.Robin, M. Lavielle, C. Vaisse, G. Celeux, JJ Daudin, BMC Bioinformatics (2005)

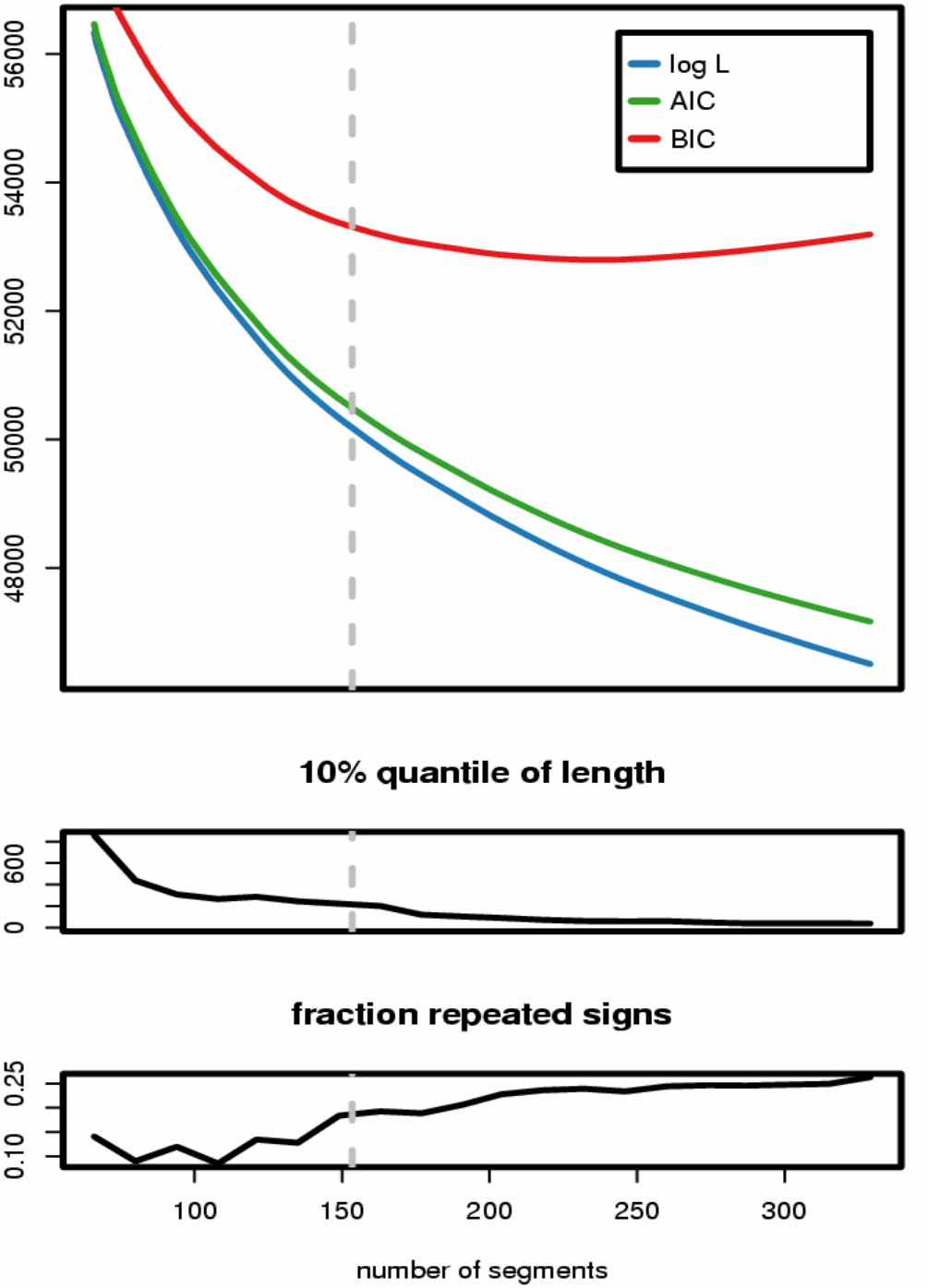
Bai+Perron, Journal of Applied Econometrics (2003)

Software: W. Huber, package tilingArray, www.bioconductor.org

A. Zeileis, package strucchange, CRAN

Model selection criteria

model family has just one parameter: no. of segments



Confidence Intervals

$$\frac{\left(\Delta_i^t Q_i \Delta_i\right)^2}{\left(\Delta_i^t \Omega_i \Delta_i\right)^2}\left(\hat{t}_i - t_i\right) \Rightarrow \arg\max_s V_i(s)$$

Δ_i **level difference**

Q_i **no. data points per unit *t***

Ω_i **error variance (allowing serial correlations)**

$t_i, \hat{t}_i$ **true and estimated change points**

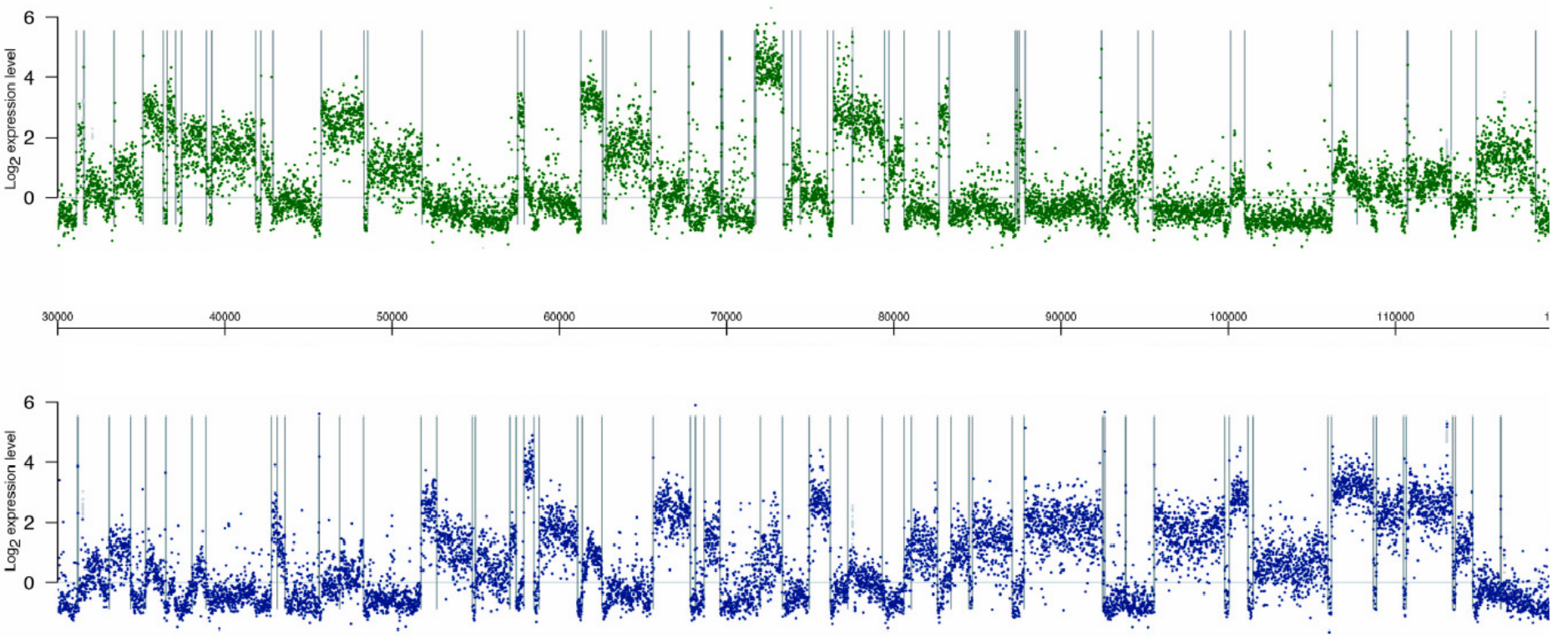
$V_i(s)$ **appropriately scaled and shifted Wiener process (Brownian motion)**

Bai and Perron, J. Appl. Econometrics 18 (2003)

Transcript Mapping by Segmentation

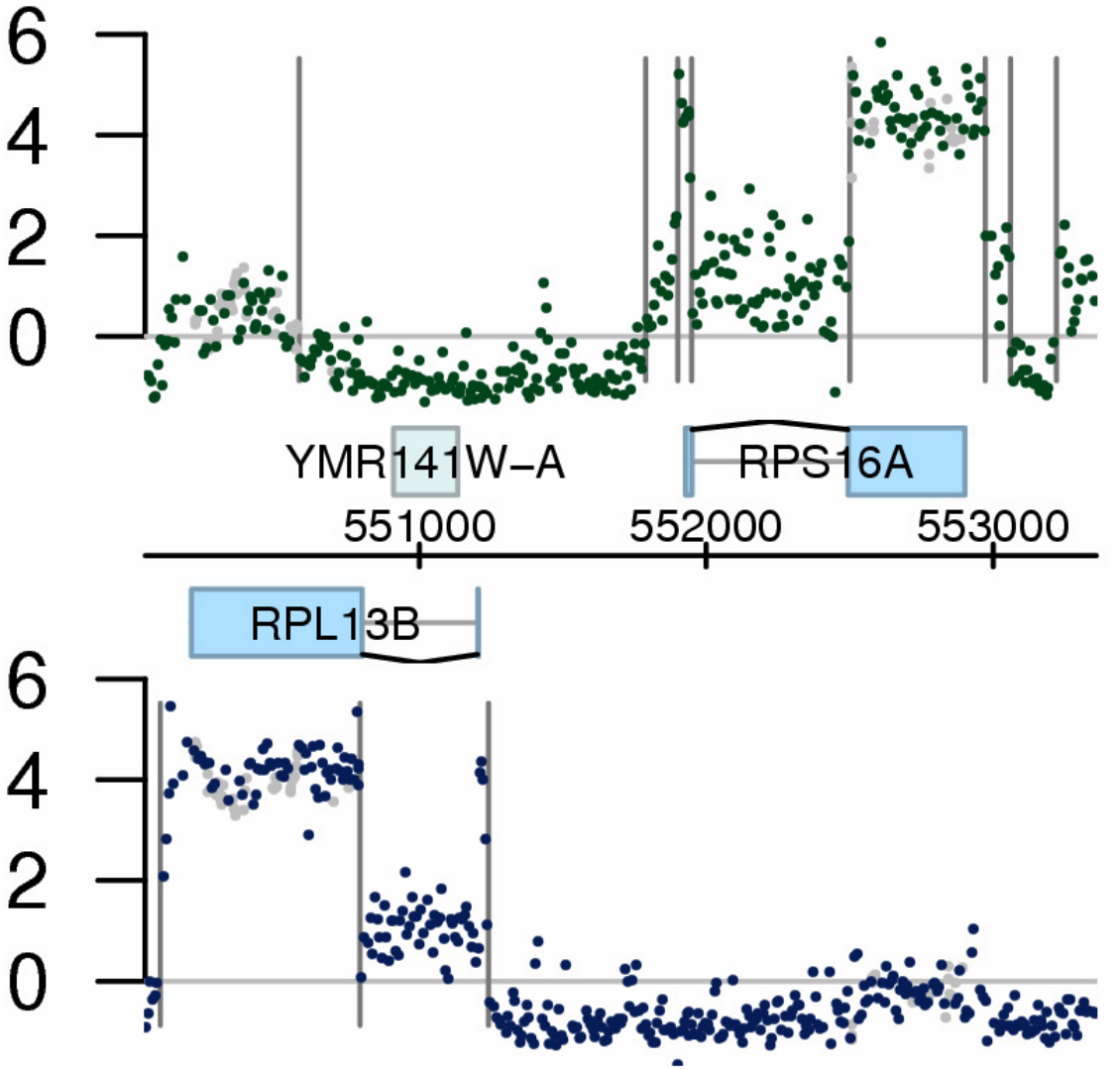
Along-chromosome plots for each strand (available in database):

www.ebi.ac.uk/huber-srv/queryGene



David *et al*., PNAS (2006)

A closer look



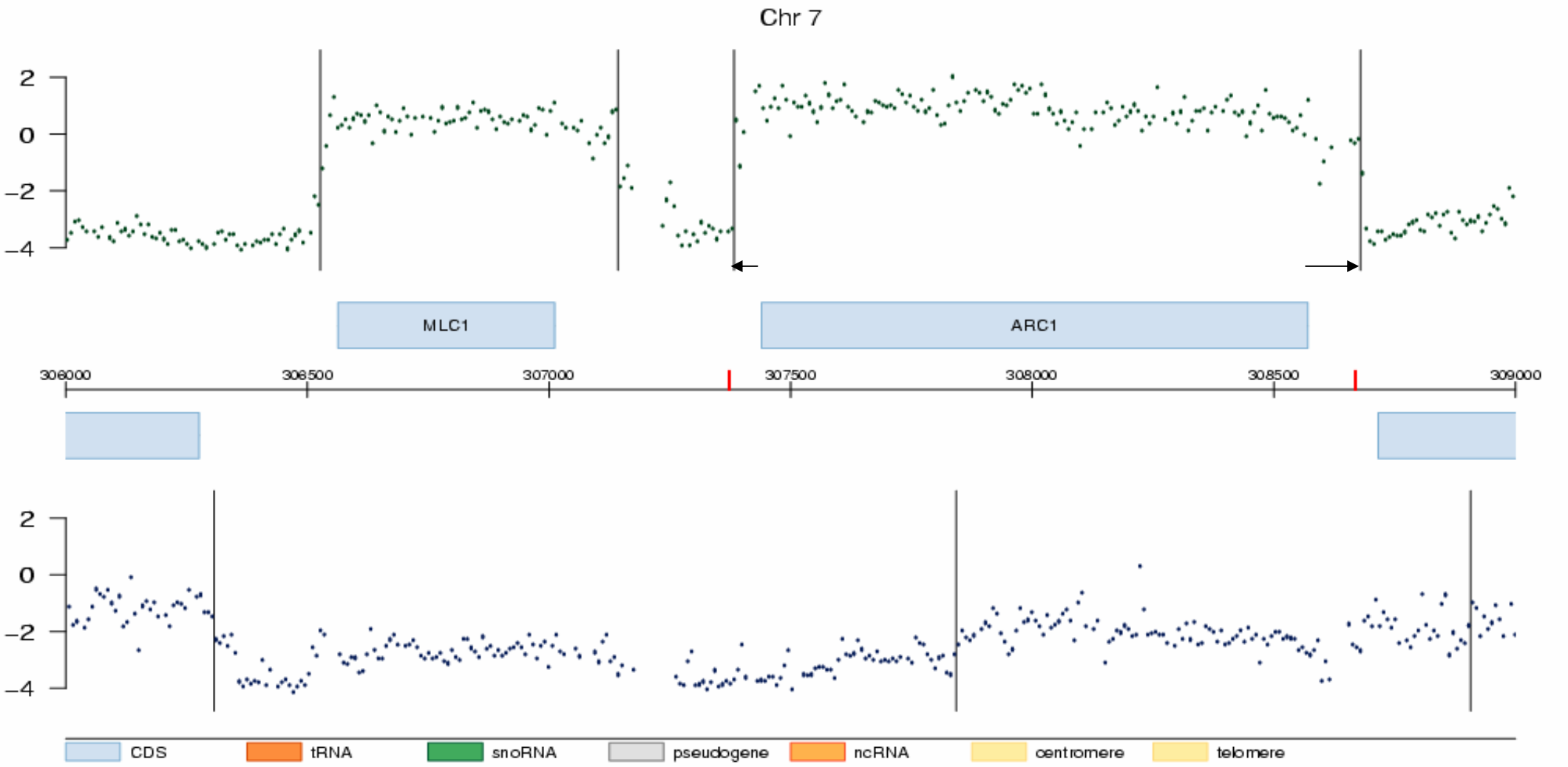
Splicing

Transcribed introns

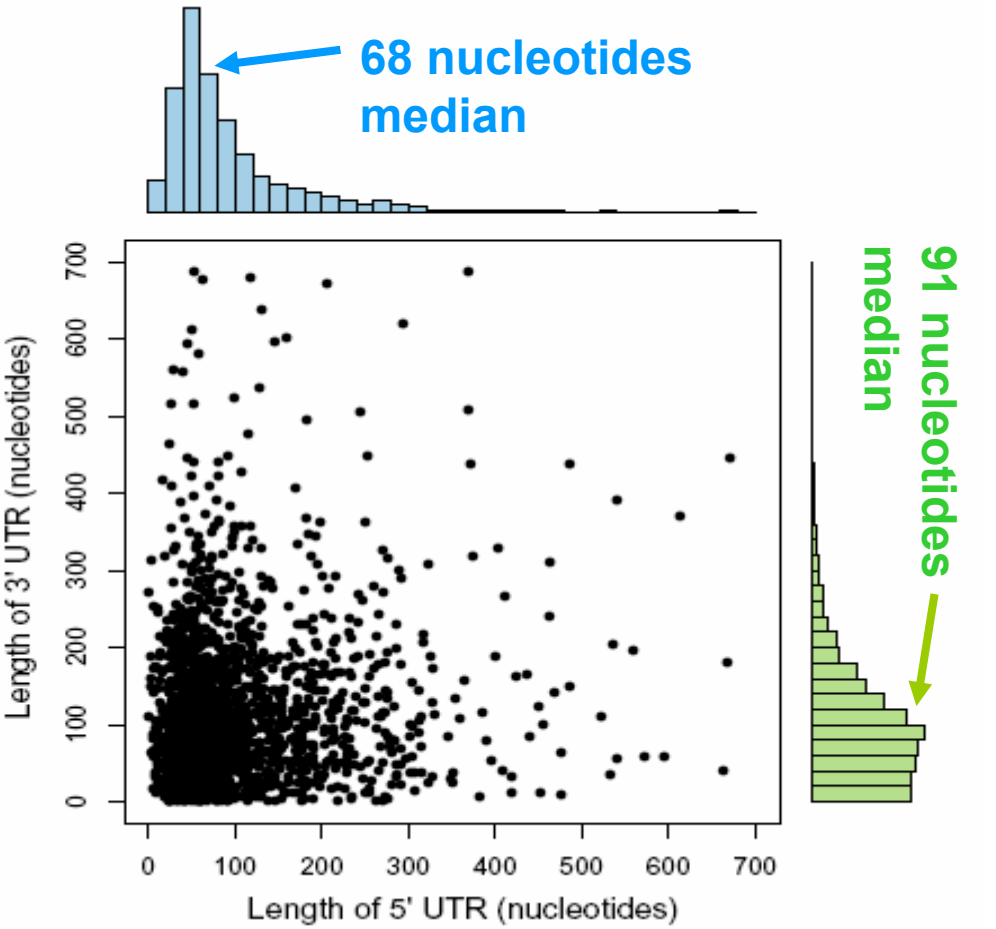
Unprecedented,

strand specific resolution

Mapping of UTRs



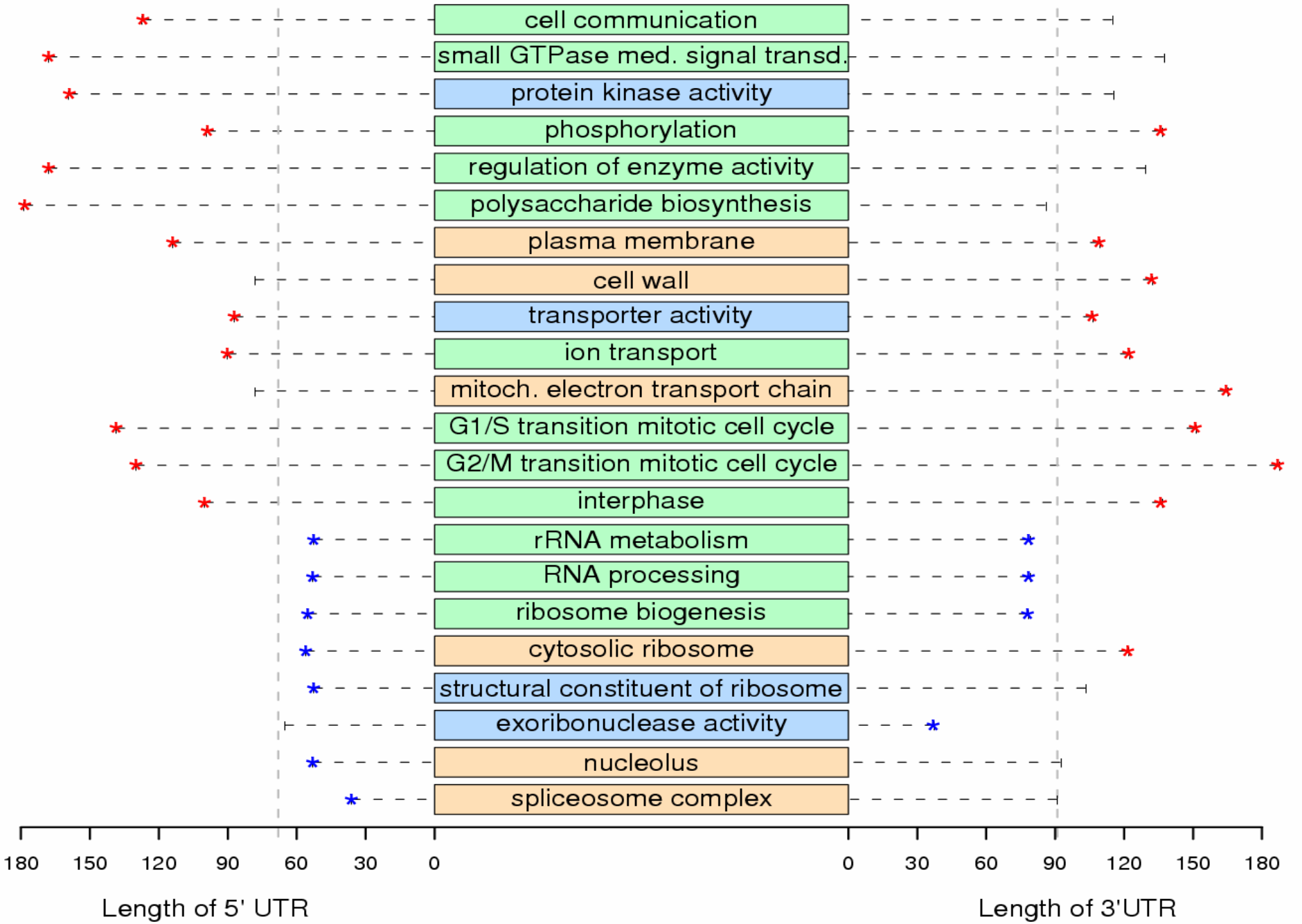
UTR lengths for 2044 ORFs



On average

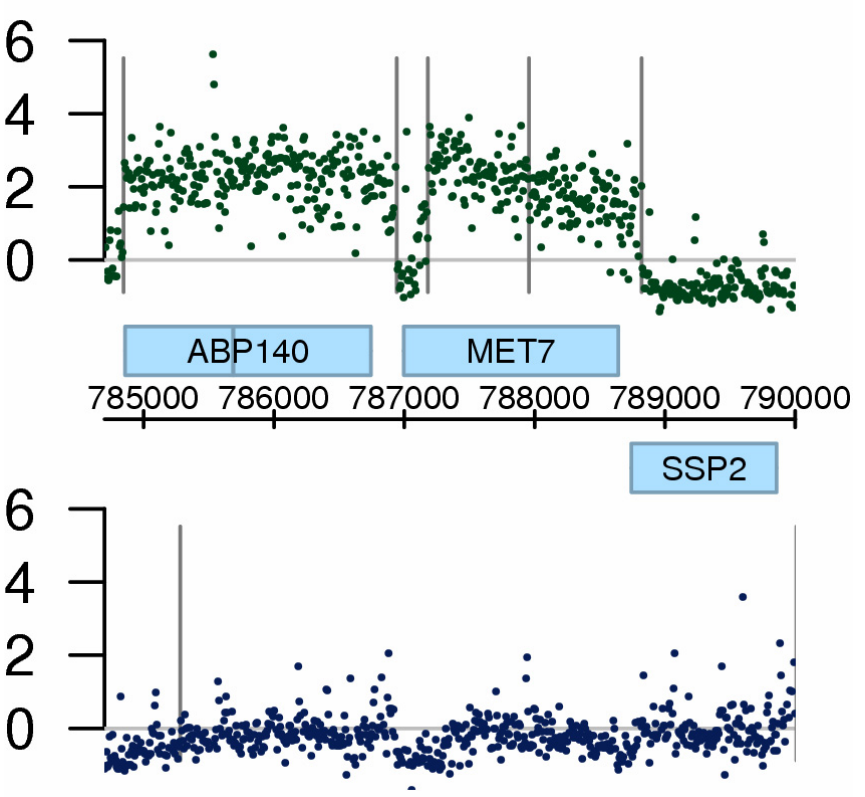
3’ UTRs are longer than 5’ UTRs

No correlation between 3’ and 5’ lengths


cell communication
small GTPase med. signal transd.
protein kinase activity
phosphorylation
regulation of enzyme activity
polysaccharide biosynthesis
plasma membrane
cell wall
transporter activity
ion transport
mitoch. electron transport chain
G1/S transition mitotic cell cycle
G2/M transition mitotic cell cycle
interphase
rRNA metabolism
RNA processing
ribosome biogenesis
cytosolic ribosome
structural constituent of ribosome
exoribonuclease activity
nucleolus
spliceosome complex
180 150 120 90 60 30 0
0 30 60 90 120 150 180
Length of 5' UTR
Length of 3'UTR

Transcriptional architectures

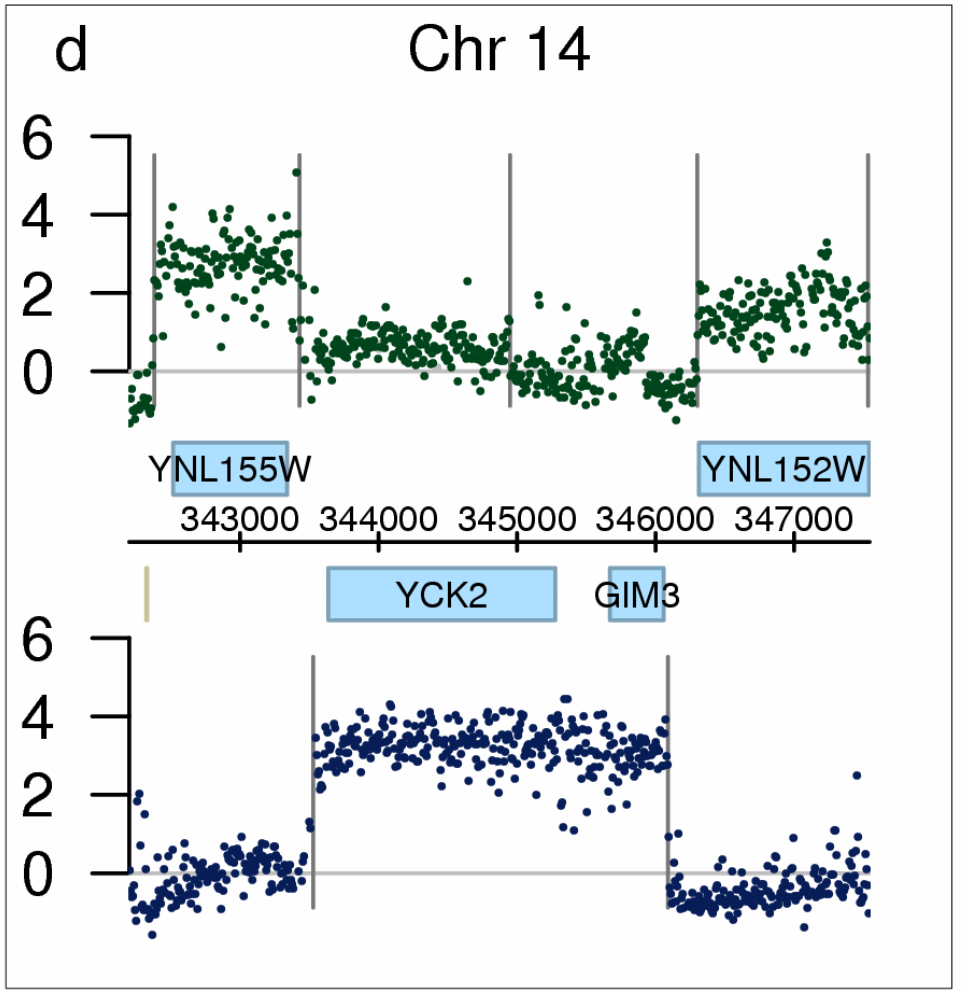
921 ORFs were divided into at least two segments




```
Symbols: * = identical : = strong similarity . = weak similarity

SGD_Scer_MET7/YOR241W   1 MHKGKKNYPNLITSFRMNLKKIILNHDRFSHPERWKTNALLRFTFVYIKF 50
MIT_Sbay_c154_23525       ----------------------------------------------------
MIT_Smik_c485_21053       ----------------------------------------------------
WashU_Sbay_Contig635.23   ----------------------------------------------------
WashU_Scas_Contig640.5    ----------------------------------------------------
WashU_Sklu_Contig2361.2   1 ----------------------------------------------------M 1
Symbols

SGD_Scer_MET7/YOR241W  51 LFDLMIIKNPLRMVGKTYRDAVTALNSLQSNYANIMAIRQTGDRKNTMTL 100
MIT_Sbay_c154_23525     1 ----MLIRNPLRMIGKTYHDAVTALNSLQSNYANIMAIRQTGDRKNIMTL  46
MIT_Smik_c485_21053     1 ----MIIRNPLRMVGKTYHDAVTALNSLQSNYANIMAIRQTGDRKNIMTL  46
WashU_Sbay_Contig635.23 1 ----MLIRNPLRMIGKTYHDAVTALNSLQSNYANIMAIRQTGDRKNIMTL  46
WashU_Scas_Contig640.5  1 ----MRLTIPLKMSTKTYRDAINSLNSLQSNYANIMAIRESGDRKNMMNI  46
WashU_Sklu_Contig2361.2 2 RFTFPLKMSITSSTKRTYQDAVTALNSLQSNYANIMAIRASGDRKNMMNI  51
Symbols                       :**:**:.:*************** :***** *.:

SGD_Scer_MET7/YOR241W  101 LEMHEWSRRIGYSASDFNKLNIVHITGTKGKGSTAAFTSSILGQYKEQLP 150
MIT_Sbay_c154_23525     47 LEMHEWSRRIGYSSSDFNKLNIVHITGTKGKGSTAAFTSSILGQYKEQLP  96
MIT_Smik_c485_21053     47 LEMHEWSRRIGYAASDFNKLNIVHITGTKGKGSTAAFTSSILGQYKEKLP  96
WashU_Sbay_Contig635.23 47 LEMHEWSRRIGYSSSDFNKLNIVHITGTKGKGSTAAFTSSILGQYKEQLP  96
WashU_Scas_Contig640.5  47 WEMKEWSRRIGYNVSEFNKLNIIHITGTKGKGSTAAFTSSILNQYKEKLP  96
WashU_Sklu_Contig2361.2 52 WEMQEWSRRIGYSTKDYNKLNIIHITGTKGKGSTAAFTQSILSQYNDRLS 101
Symbols                     **:********  .::******:***************.***.**:::*.
```

Operon-like structures

123 segments contained ORFs of more than one protein-coding gene

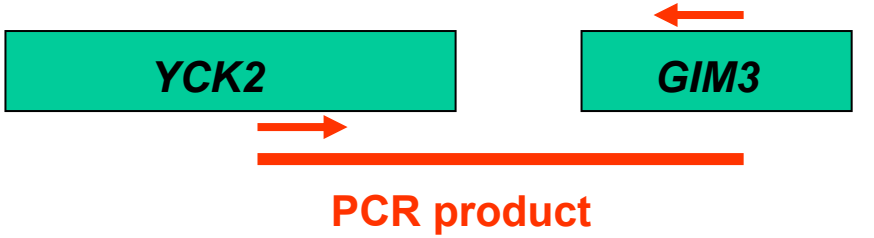


YCK2

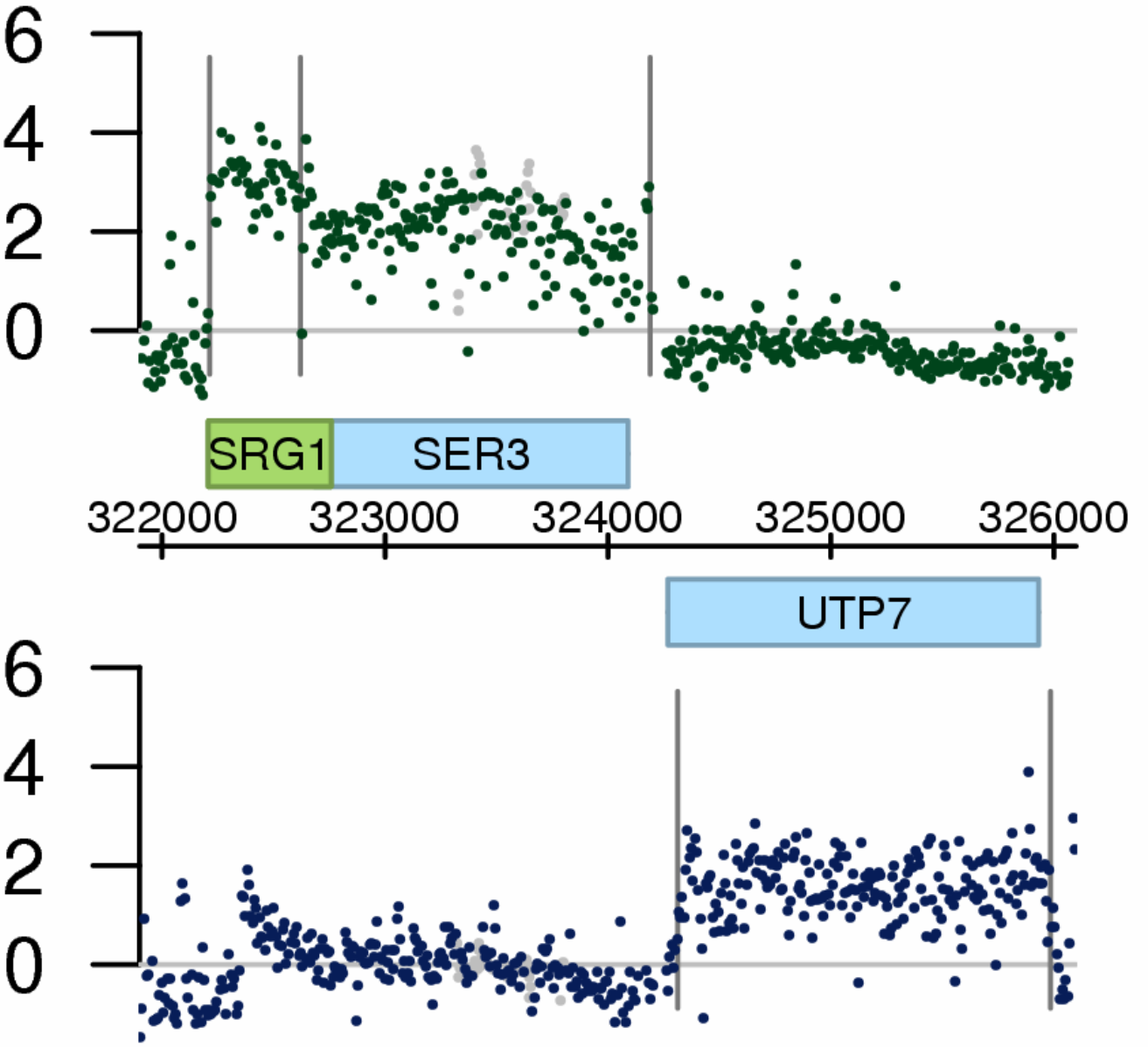
casein kinase I, involved in cytokinesis

GIM3

tubulin binding, involved in microtubule biogenesis

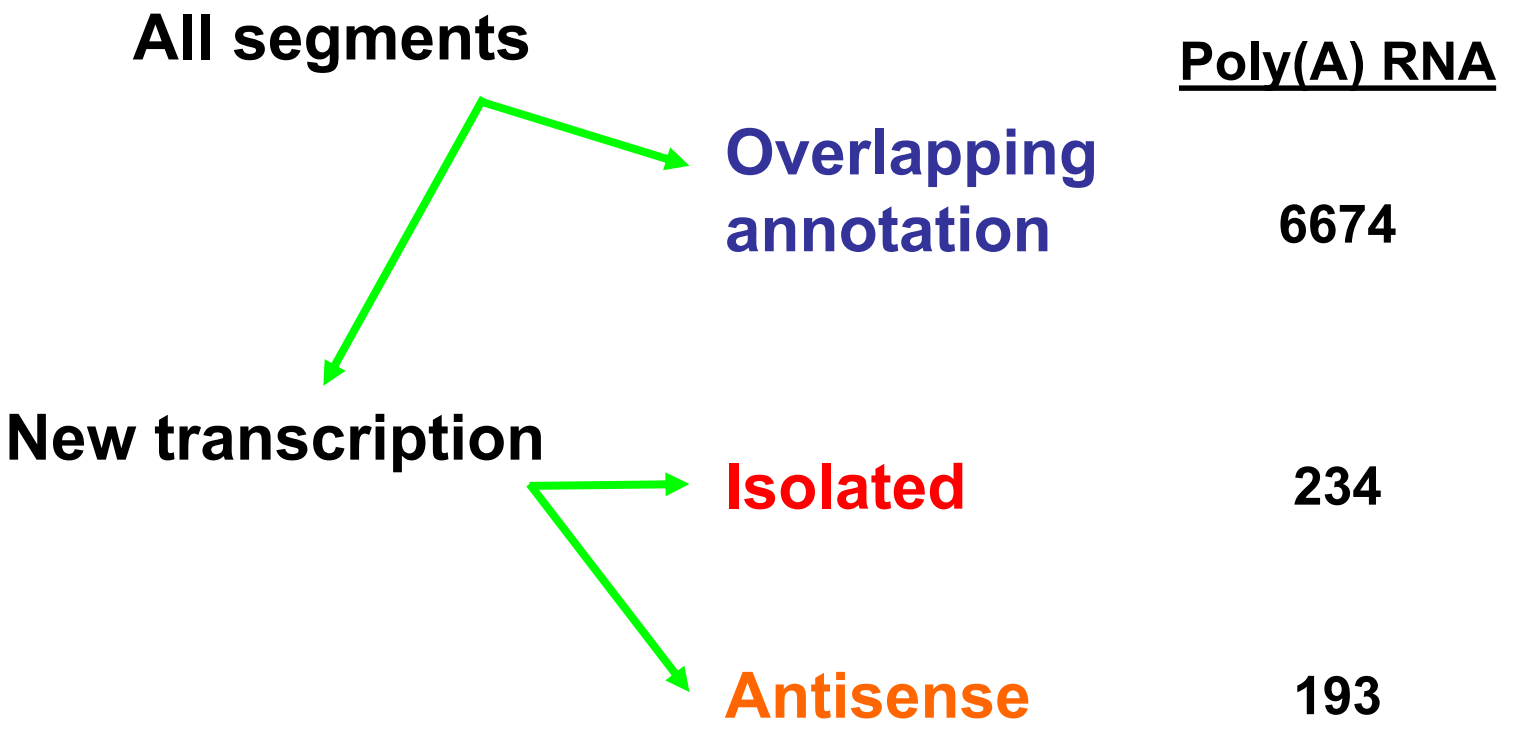


Transcription over active promoters

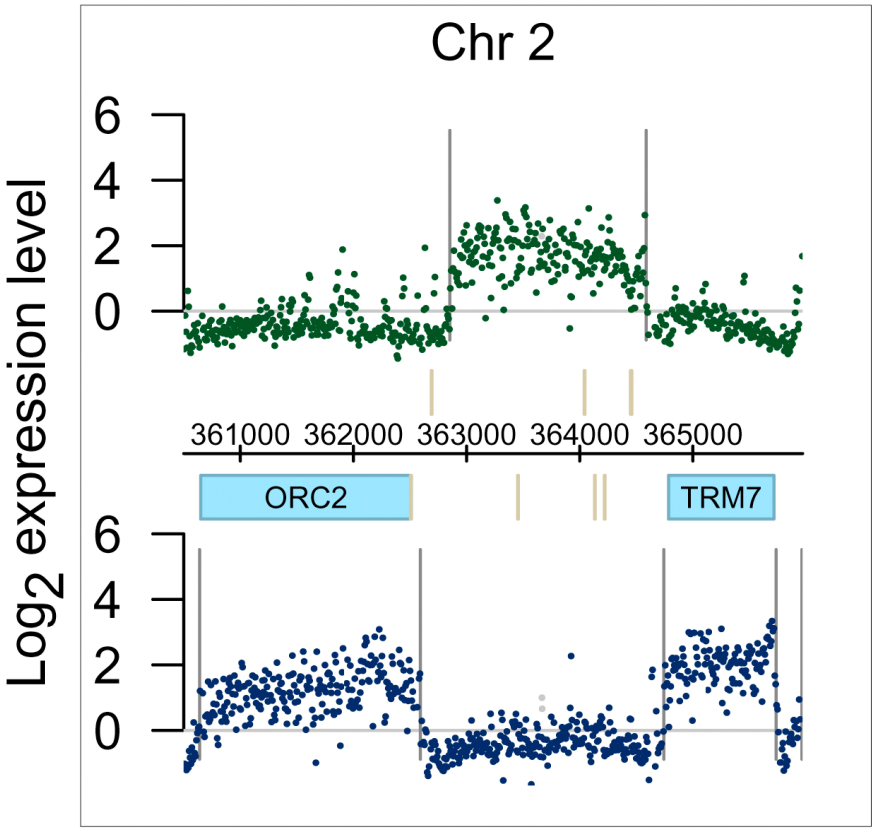


Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature* **429,** 571-574 (2004).

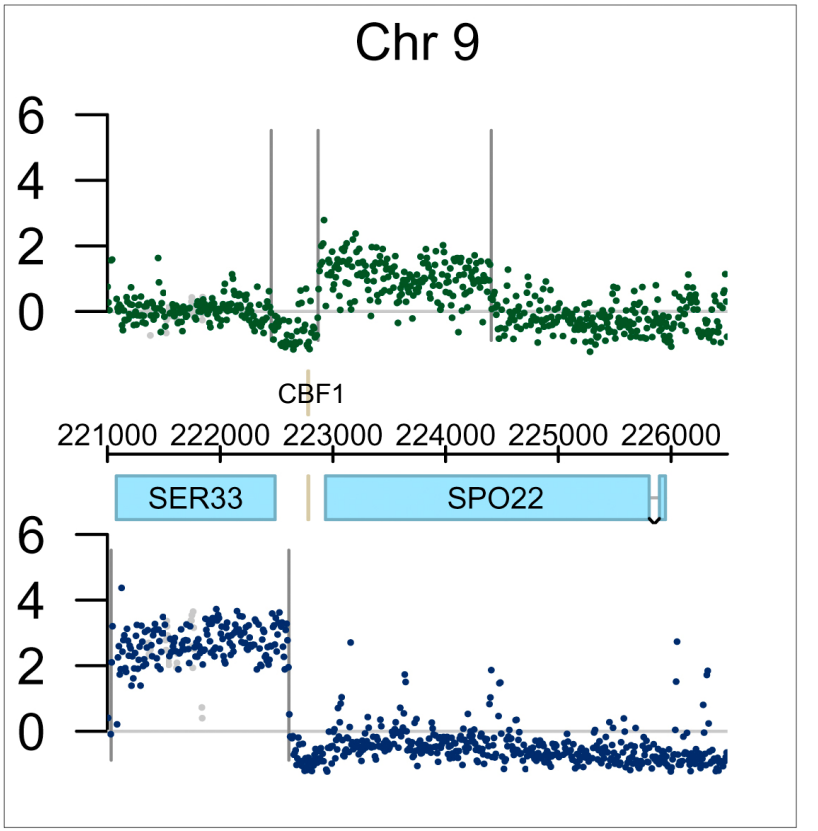
Looking for New Transcripts



Isolated



Antisense



***CBF1*: important for growth in rich media**

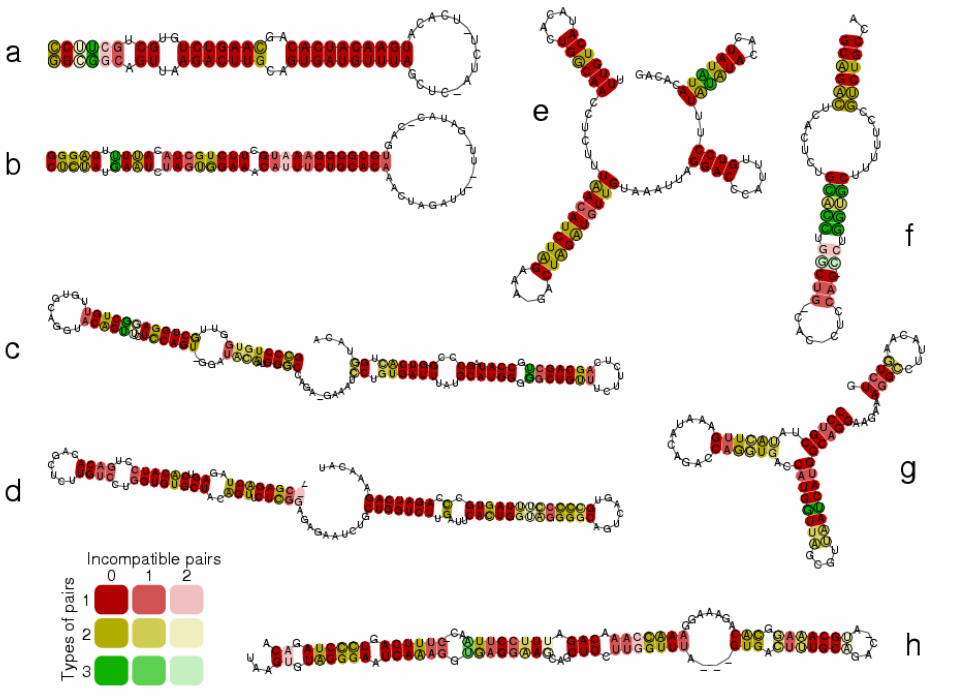
GO of Genes with antisense: Cell wall, transcriptional regulation, meiotic cell cycle…

Novel transcripts

Sequence conservation (with other yeast species) not more than for other intergenic sequence

No codon signature (3-periodicity of mutation frequencies)

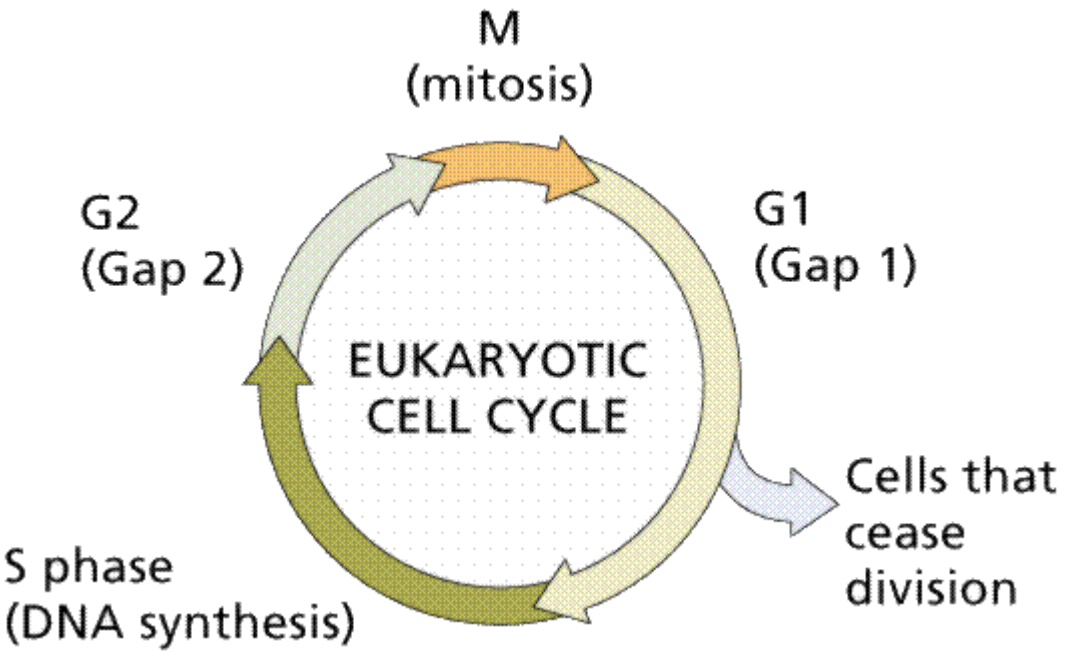
But: conservation of predicted RNA secondary structures



with Lee Bofkin, Nick Goldman

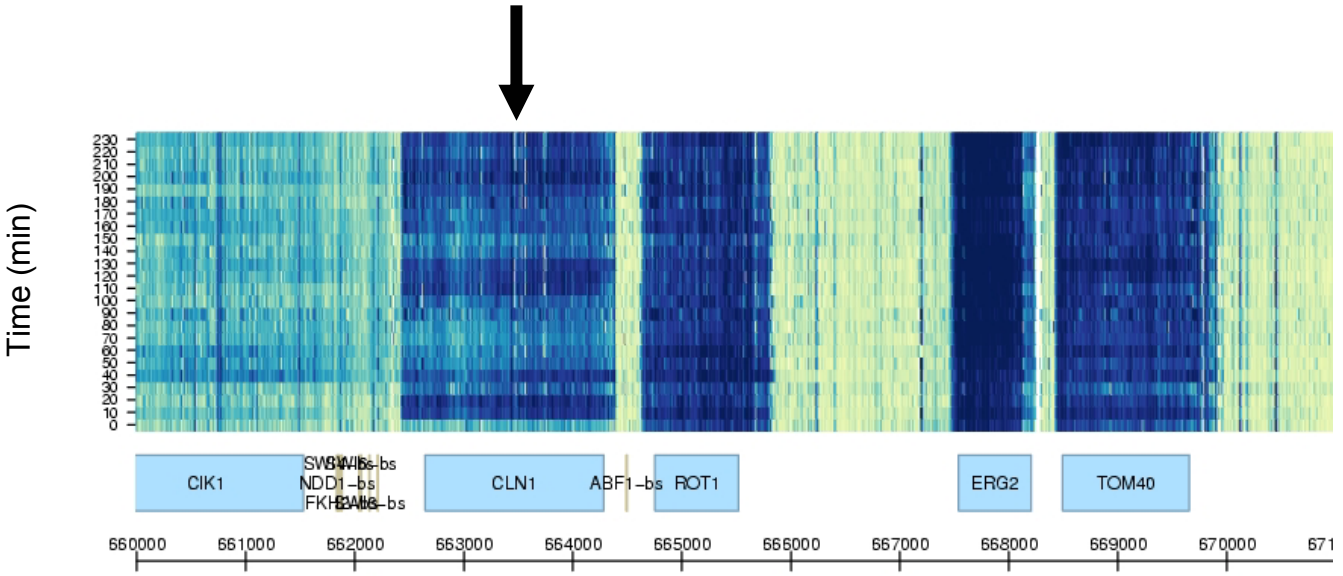
Stephan Steigele, Kay Nieselt
Peter Stadler

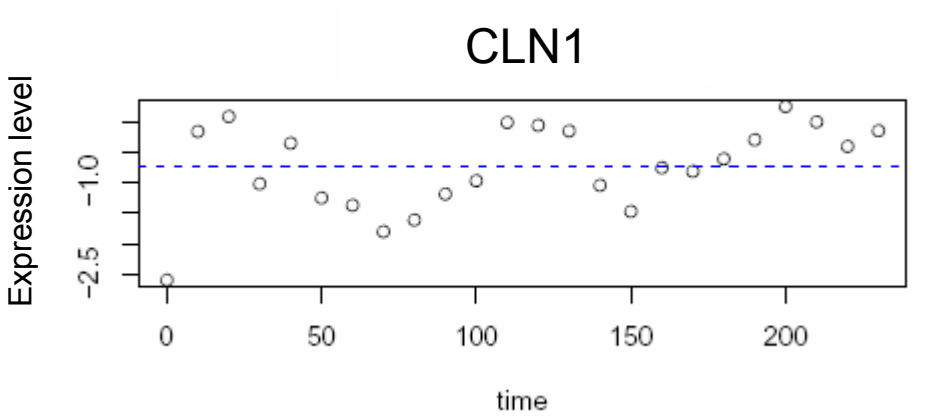
Cell Cycle

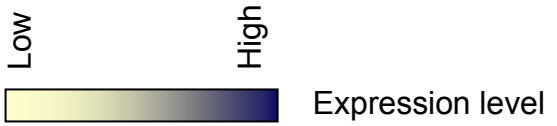


Temperature sensitive cdc28 – arrest at G1
Monitored at 10 min intervals for 230 min in total
(~3 cell cycles)

Cycling of known transcript

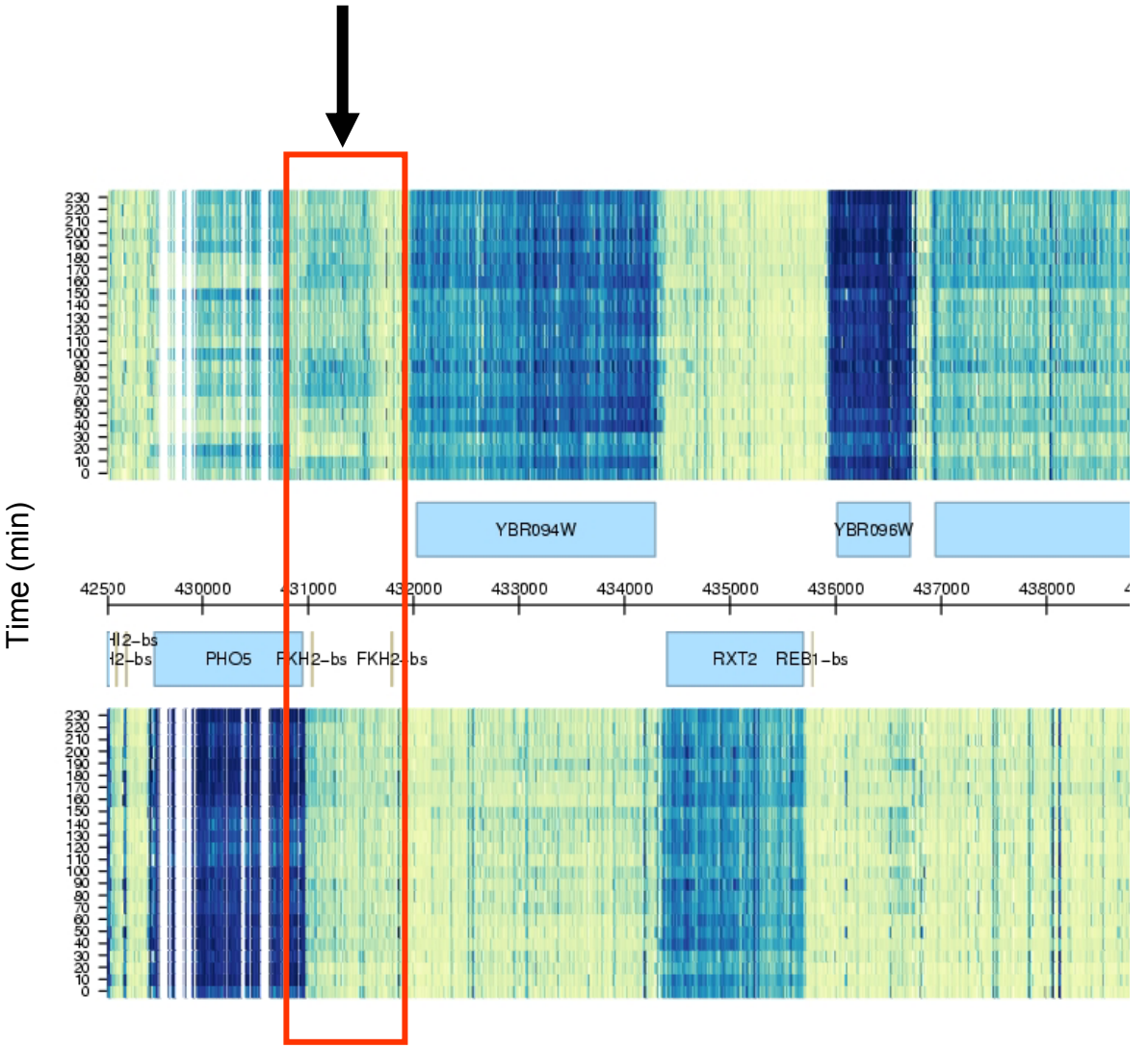


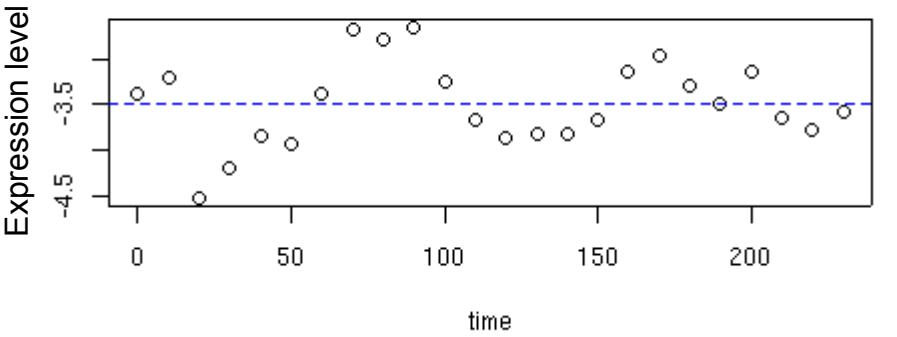




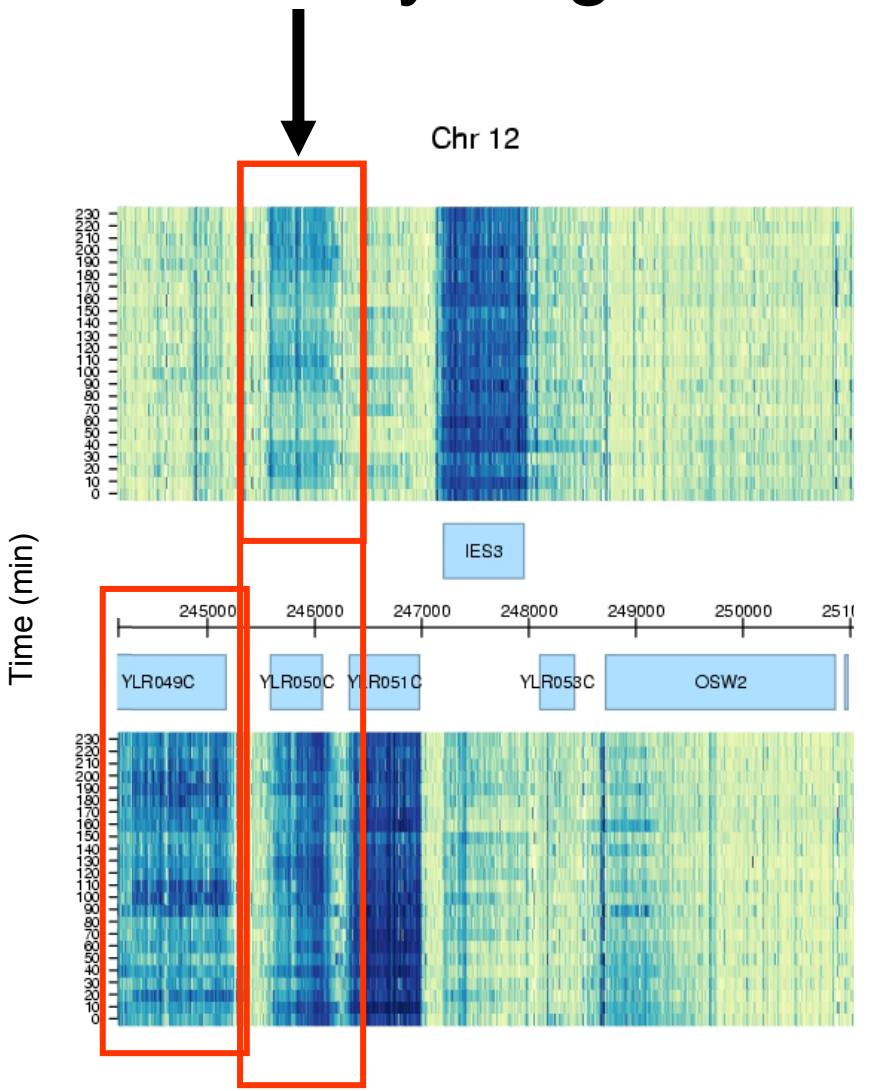
G1 cyclin involved in regulation of the cell cycle; activates Cdc28p kinase to promote the G1 to S phase transition

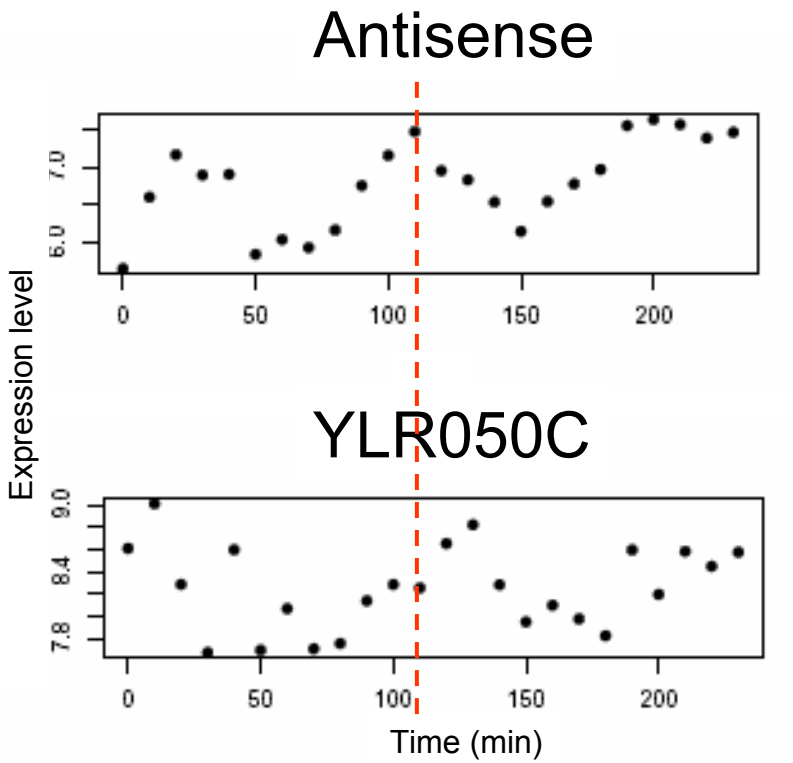
Cycling of novel transcript





Cycling of antisense transcript





RNA mediated regulation

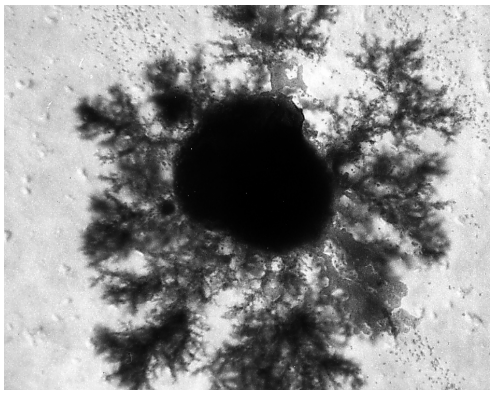
- **UTR lengths associated with function, localization, regulation**
- **Antisense found predominantly to 3' UTRs and longer UTRs**
- **Antisense correlated with GO categories**
- **Similar to patterns for miRNAs in other species**

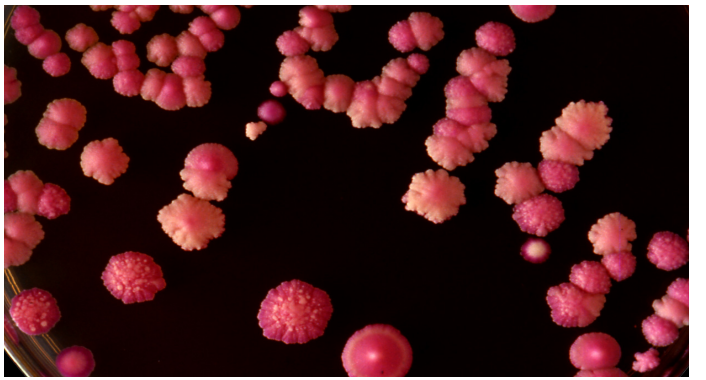
Suggests a functional role for antisense in *S. cerevisiae*

A Clinical Isolate of *S. cerevisiae*: YJM789

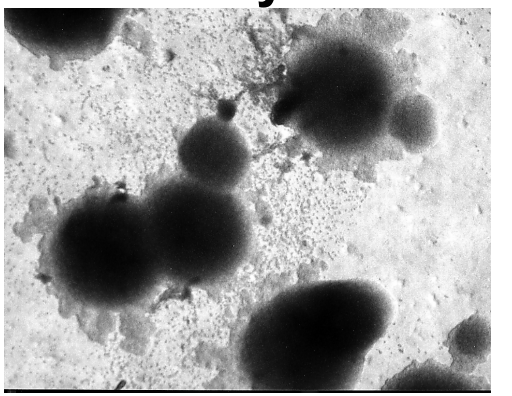
- Isolated from lung of an AIDS patient
- Pathogenic in mouse model
- Forms pseudohyphae; undergoes colony-morphology switching
- Able to grow at 42ºC

YJM789

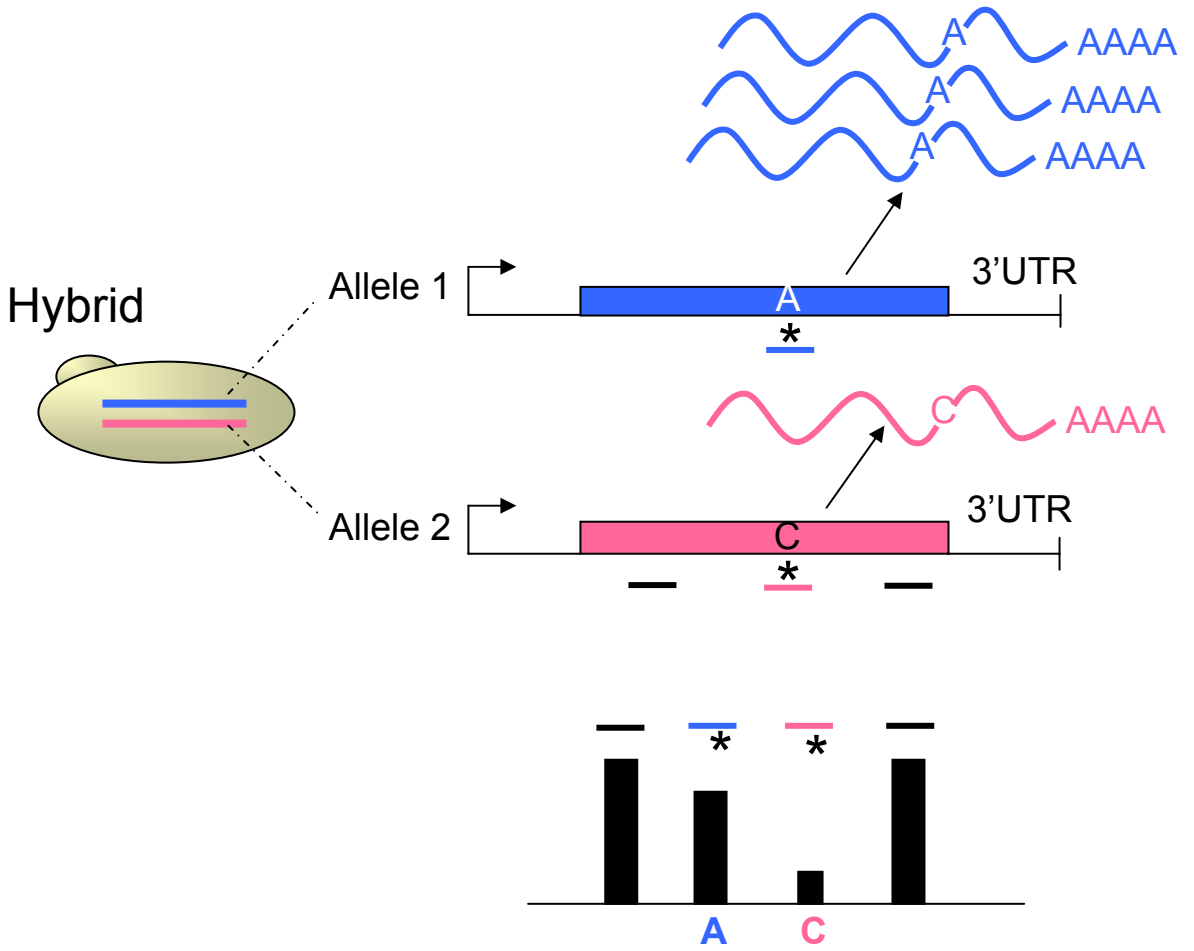




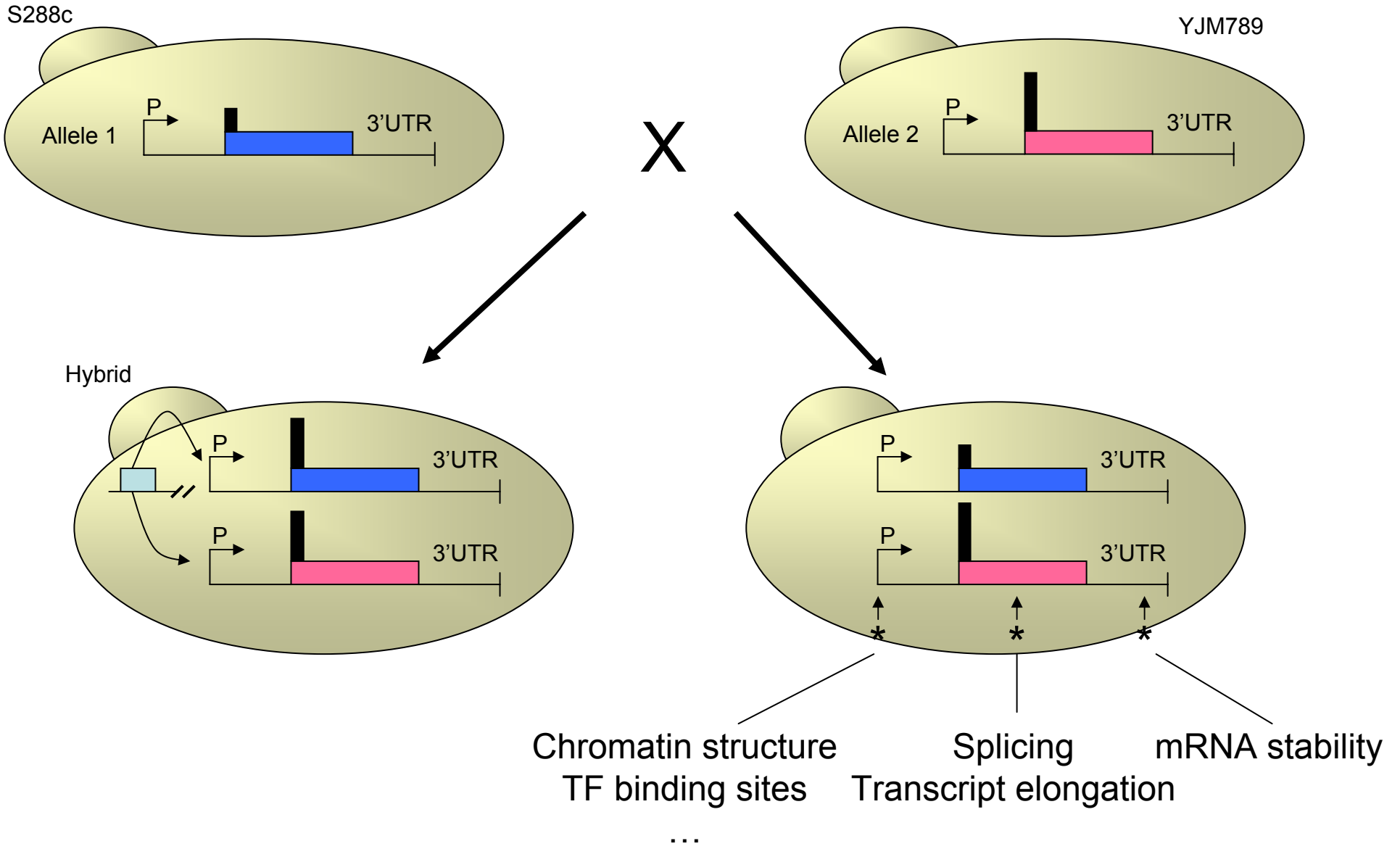
Laboratory strain



Two Genomes on One Array: Quantifying Allelic Transcription

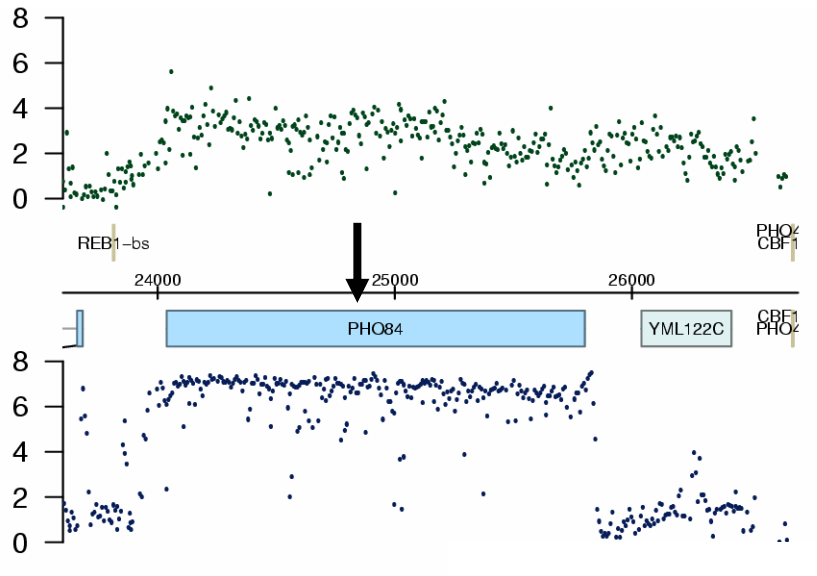


Local vs distant QTL
..*trans* vs. *cis* regulation

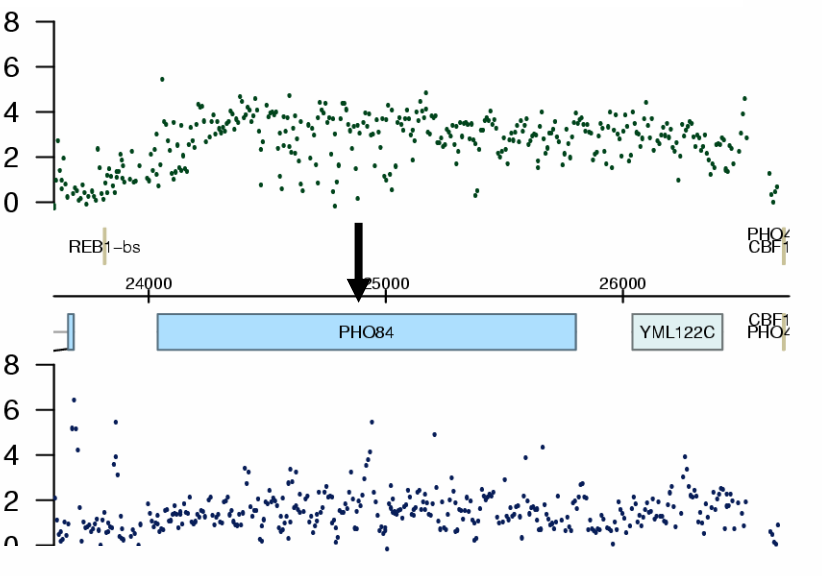


PHO84 Allele-specific Expression

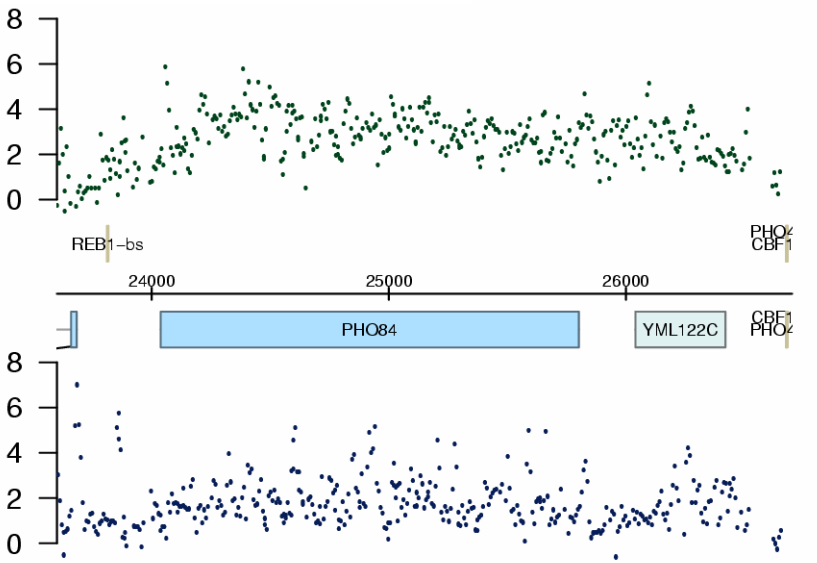
S288c

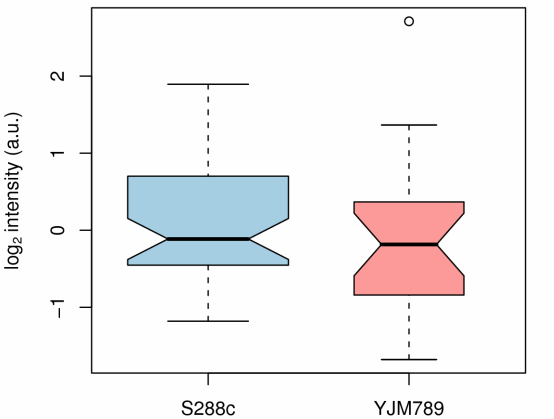


YJM789



Hybrid (S288c/YJM789)



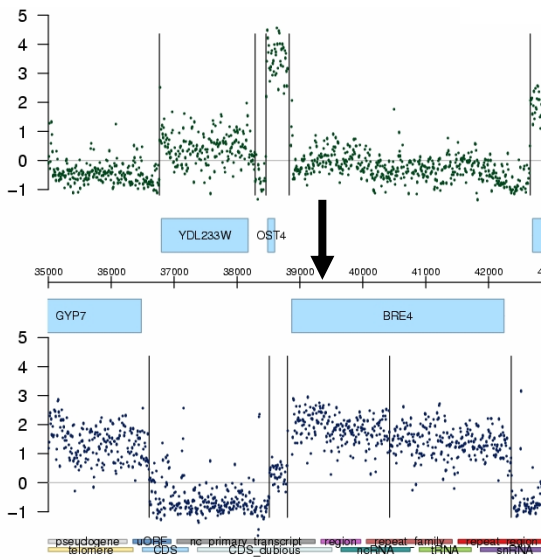


Both alleles off
=> *Distant QTL*

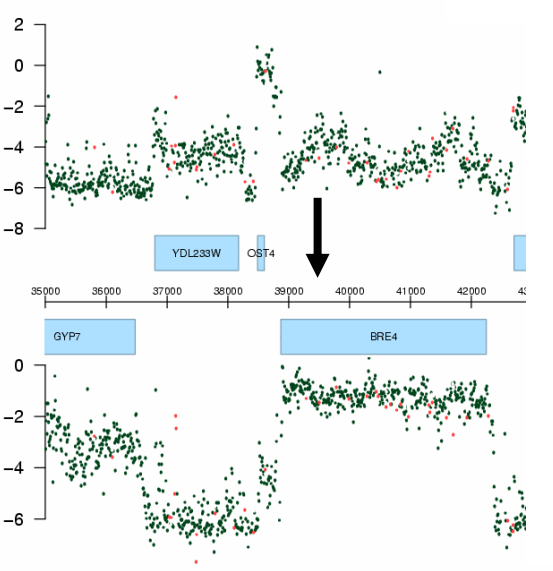
PHO84 – inorganic phosphate (Pi) transporter

BRE4 Allele-specific Expression

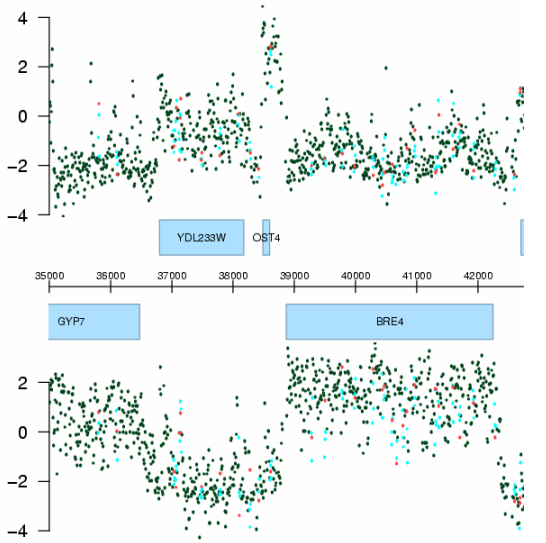
S288c

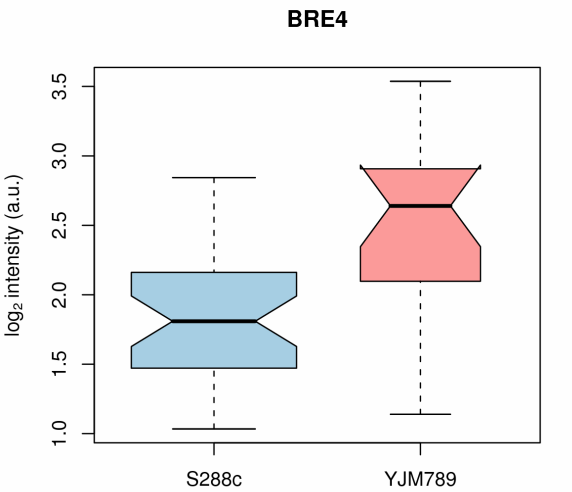


YJM789



Hybrid (S288c/YJM789)

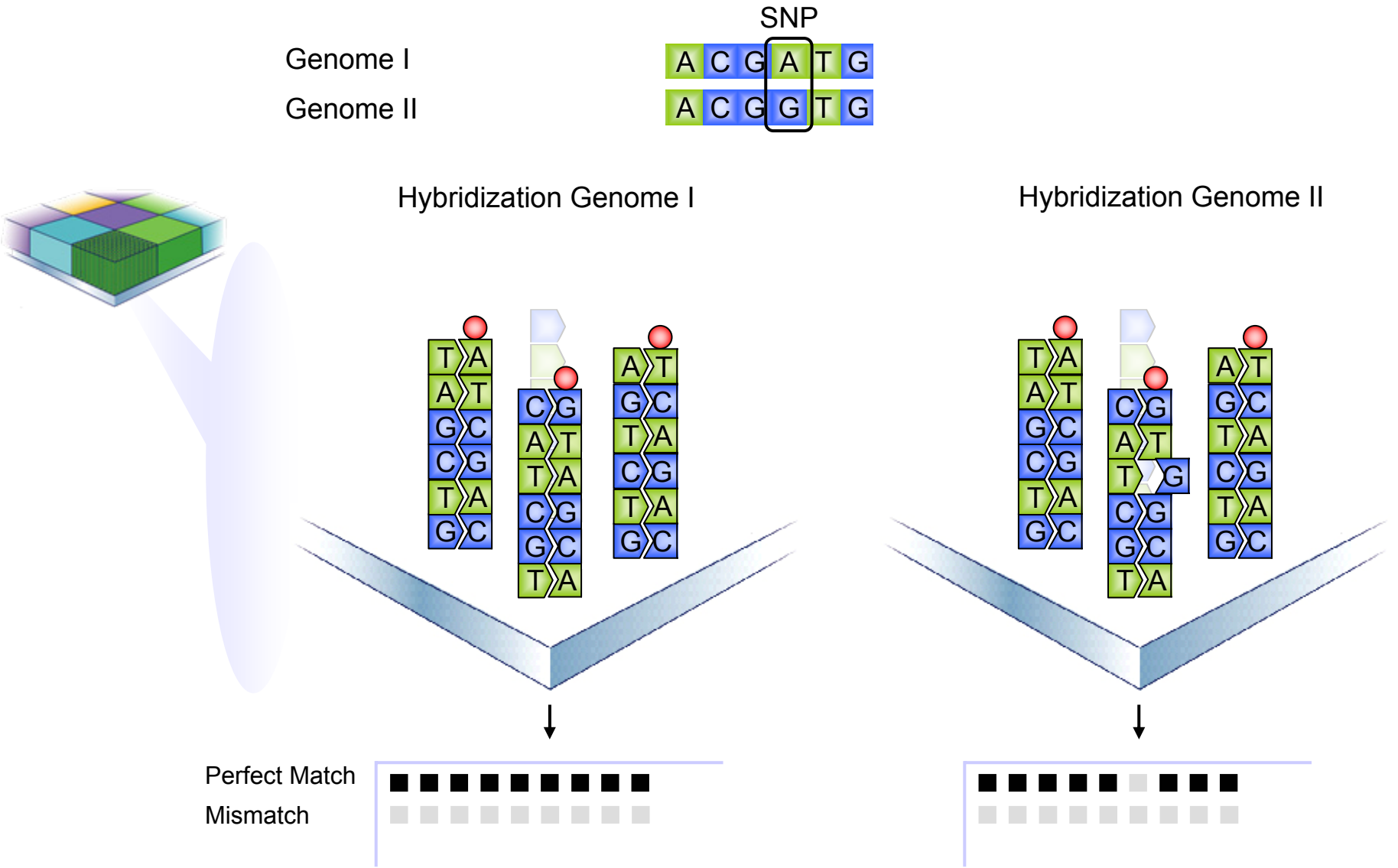




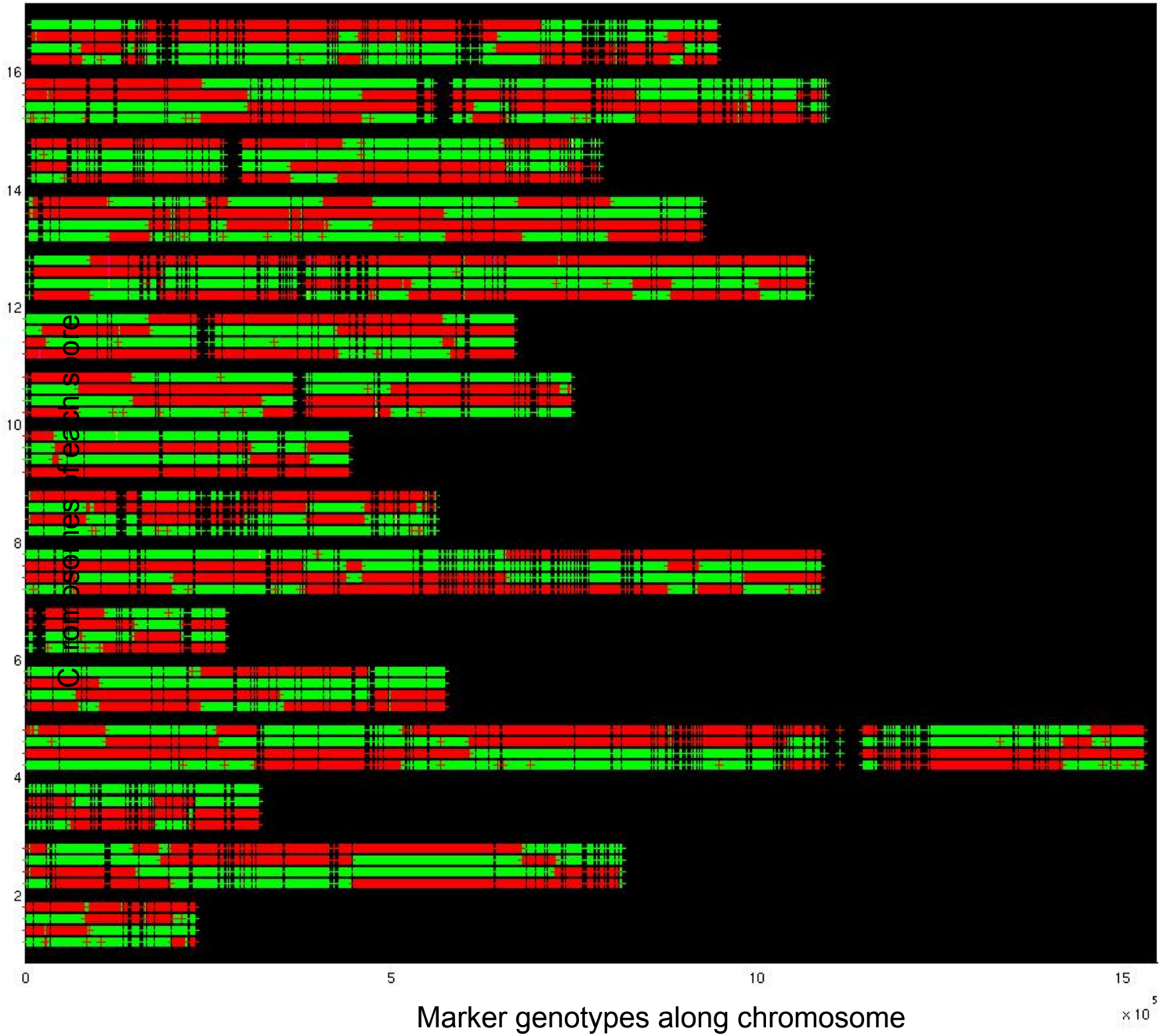
YJM789 higher
=> *cis*

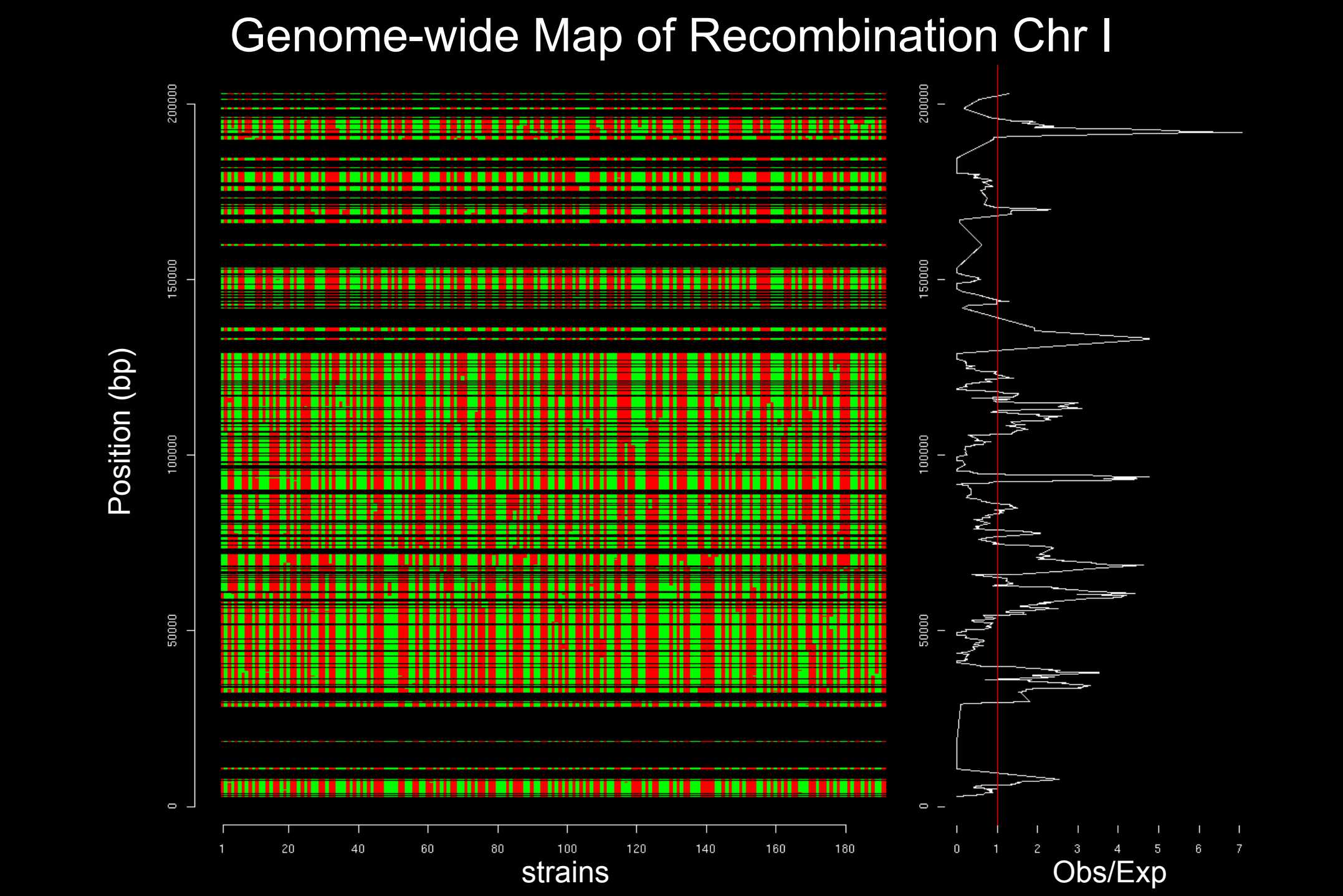
BRE4 – unknown, mutant sensitive to drug Brefeldin A

Genotyping with Microarrays



Segregation of 50,000 Markers in a Tetrad




Genome-wide Map of Recombination Chr I
Position (bp)
0
50000
100000
150000
200000
1
20
40
60
80
100
120
140
160
180
strains
0
1
2
3
4
5
6
7
Obs/Exp

Conclusion

- Transcriptional complexity goes far beyond current annotation
- Hundreds of antisense transcripts found
- Allelic variation in transcription detected
- Map of recombination breakpoints

Acknowledgements

Lars Steinmetz — **EMBL Heidelberg &**

Lior David — **Stanford Genome Tech. Center**

Eugenio Mancera, Fabiana Perocchi, Sandra Clauder-Münster

Marina Granovskaia — **EMBL Heidelberg**

Richard Bourgon, Paul McGettigan, Matt Ritchie, Jörn Tödling,

BIOCONDUCTOR

Robert Gentleman — **Ben Bolstad**

Vince Carey — **Paul Murrell**

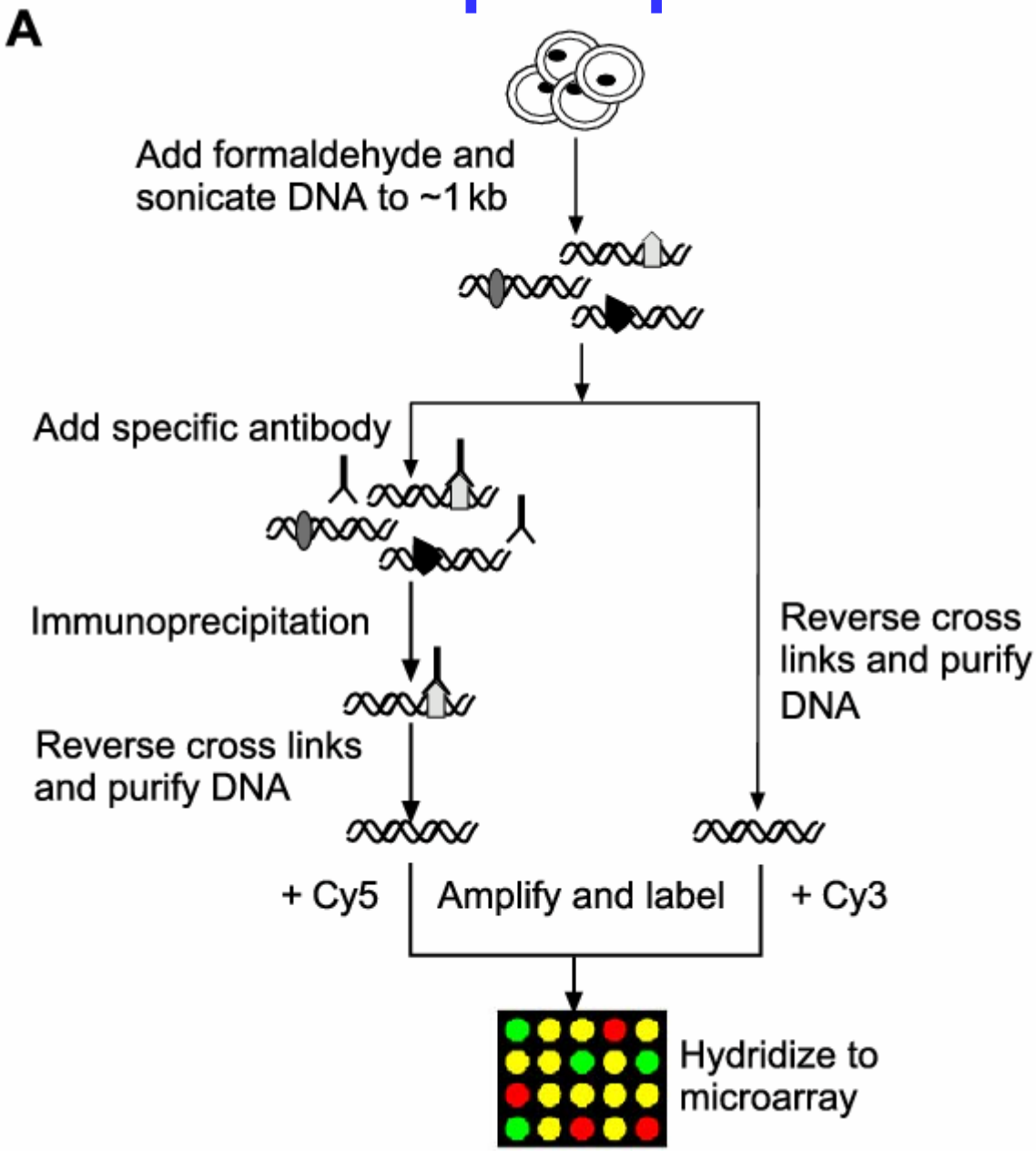
Rafael Irizarry — **Achim Zeileis**

Analysis of ChIP-chip experiments

NGFN Course 2006

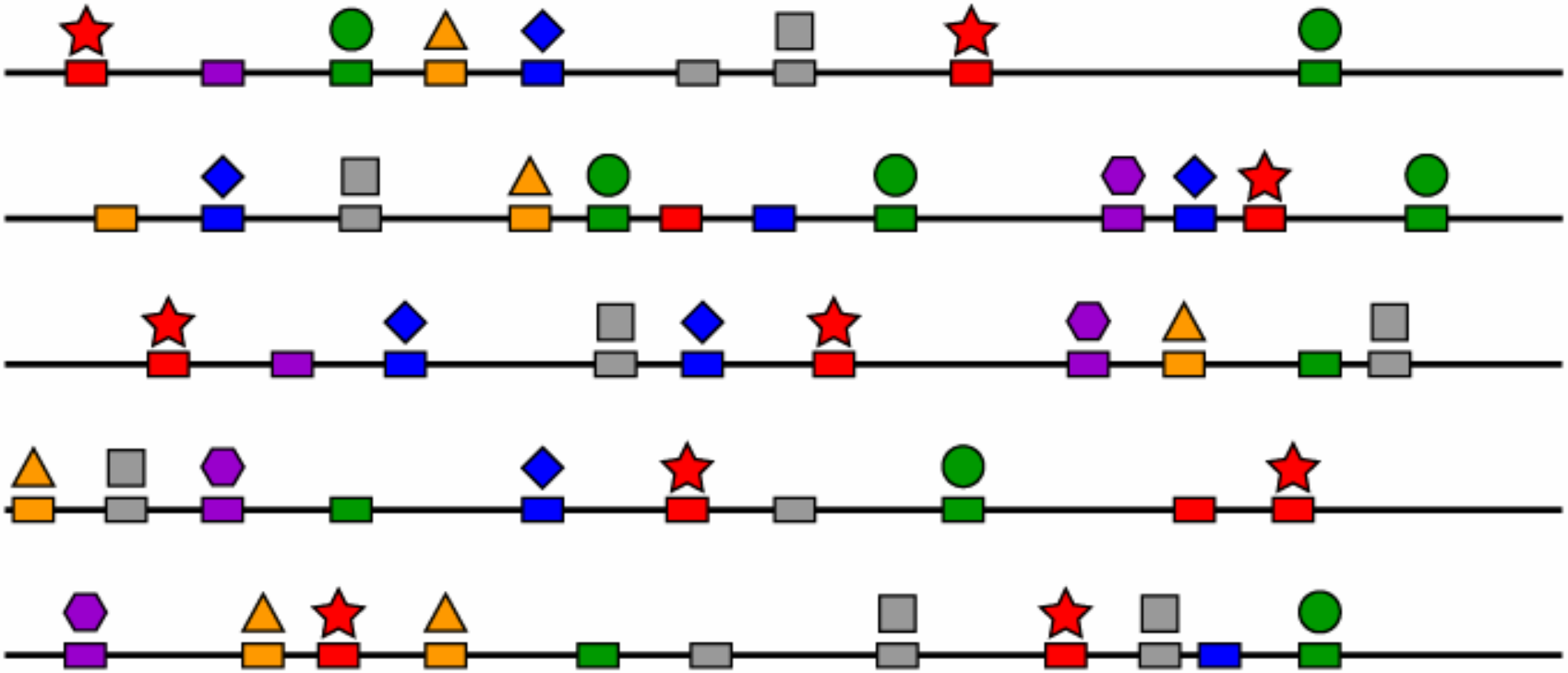
Wolfgang Huber EMBL/EBI

Chromatin immunoprecipation and DNA-chip

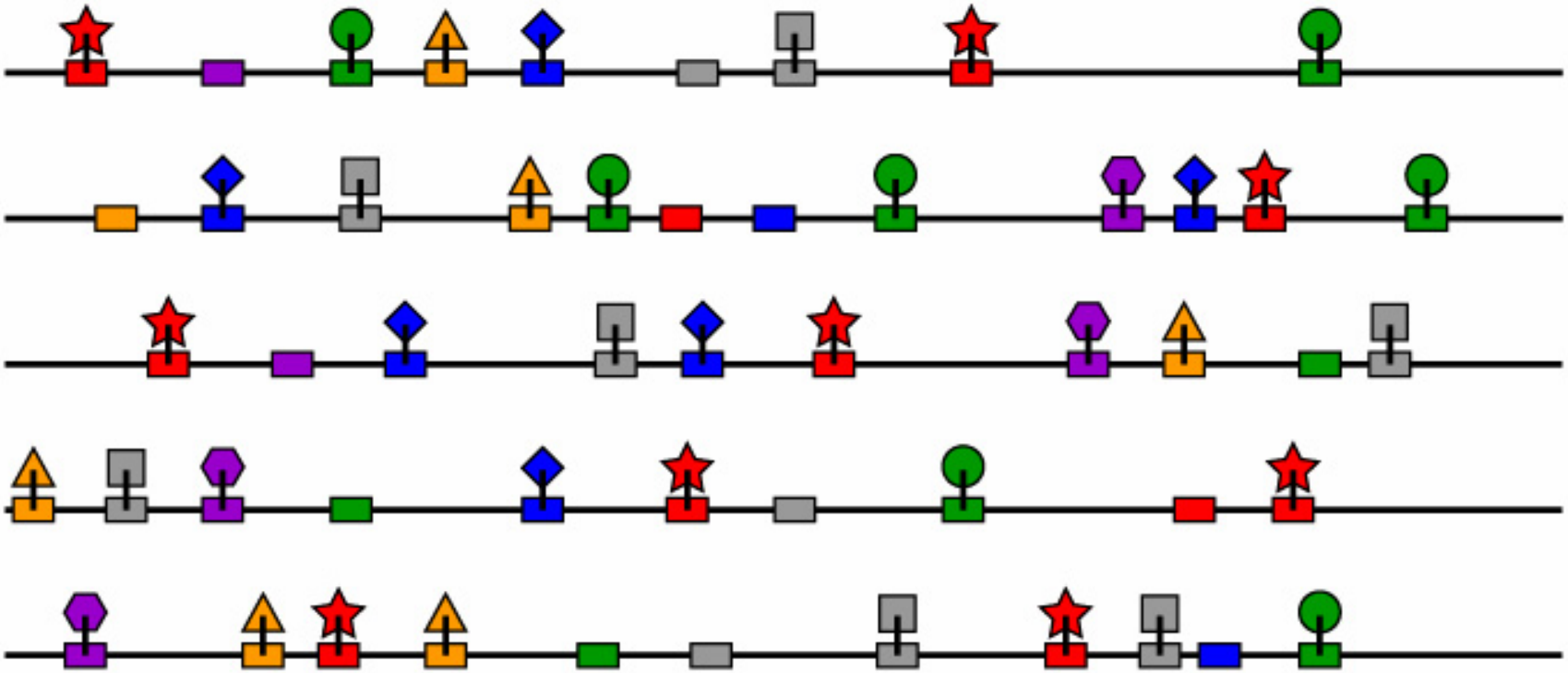


MJ Buck and JD Lieb, Genomics 83 (2004)

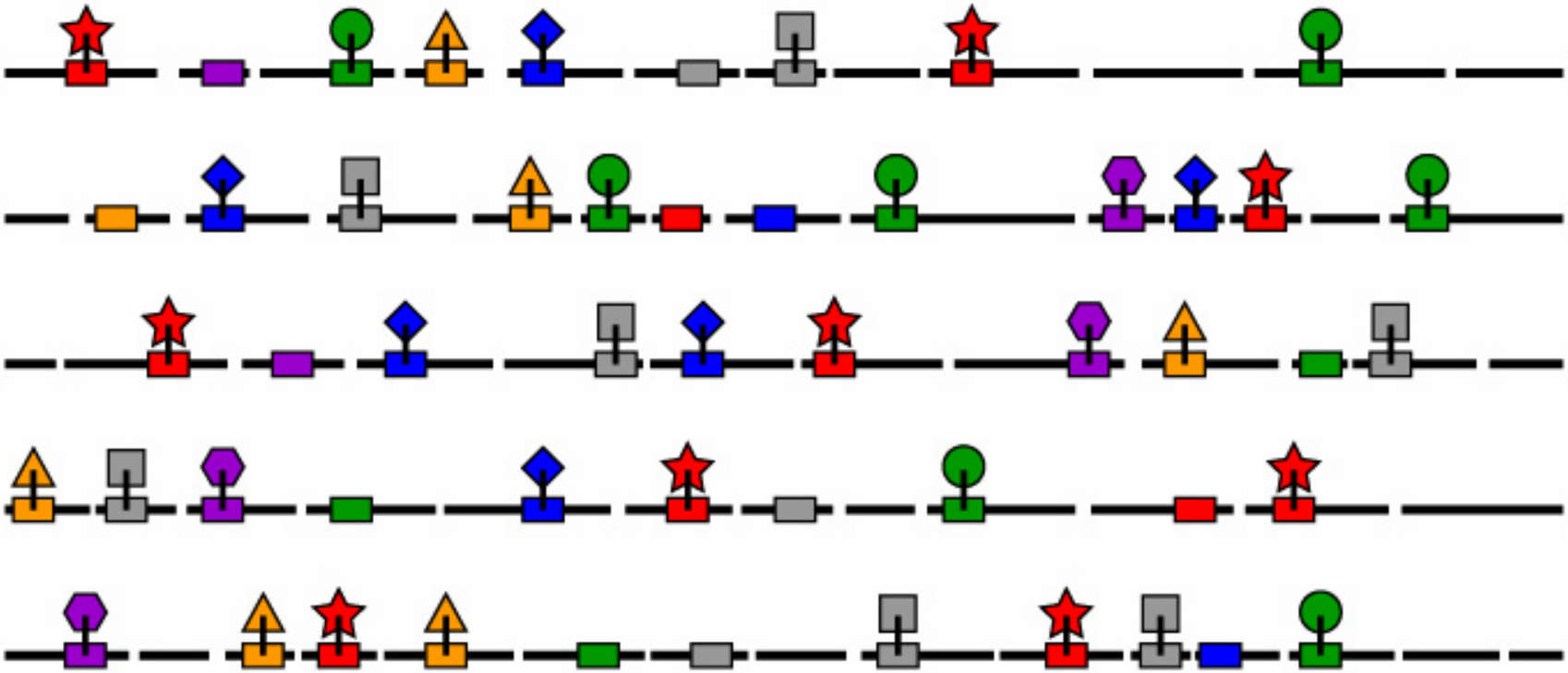
Chromatin immunoprecipitation (ChIP)



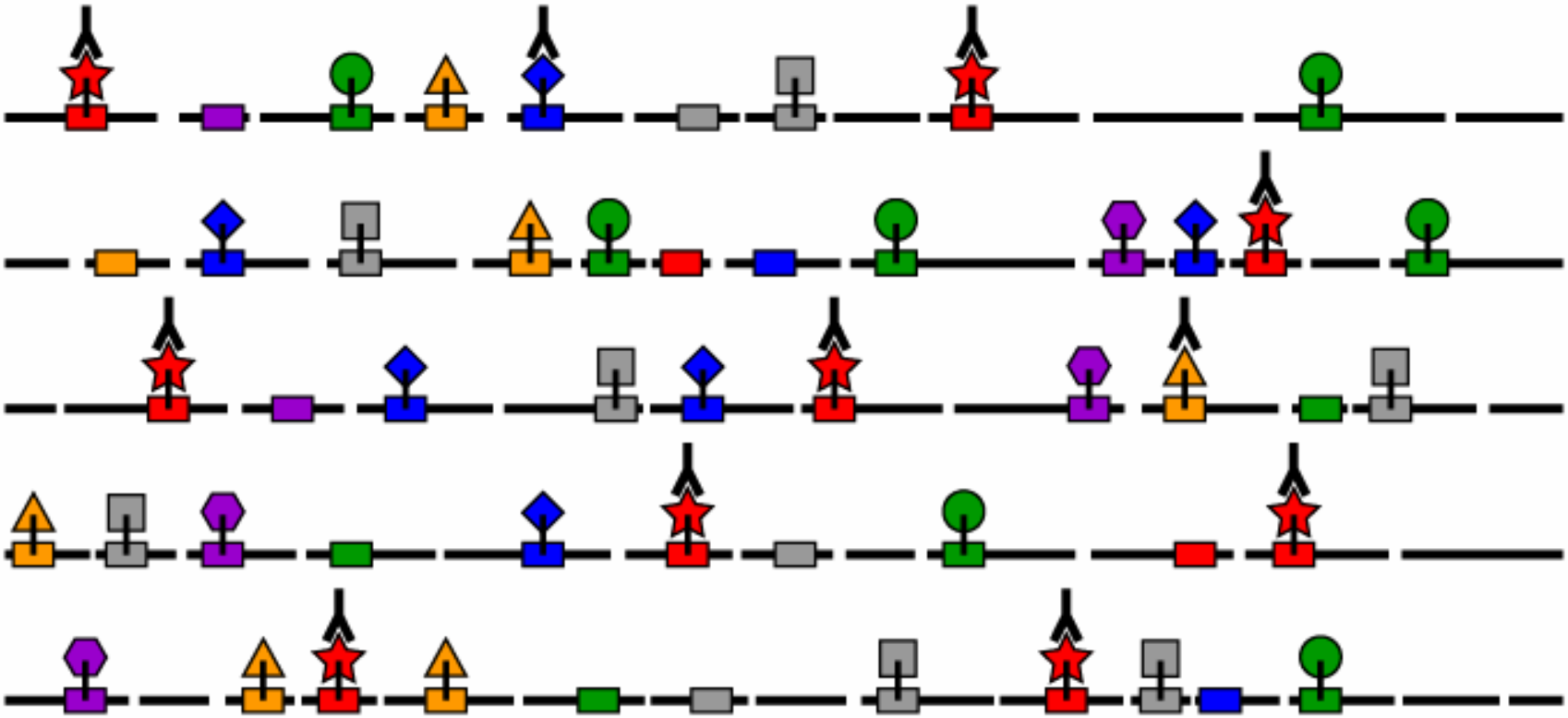
TF/DNA crosslinking *in vivo*



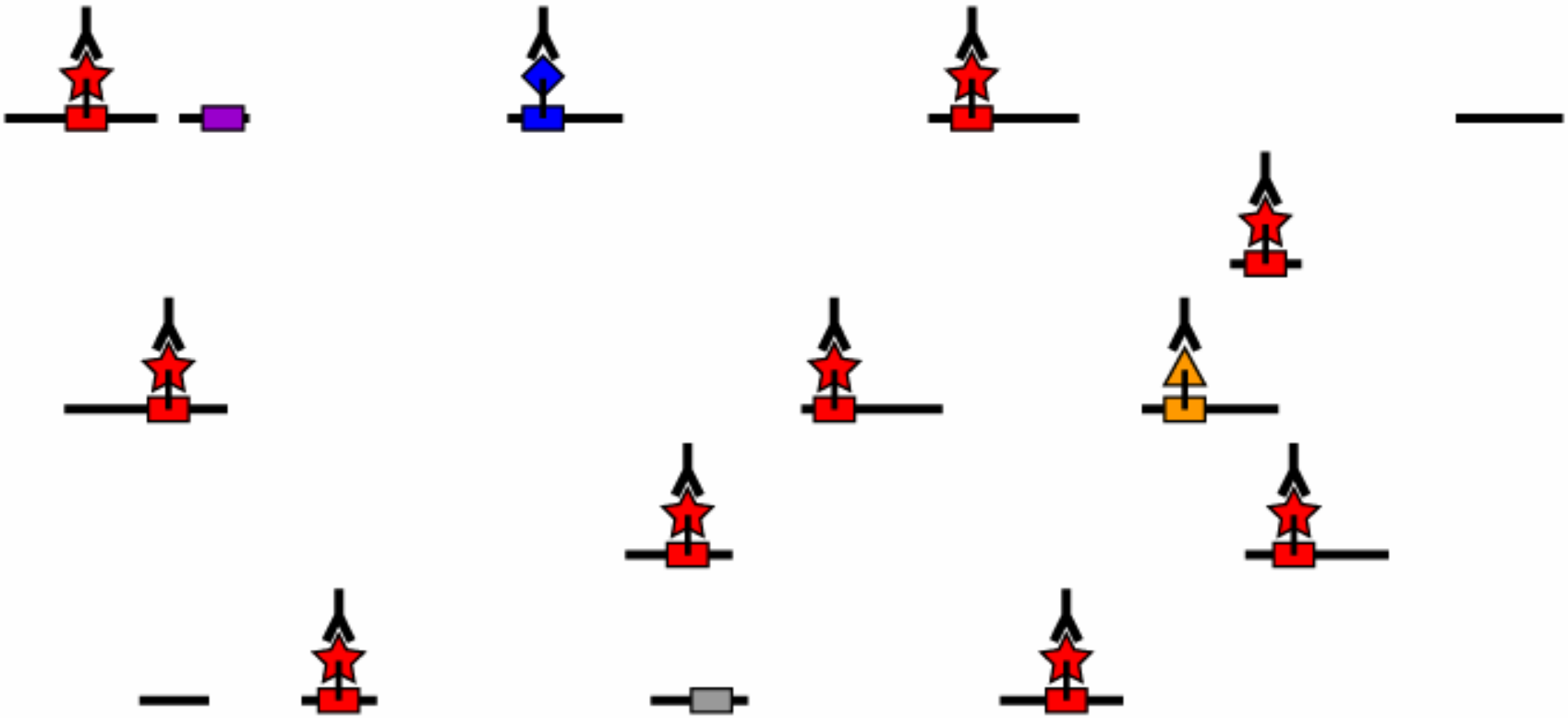
Sonication



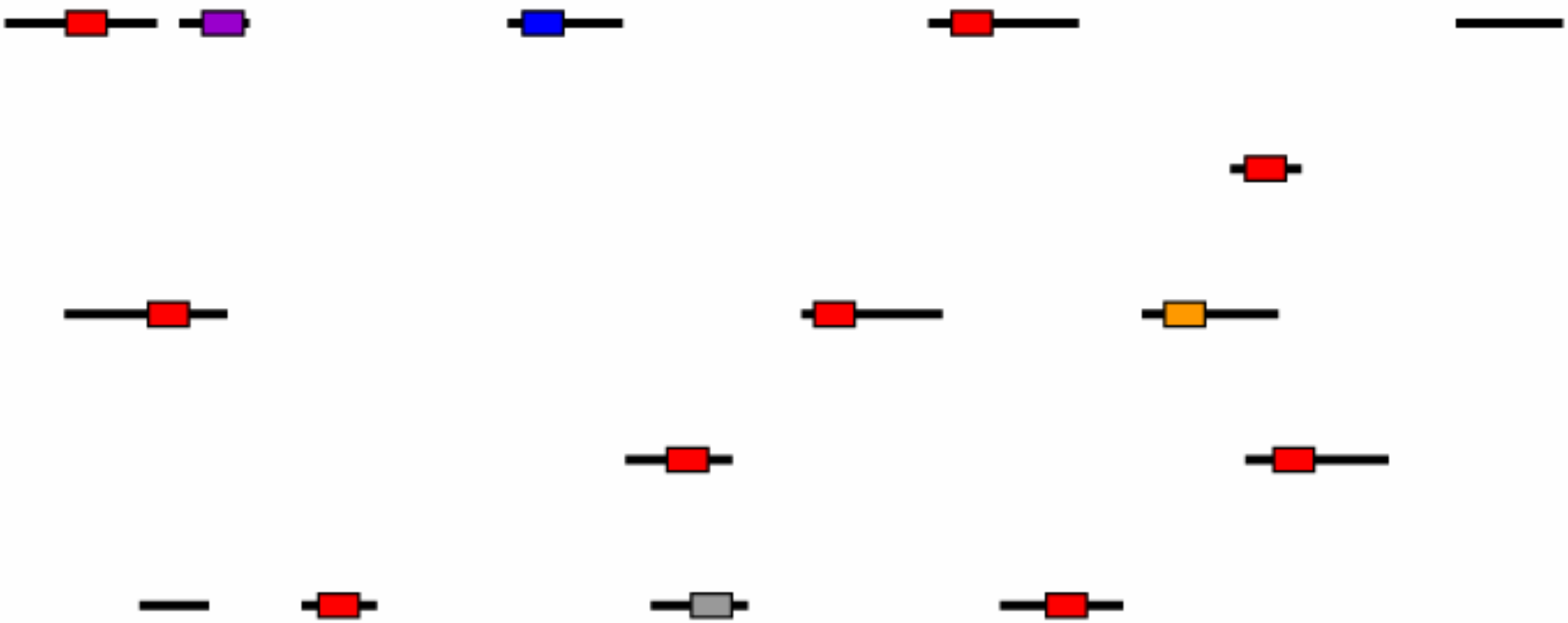
TF-specific antibody



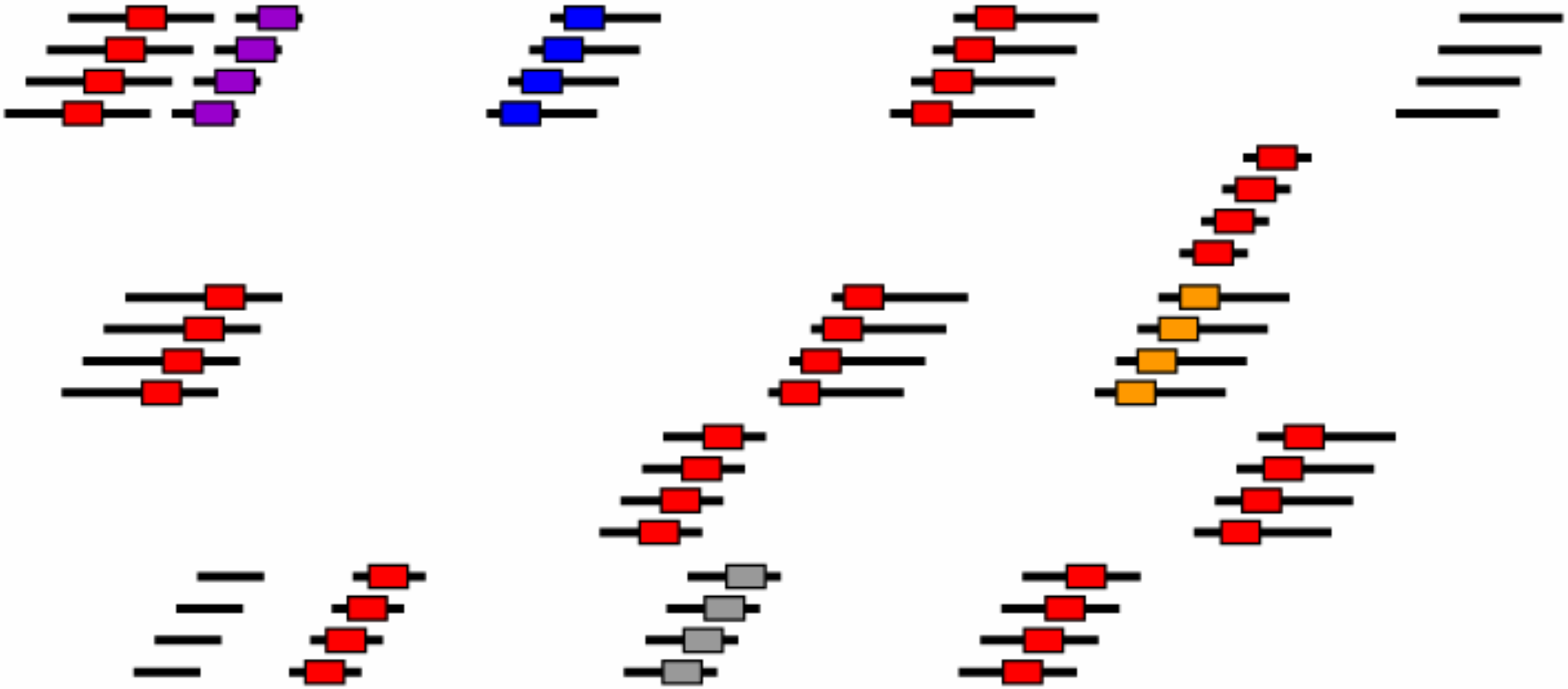
Immunoprecipitation



Crosslink reversal and DNA purification



Amplification



Brief history

Yeast:

DNA-binding sites of individual TFs (Ste12, Gal4, Swi4, Swi6, Rap1): 2000, 2001

"Regulatory network" of 106 TFs: Lee et al. 2002

Condition dependence: Harbison et al. 2004

Mammals:

Early experiments 2002, 2003

Limitations: size of genome / no. probes on array; repeated sequence

See Hanlon/Lieb review for more

Protein-DNA interactions

not just transcription factors...

DNA replication
Recombination
DNA repair

Technological options

Enrichment of the protein of interest:

Specific antibody

Tag protein and use tag-specific antibody

Tag protein and use tandem affinity purification

Array probes

Spotted PCR product

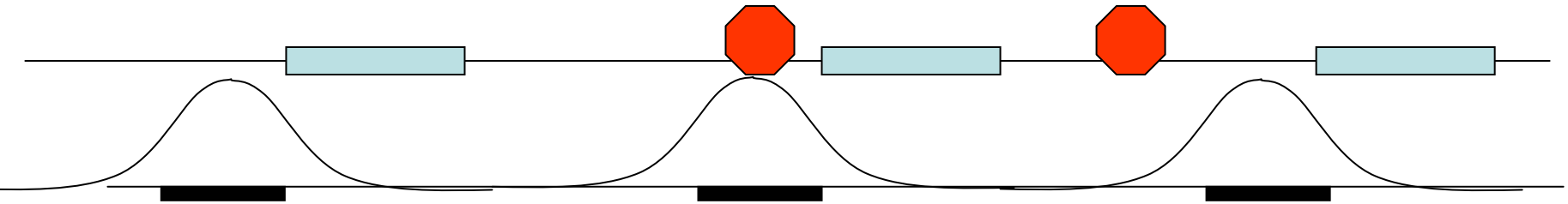
Spotted oligonucleotides

In-situ synthesized oligonucleotides (Affymetrix, Nimblegen)

Low-resolution arrays

Low spatial resolution (e.g. 1 data point per 5' upstream region)

Confounding between binding affinity / occupancy and distance from probe to binding site



Controls

Sample:
ChIP of interest

There are (at least) two types of control:

Control for array & hybridization variability (esp. in two-color technique), for probe effects (esp. in short oligo technique):
genomic DNA ("Input")

Biological control, for sample handling, differential PCR, antibody unspecificity
ideal: cells lacking AB epitope but otherwise identical
2nd best: mock IP (no AB)

How should the controls be used?

Hybridization controls: see later - probe response normalization

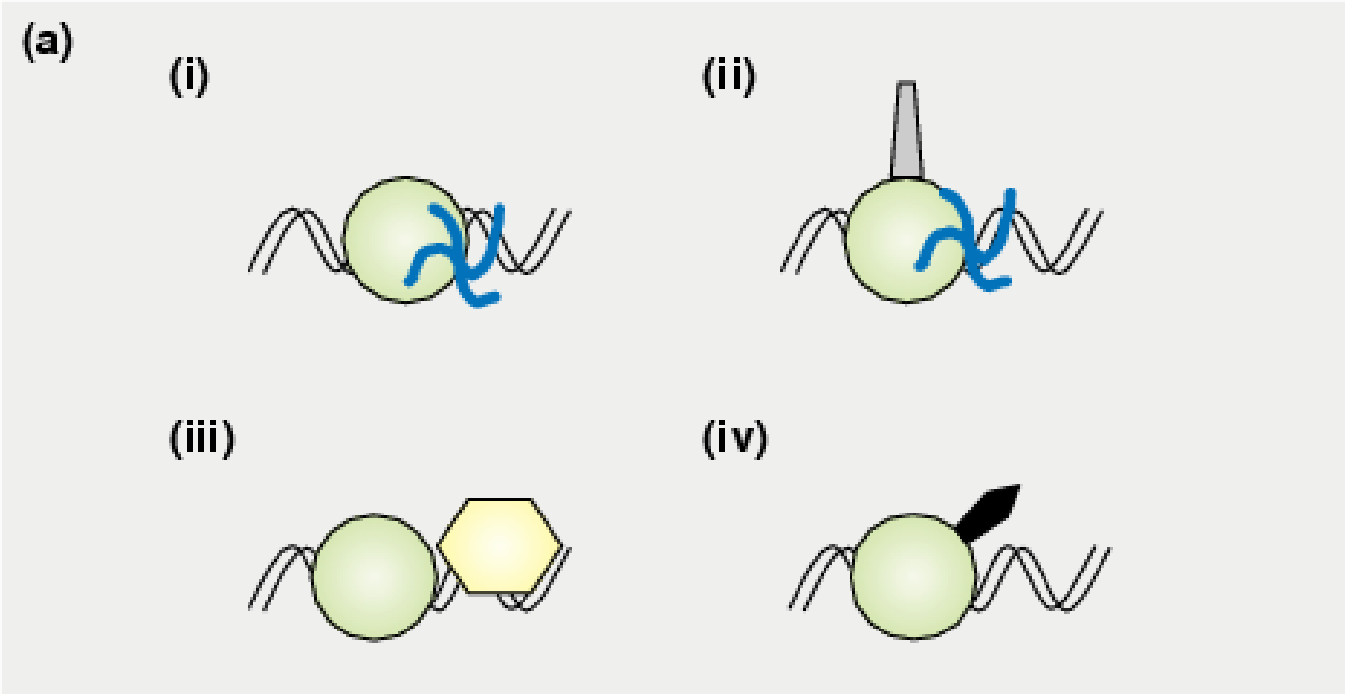
Biological controls:

1. Log-ratio: call enrichment if *log(IP) - log(control)* is large

2. Conditional: call enrichment if *log(IP)* is large and *log(control)* is small

Biologists seem to prefer 2.

Complications I: Variable formation of DNA-protein crosslinks



(i) Unmodified protein
(ii) Modified protein
(iii) Another interacting protein is in the way
(iv) modified lysine

From Hanlon and Lieb 2004

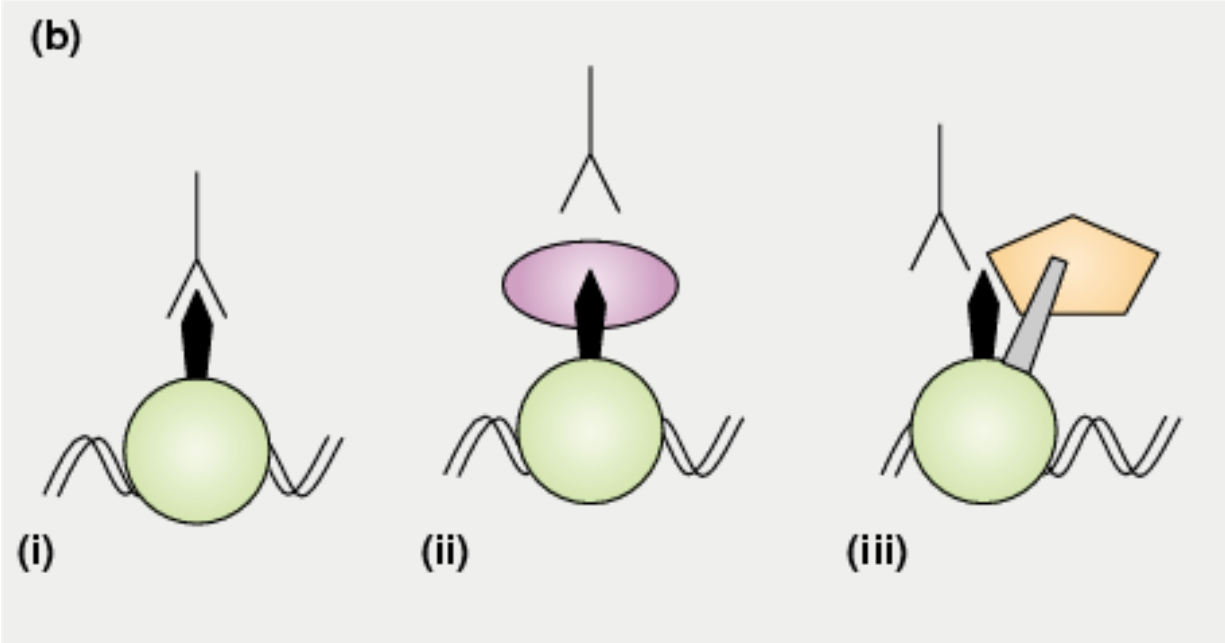
Formaldehyde:

crosslink proteins by linking amino-group of lysines with adjacent peptide bonds

also DNA-protein (if DNA is partially denatured)

what happens exactly when chromatin is crosslinked in vivo with formaldehyde is not well known.

Complications II: Variable epitope accessibility



(i) epitope detected by ChIP
(ii) direct competition
(iii) blockage

From Hanlon and Lieb 2004

A physical model for ChIP-chip data

Slides from

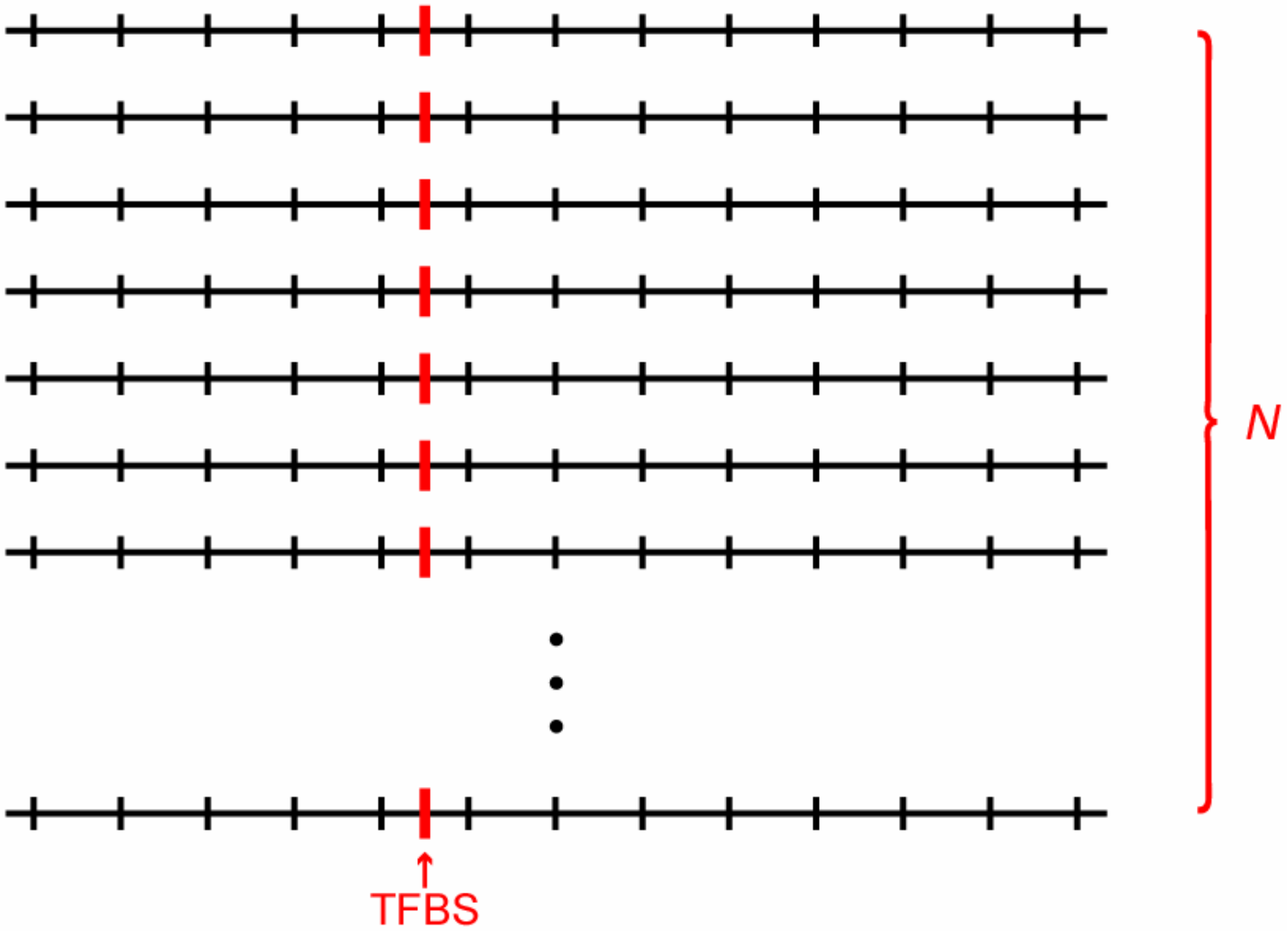
Richard Bourgon + Terry Speed

Department of Statistics, UCB

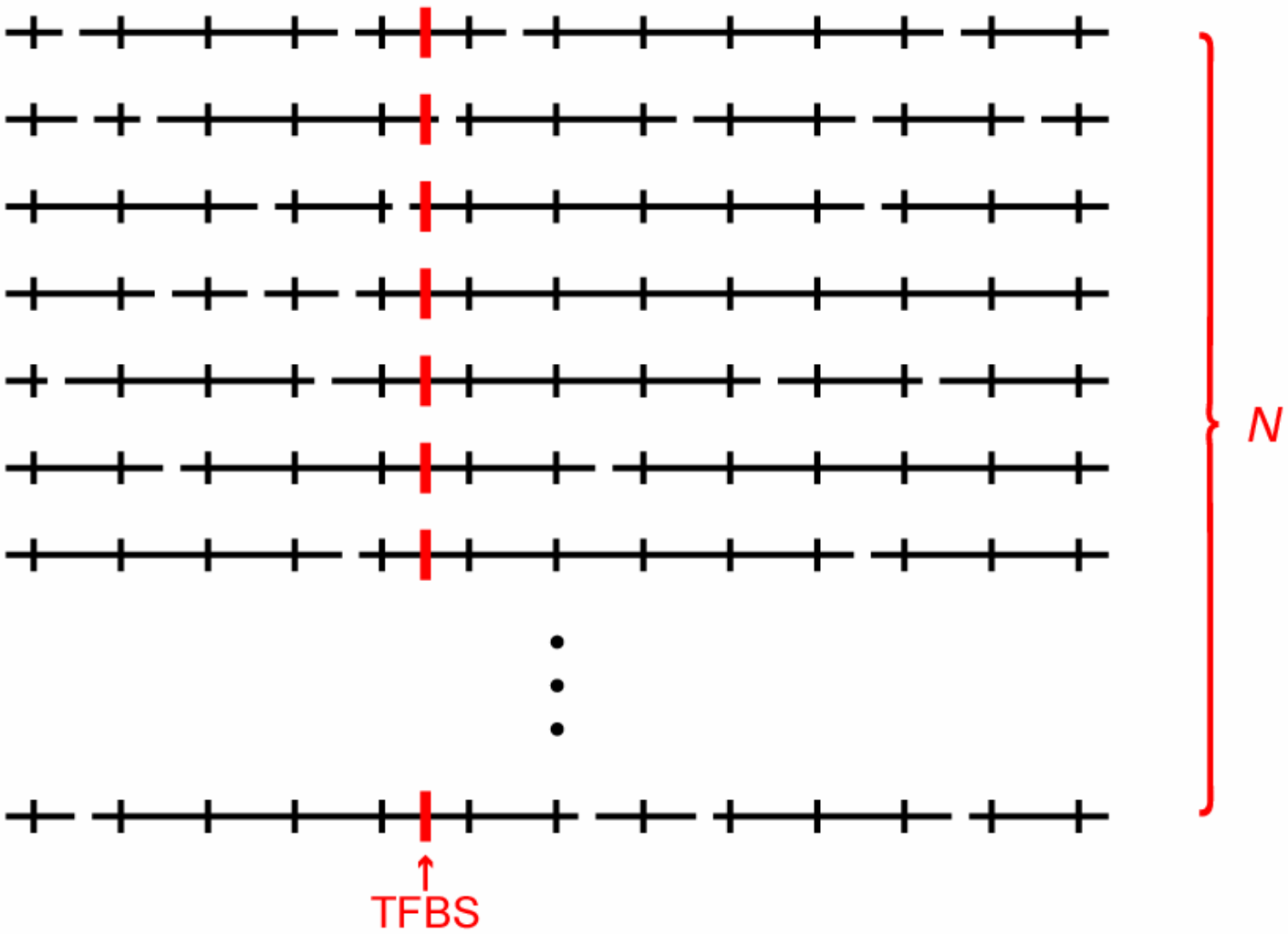
A physical/statistical model for the assay

Step	*Model*
Source material	N strands of extracted DNA.
Sonication	Uniform fragmentation of chromatin, with no interference. Probability of a break at any base is θ.
IP	Fragments with no binding site pass with probability ϕ; fragments with a binding site pass with probability ϕ', and $\phi' \gg \phi$.
Amplification	Z, a random multiplier for each fragment passing IP. (For PCR, Z is a branching process with t cycles and efficiency p.)

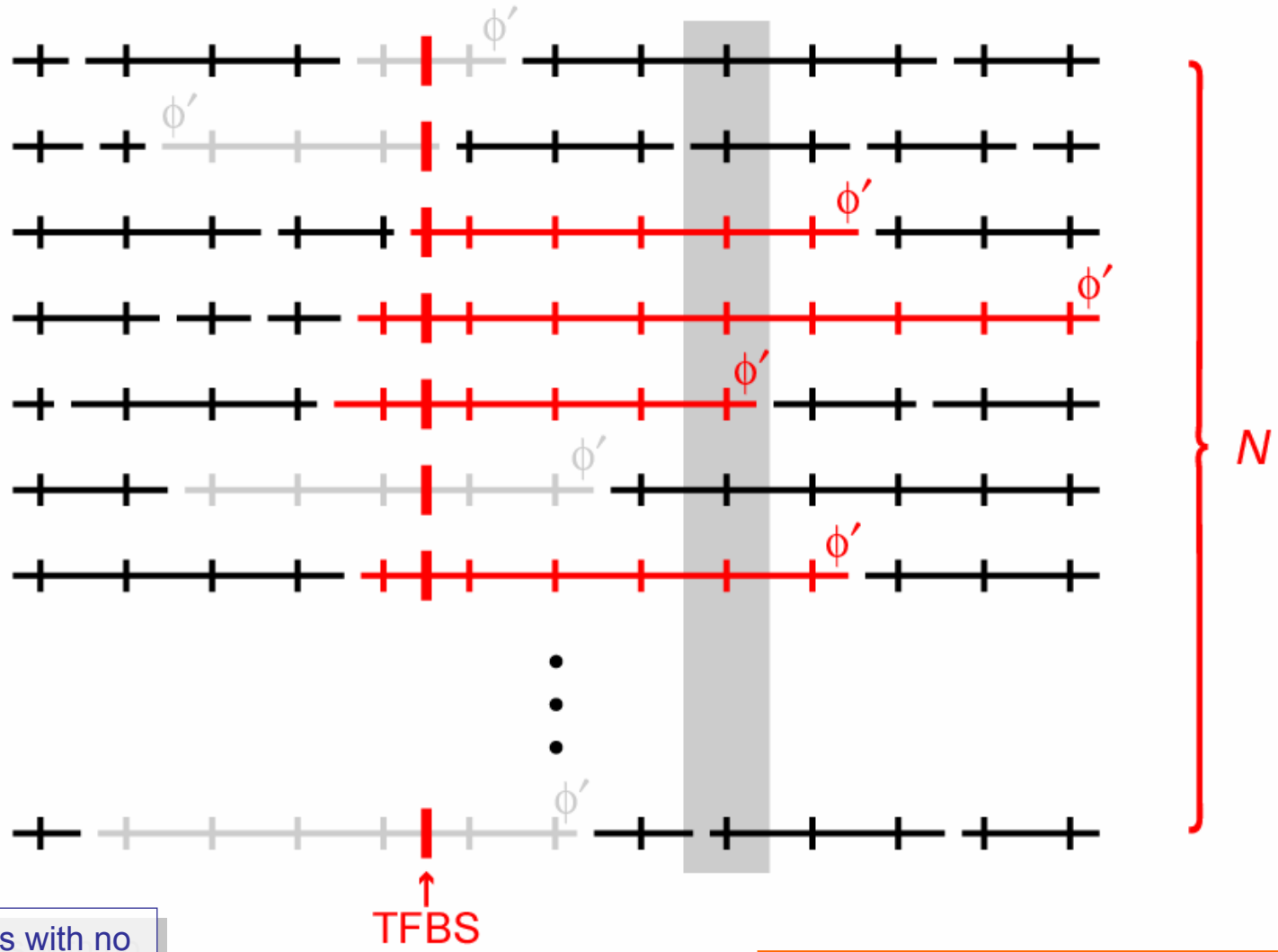
N source DNA strands



Sonication

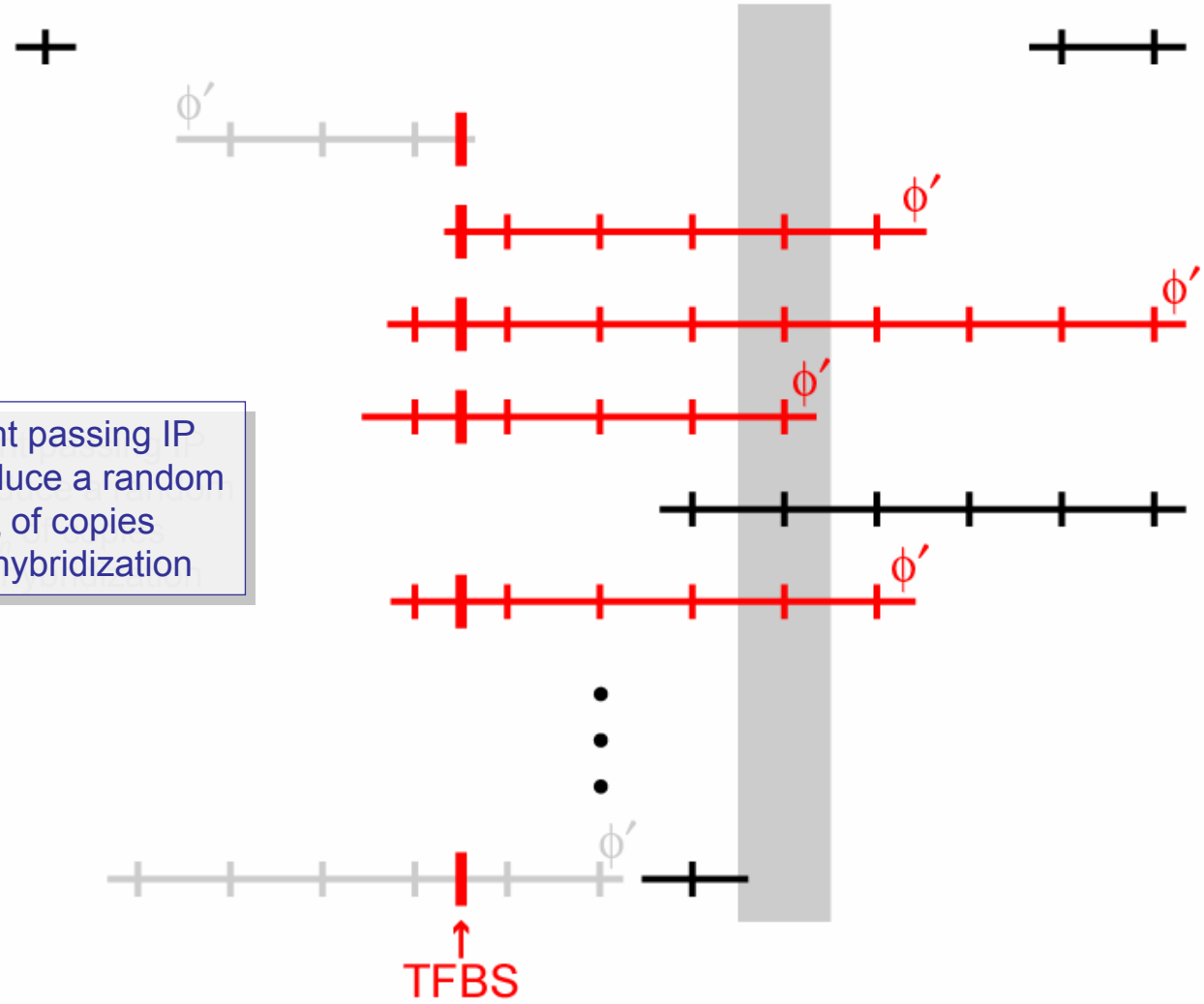


Immunoprecipitation



All fragments with no TF binding site pass with probability ϕ

Fragments passing IP



Each fragment passing IP goes on to produce a random number, Z_n of copies available for hybridization

Implications

This model has implications for…

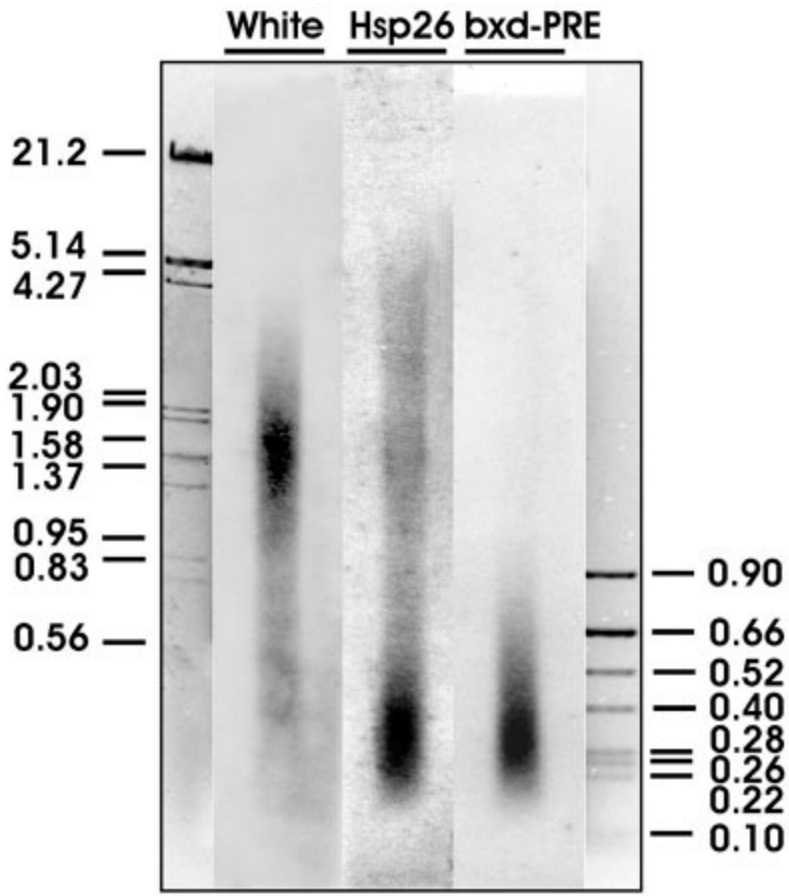
- **Target abundance: the shape and size of signal near a binding site.**
- **Spatial correlation: correlation between nearby observations, both near binding sites and also, significantly, in "background" regions far from binding sites.**

Sonication: uniform fragmentation?

Schwarz, Kahn, and Pirrotta (2005), *Mol. Cell Biol.* 25:432-39.

suggest that the *bxd* PRE and *hsp*26 promoter regions lie in lower density chromatin than the *white* coding region, and as a consequence are more sensitive to sonication.

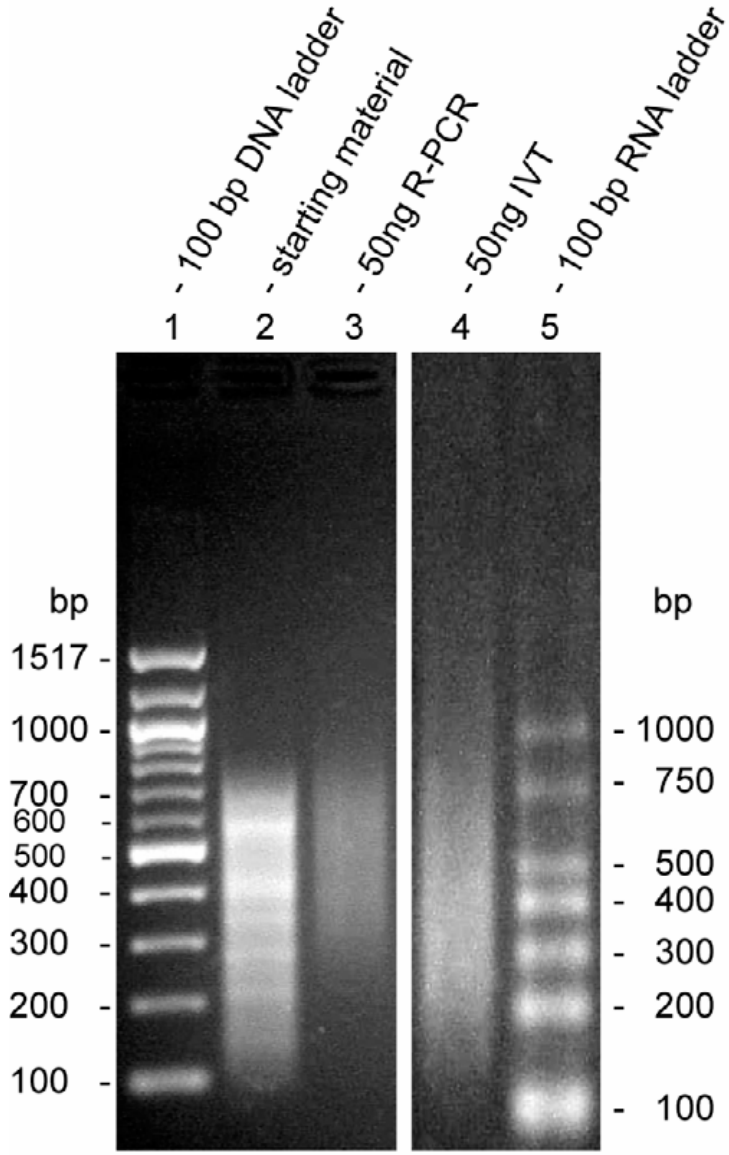
(Marker fragment sizes are in kb.)



PCR: *i.i.d.* amplification multipliers?

Liu, Schreiber, and Bernstein (2003), *BMC Genomics* 4:19-39.

Biases with respect to sequence and fragment length exist, and will be amplified exponentially.



Target abundance

Dependence of expected signal on distance from binding site to probe site

Target abundance

Suppose probe i is Δ bases from a binding site τ. Probability that fragment is available for binding to probe:

$$\mathbb{P}(i \sim \tau)\mathbb{P}(X_{in} = 1 \mid i \sim \tau) + \mathbb{P}(i \not\sim \tau)\mathbb{P}(X_{in} = 1 \mid i \not\sim \tau)$$

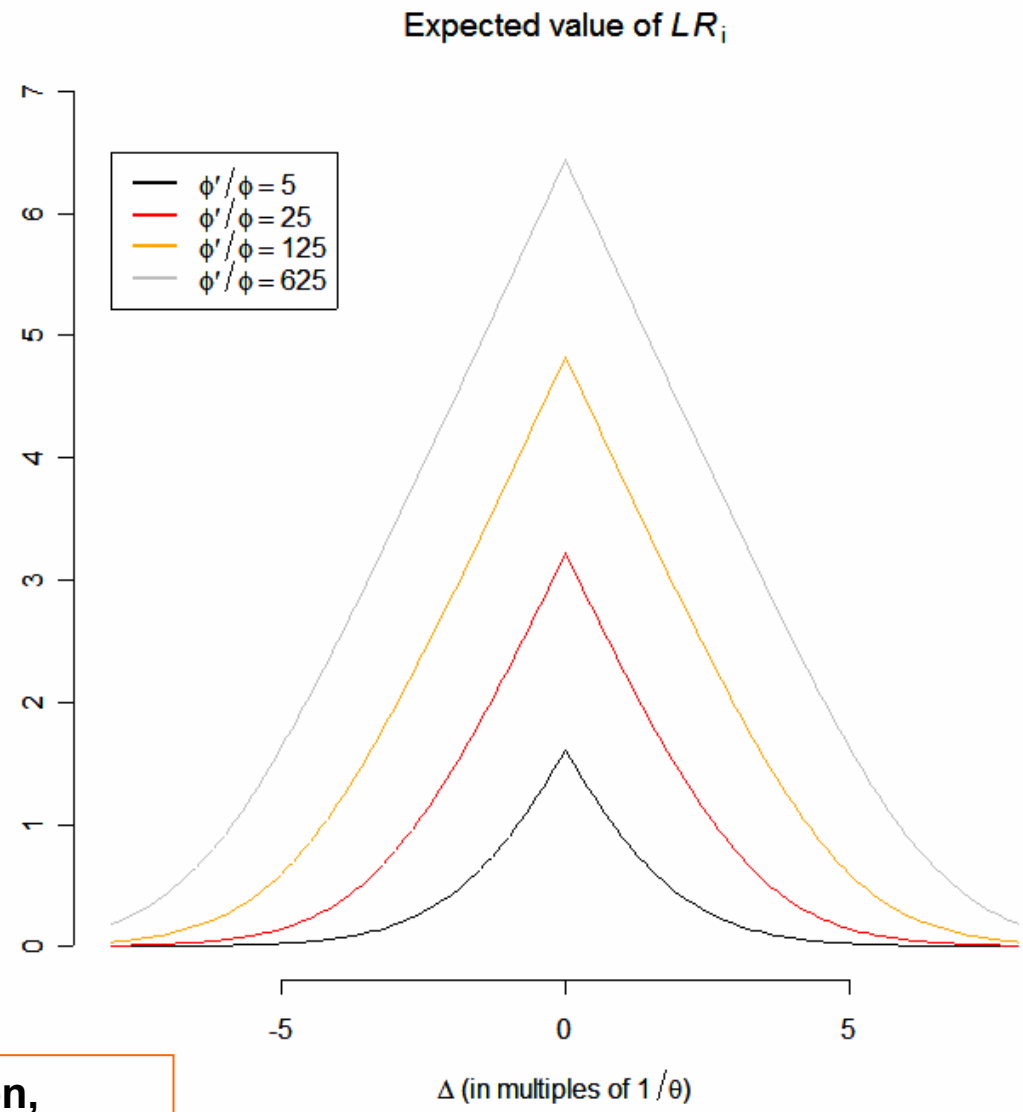
$$= (1-\theta)^{\Delta}\phi' + \left(1-(1-\theta)^{\Delta}\right)\phi$$

Note exponential decay from ϕ' to ϕ.

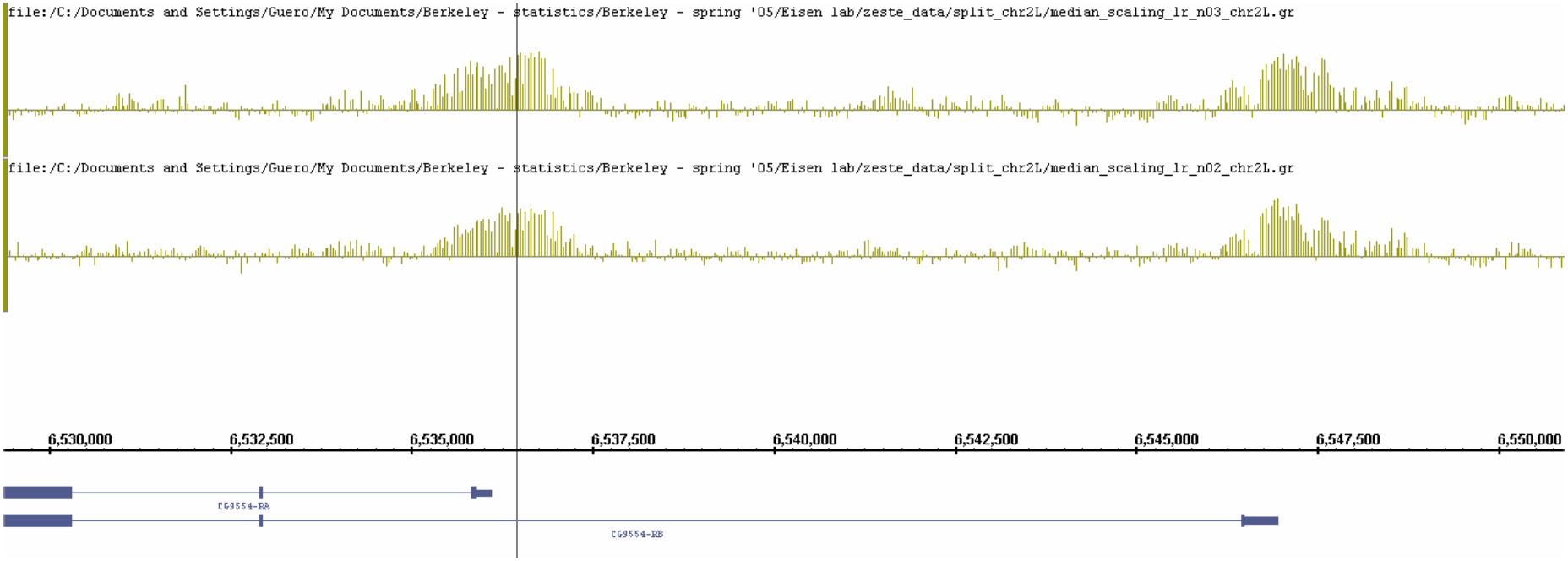
Expected log-ratio

Peak amplitude *and* width depend on the efficiency ratio:

$$\phi'/\phi$$



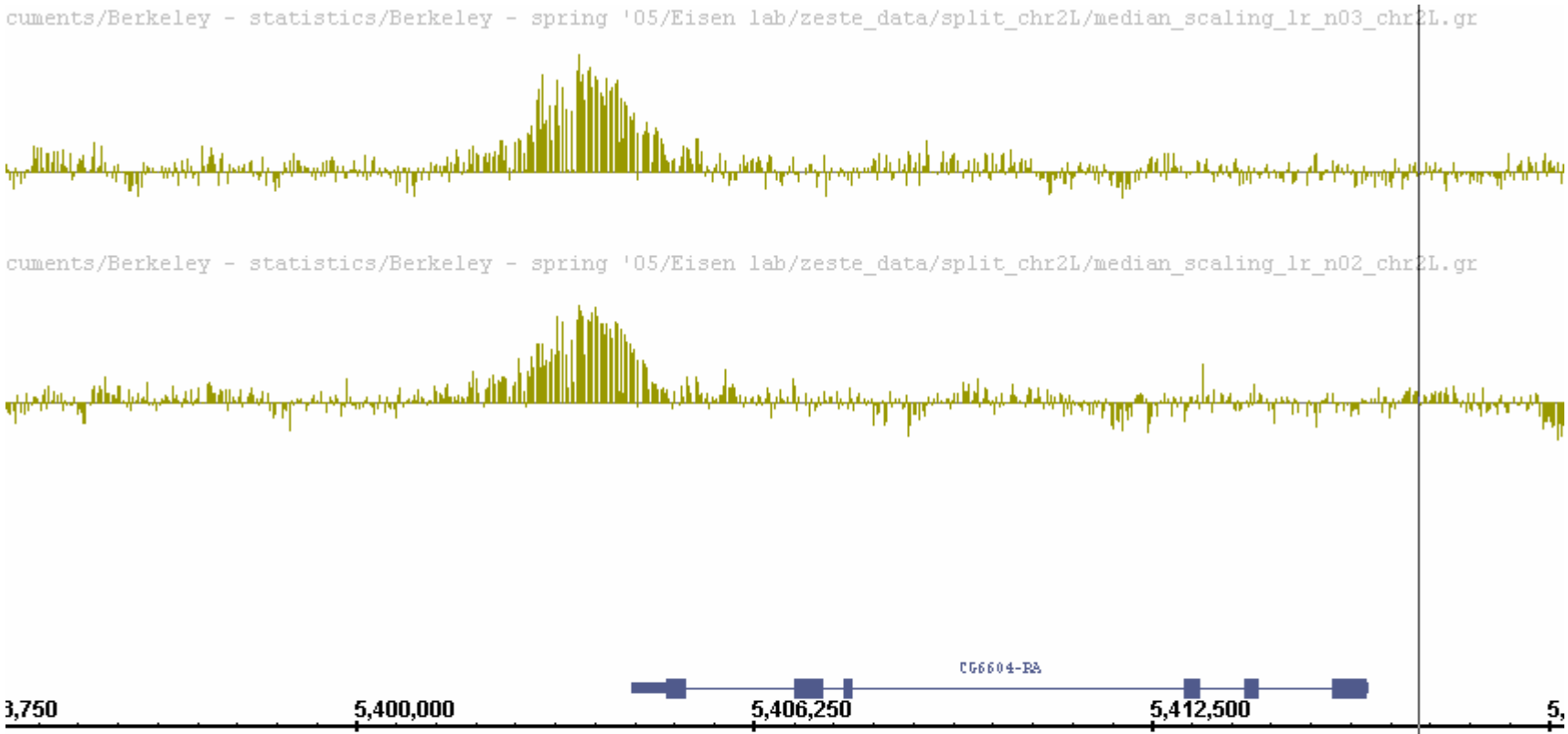
CG69554-RA and RB



Chromosome 2L

Log ratios (unsmoothed) from 3 vs. 3 comparisons, two different IP/PCR/hybridization groups.

CG6604



Summary: shape of expected signal

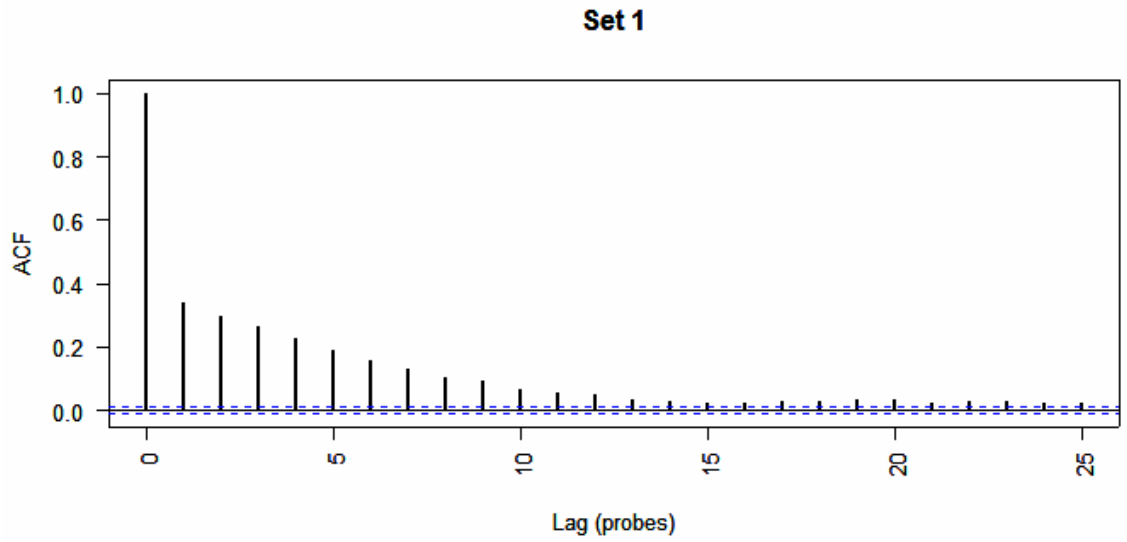
Binding sites produce high expected signal in multiple, consecutive probes.

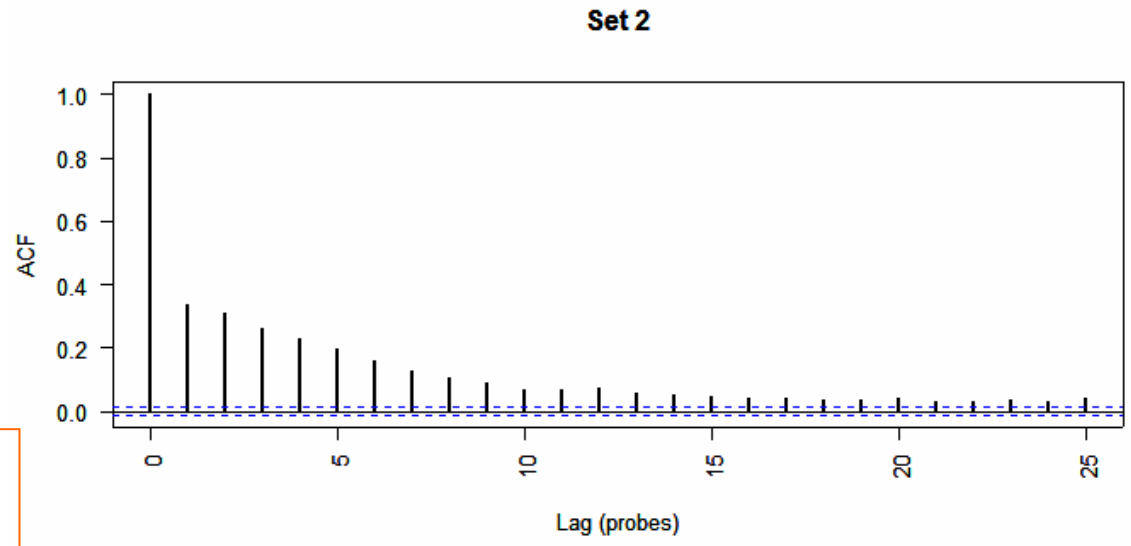
- Shape permits localization of the binding site! In the absence of noise, the binding site coincides with the peak.
- Methods which take shape into account may be more powerful than those that do not.
- Windowed enrichment estimates will be downwardly biased.
- Calibration/test data with a different signal shape will not accurately represent behavior of true signal.

Spatial correlation in log-ratio

For simplicity, ignore irregularity of probe spacing.

Compute auto-correlation at various lags.





For both data sets, there is statistically significant auto-correlation up to a lag of ≈15 positions.

Spatial correlation in log-ratio

in unbound regions

$$\mathrm{cor}(LR_i, LR_j) \propto (1-\theta)^{|i-j|}$$

θ related to fragment length distribution

Summary: spatial correlation

Under the model, both target abundance and the intensity log-ratio exhibit positive spatial correlation.

Correlation arises from the relationship between fragment size and probe spacing.

Here we have used the log-ratio statistic to confirm spatial correlation. Correlation in abundance, however, will impact *any* statistical procedure.

Ignoring spatial correlation can produce false positives, producing spurious hits *in both directions.*

Acknowledgments

– U.C. Berkeley
  - Richard Bourgon
  - Terry Speed, support from VIGRE (NSF).

– LBNL
  - Mike Eisen, Mark Biggin, David Nix, Xiaoyong Li.

– Affymetrix
  - Simon Cawley, Stefan Bekiranov, Antonio Piccolboni, Dione Bailey, Srinka Ghosh.