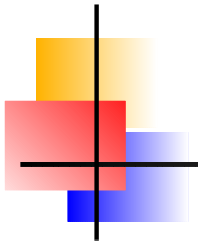


Advanced Statistical Methods for the Analysis of Gene Expression and Proteomics

Lecture 10 – Bayesian multiple comparison; ArrayCGH analysis

Yuan Ji

yuanji@mdanderson.org

Department of Bioinformatics and Computational Biology
University of Texas M.D. Anderson Cancer Center

Bayesian Multiple Testing Based on Test Statistics

Yuan Ji

January 28, 2008

Outline

1. The problem of multiple testing
2. The proposed approach
   1. A hierarchical modeling approach for multiple testing
   2. An illustrative example – F-tests
   3. A model assessment tool
3. Application
4. A brief discussion

A Bayesian framework

- Suppose a sequence of m null hypotheses H_{0i} is tested against a corresponding sequence of alternative H_{1i} for $i = 1, \ldots, m$.
- A Bayesian procedure for this problem:
  - Construct a latent indicator $J_i = 0$ if H_{0i} is true and $J_i = 1$ if H_{1i} is true.
  - Compute the marginal posterior probability $\Pr(J_i = 1|\text{data})$ based on some appropriate models.
  - Adjust for multiplicity using the marginal posterior probabilities.

Bayesian modeling

For test i, observed data y_i. A Bayesian hierarchical model consists of

- Probability distribution $p(y_i|J_i = k) = p_k(y_i; \boldsymbol{\theta_k})$, $k = 0, 1$.
- The likelihood function:

$$p_0(y_i; \boldsymbol{\theta_0})^{1-J_i} p_1(y_i; \boldsymbol{\theta_1})^{J_i}.$$

- Priors for $\boldsymbol{\theta_k}$ is $f_k(\boldsymbol{\theta_k})$; prior $\Pr(J_i = 1) = \pi$.
- Hyperpriors for the parameters in the priors (e.g., π).

Compute

$$r_i = \Pr(J_i = 1|y_1, \ldots, y_m)$$

the marginal posterior probability that H_{1i} is true.

Multiplicity

- Probabilities r_i adjust for multiplicities automatically as long as
  - $\Pr(J_i = 1) > 0$ for all $i = 1, \ldots, m$;
  - $\pi \sim p(\pi)$, rather than fixed.
  - *Ref. Scott and Berger (2003); Müller et al. (2006)*
- Optimal decision (Müller et al., 2004) is

$$I(r_i > t),$$

  to reject all the null hypotheses with $r_i > t$ for some fixed value t.
- Choice of t depends on choice of loss functions.

Motivation

- Construction of appropriate Bayesian models can be difficult. (e.g., construction of priors for $\boldsymbol{\theta_k}$).
- Values of posterior probabilities r_i are often sensitive to the prior densities.
- MCMC computation can be intensive, especially for high-dimensional data (e.g., genomics/proteomics data).

Hierarchical model based on test statistics

- Johnson (2005) proposed computing posterior probabilities r_j based on test statistics.
- Main idea:
  - Base the models on the sampling distributions of test statistics.
  - The null distributions are often completely specified – no need for prior specification.
  - The alternative distributions of test statistics can often be described with a parsimonious parametrization.

Hierarchical model based on test statistics (cont)

Therefore,

- Models under the null $p_0(y_i)$ are free of parameters.
- Models under the alternative $p_1(y_i; \boldsymbol{\theta_1})$ depend on few parameters (often just one).
- $\Pr(J_i = 1 | y_1, \ldots, y_m; \boldsymbol{\theta_1})$ has a closed-form solution – easy to sample.

Probability model

Let f_i be the test statistic (e.g., χ^2-, $F-$, $t-$ or $z-$statistic) for null H_{0i} vs. H_{1i};

- Likelihood $p(f_i|J_i,\tau) = p_0(f_i)^{1-J_i} p_1(f_i|\tau)^{J_i}$;
- Prior of $J_i \sim \text{Bin}(1,\pi)$;
  - Hyperprior of $\pi \sim \text{Beta}(p_0,(1-p_0))$, where p_0 is fixed.
- Prior of $1/\tau \sim \text{Gamma}(1,2)$;

MCMC

MCMC algorithm for $\{\pi, \tau, J_1, \ldots, J_m\}$

- Full conditional

$$\Pr(J_i = 1 | f_1, \ldots, f_m, \tau, \pi) = \frac{p_1(f_i|\tau)\pi}{p_1(f_i|\tau)\pi + p_0(f_i)(1-\pi)}$$

- $\pi | J_1, \ldots, J_m \sim \text{Beta}(p_0 + \sum J_i, (1 - p_0) + m - \sum J_i)$.
- Sample τ, e.g., using random-walk Metropolis-Hastings.

A special case – $F-$tests

Suppose

$$\mathbf{y}_i|\beta_i, \sigma_i^2 \sim N_n(\mathbf{X}_i\beta_i, \sigma_i^2\mathbf{I}).$$

For testing the validity of linear constraint $H_{0i} : \mathbf{Q}'\beta_i = \xi$, the classical $F-$ statistic f_i is the ratio of average sums of squares.

- $p_0(f_i)$ is a central $F-$ distribution;
- Suppose alternative H_{1i} assumes that

$$\beta_i \sim N(\beta_i^*, \tau\sigma_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1})$$

  where β_i^* is a value satisfying H_{0i},
- then $p_1(f_i|\tau) \sim (1+\tau)p_0(f_i)$.

Decision rules

- Posterior probability $r_i = \Pr(J_i = 1 | f_1, \ldots, f_m)$ is computed using the MCMC sample.
- Reject H_{0i} if $r_i > t$ for some value of t (more discussion later)

qqplot

A quantile-quantile plot is proposed to check the model fitting.

- Suppose $\{\tau^1, \ldots, \tau^B\}$ is the MCMC sample.
- Randomly draw τ^s.
- Obtain the corresponding posterior sample $\{J_1^s, \ldots, J_m^s\}$ from the s^{th} iteration of the MCMC.
- Assign the test statistics f_i to the null group if $J_i^s = 0$, and to the alternative group if $J_i^s = 1$.

qqplot continued

- Plot the sample quantiles of f_i in the null group against the theoretical quantiles based on the distribution $p_0(f_i)$;
- Plot the sample quantiles of f_i in the alternative group against the theoretical quantiles based on the distribution $p_1(f_i|\tau^k)$;
- Compare the curves with the 45 degree line.

This procedure only works for quantities of which the sampling distributions are free of parameters – such as the F-statistics (its distribution only depends on two degrees of freedom).

Simulation 1

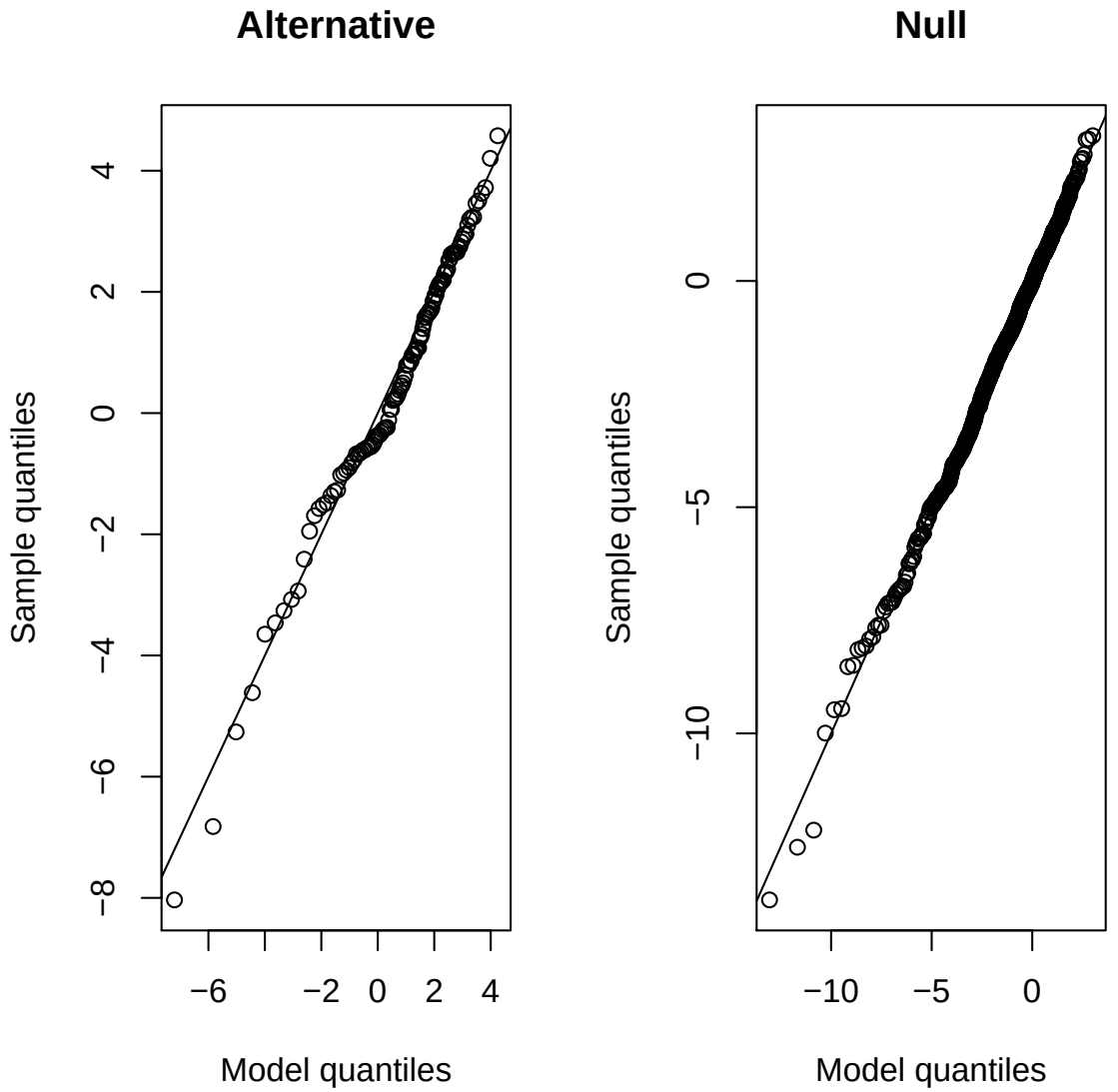
Consider one-sample t-tests $H_{0i} : \mu_i = 0$, $i = 1, \ldots, m$.

Observed data for test i are samples $\{y_{i1}, \ldots, y_{in}\}$. The $F-$statistic f_i is the square of the one-sample $t-$statistic.

- We generated $m = 1000$ tests.
- Sample sizes per test $n = 11$.
- Under H_{0i}, $f_i \sim F_{1,10}$ and under alternative $f_i \sim (1 + \tau)F_{1,10}$.

Simulation scheme consists of sampling τ, π, $J_i|\pi$, and $f_i|J_i, \pi$ (in this order), from their true distributions under the proposed model.

qq-plots 1


Alternative
Sample quantiles
Model quantiles
−8
−6
−4
−2
0
2
4
−6
−2
0
2
4
Null
Sample quantiles
Model quantiles
−10
−5
0
−10
−5
0

Simulation 2

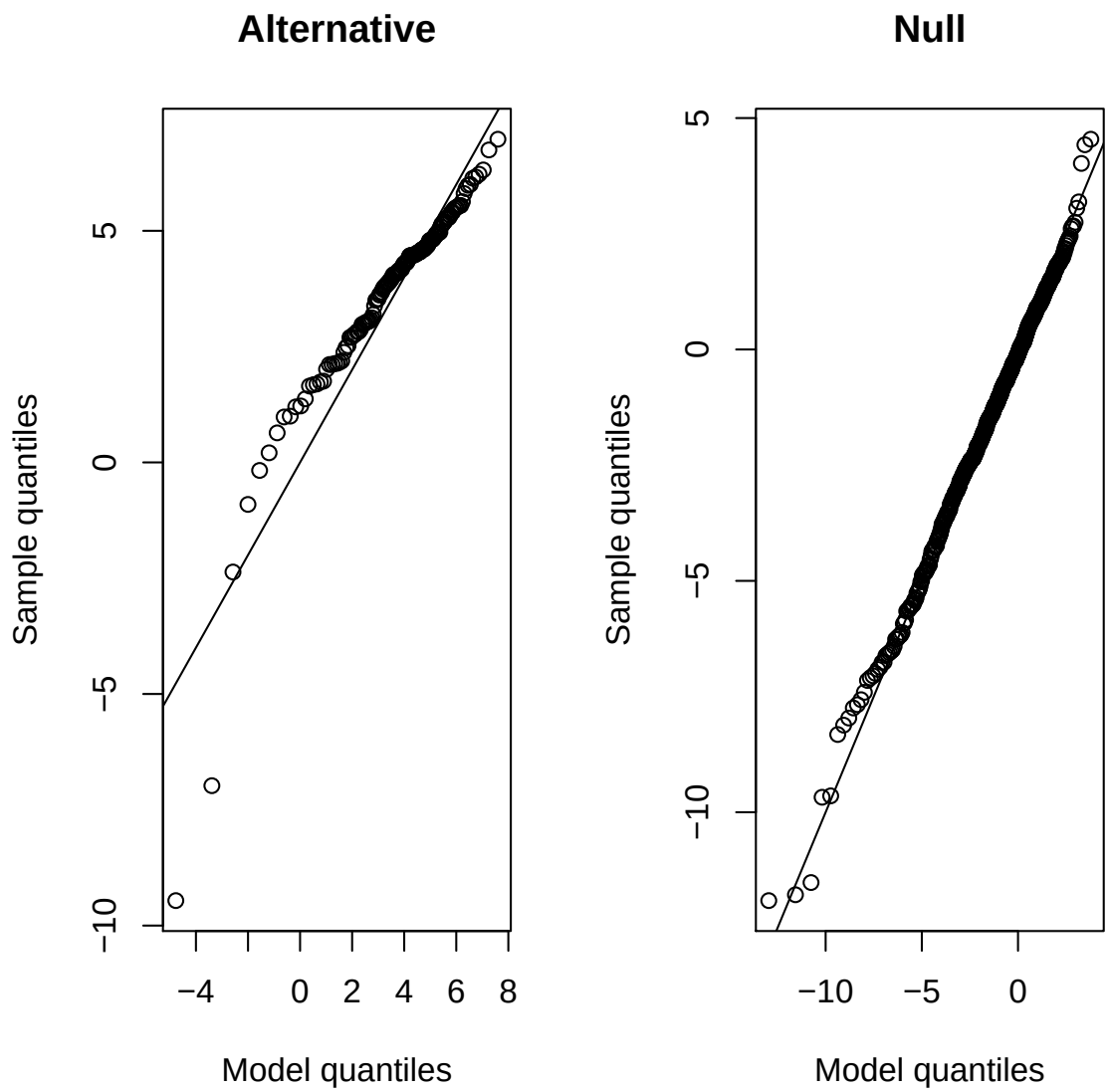
- Sample $y_{i1}, \ldots, y_{in} \overset{iid}{\sim} N(3,1)$ for $i = 1, \ldots, 100$;
- Sample $y_{i1}, \ldots, y_{in} \overset{iid}{\sim} N(0,1)$ for $i = 101, \ldots, 1000$;
- $H_{0i} : \mu_i = 0$
- Compute

$$t_i = \frac{\bar{y}_{i.}}{\hat{\sigma}/\sqrt{n}}$$

where $\bar{y}_{i.}$ is the sample mean and $\hat{\sigma}$ is the sample standard deviation.

After applying the proposed method,

qqplots 2

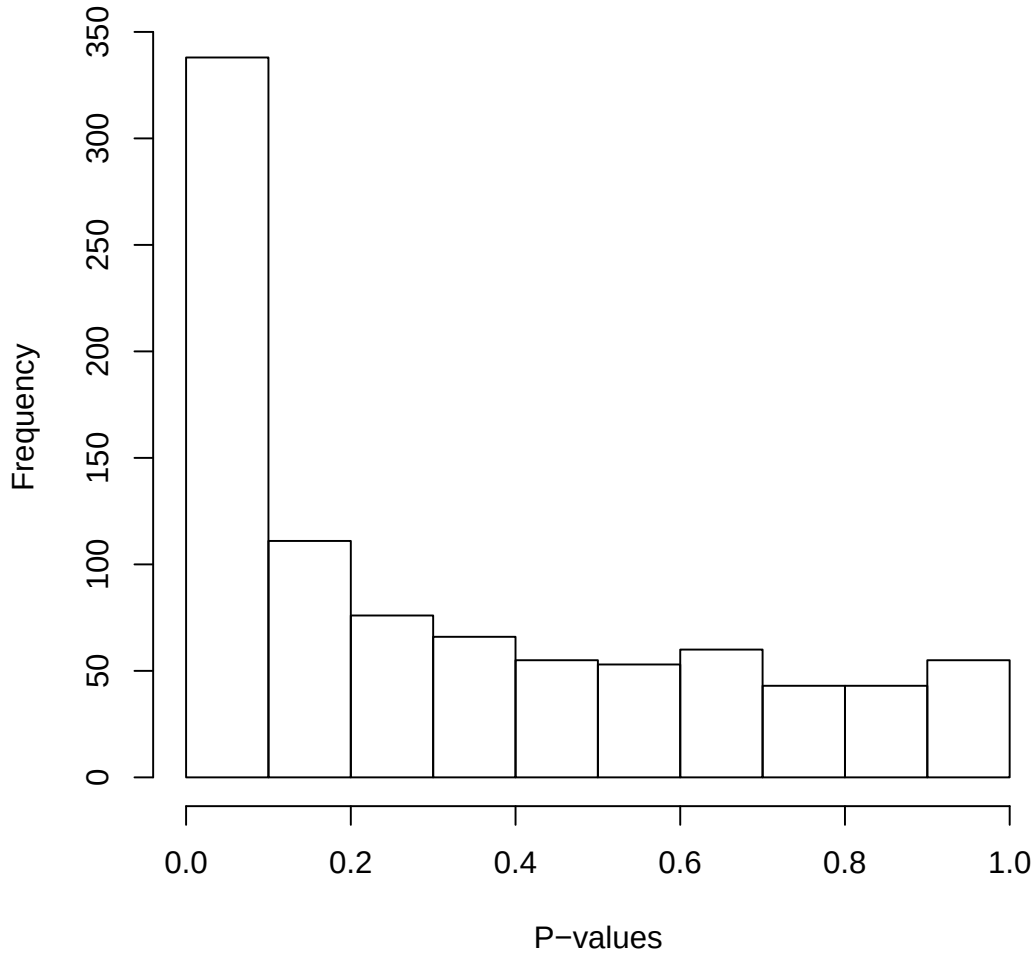

Alternative
Sample quantiles
Model quantiles
−10
−5
0
5
−4
0
2
4
6
8
Null
Sample quantiles
Model quantiles
−10
−5
0
5
−10
−5
0

siRNA screening

An siRNA screening experiment conducted by Gordon Mills and his lab.

- A kinase library of about 900 siRNA's are screened for their silencing properties.
- A functional silencing siRNA significantly reduced cell viability (measured as a continuous variable).
- Using 96-well plates, the library is screened with 30 plates in triplicates.
- F-statistics f_i are computed for all 900 siRNA's with degrees of freedom $(1, 4)$.

Histogram of p-values

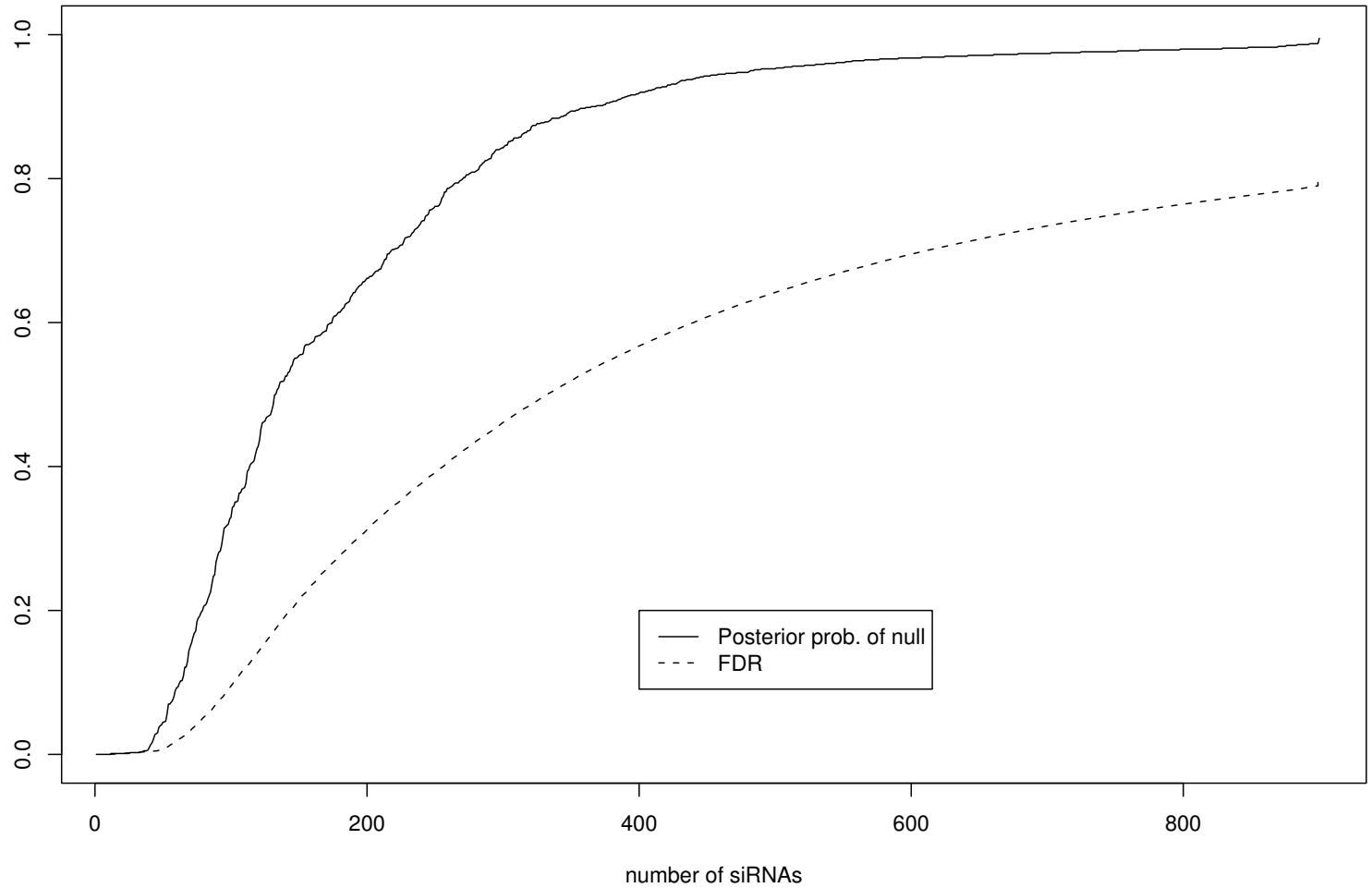


Results using the Bayesian procedure

We applied the proposed method for the 900 F-statistics f_i.

- Assume $\pi \sim \text{Beta}(0.5, 0.5)$.
- Assume $1/\tau \sim \text{Gamma}(1, 2)$.
- Under null, $f_i \sim F(1, 4)$.
- Under alternative, $f_i \sim (1 + \tau)F(1, 4)$.

Posterior probability and FDR



Decision rules

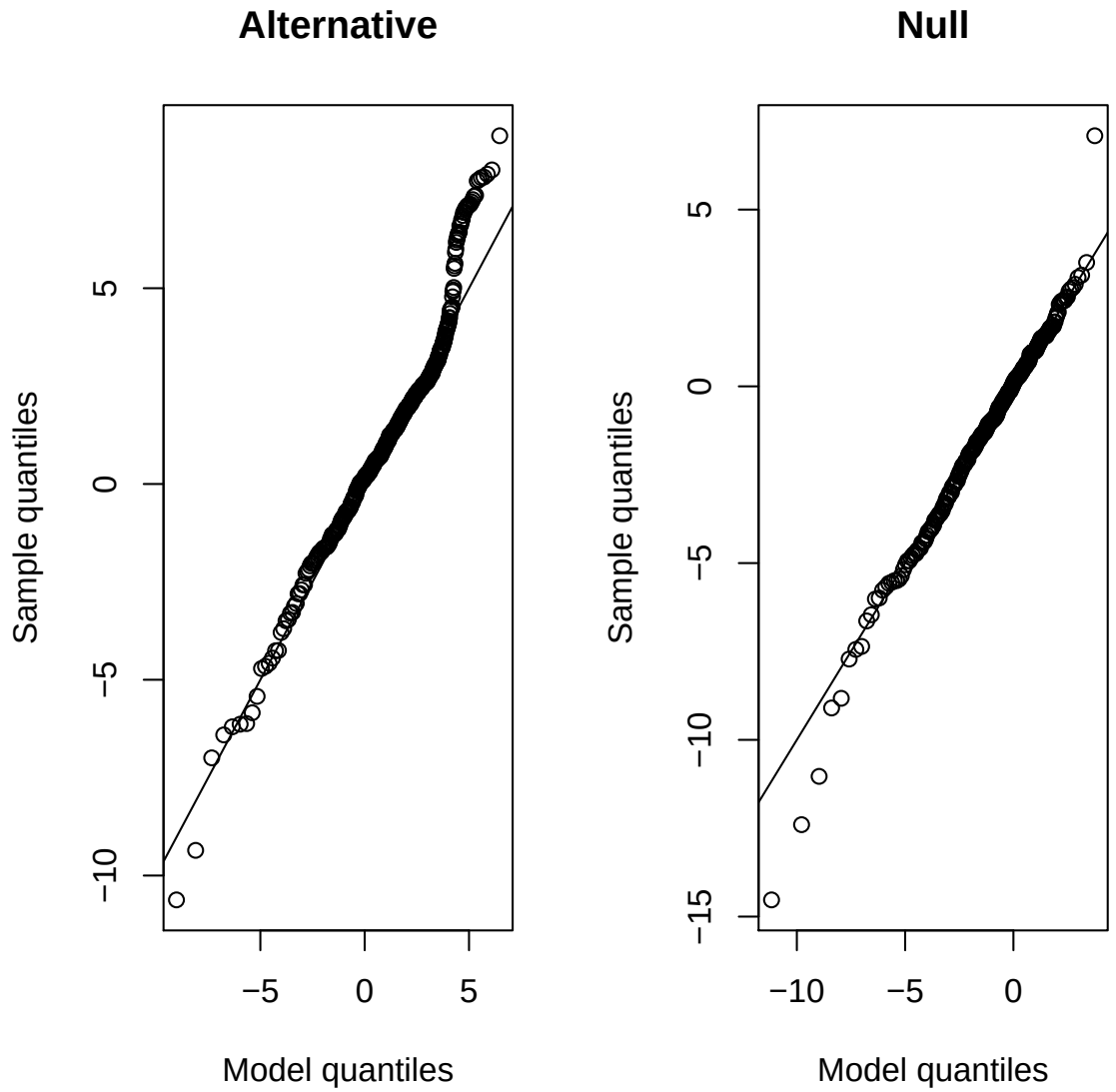
The optimal decision takes the form (M̈(u)ller et al., 2004)

$$d_i = I(r_i \leq t)$$

- If the "goal" (loss function) is to minimize FNR subject to FDR $\leq \alpha$, then t equals the largest r_i such that the corresponding posterior expected FDR (by rejecting all the $r_j \leq r_i$) is $\leq \alpha$.

- In the above plot, draw a horizontal line at y-axis $= 0.2$. Draw a vertical line at the intersection between the horizontal line and the dotted curve. The intersection between the vertical line and the solid curve is the optimal t value in d_i.

Model assessment



A gene expression experiment

Khodarev et al. (2005) studied the association between progression of Barrett's Metaplasia to Adenocarcinoma and gene expression levels. Three conditions are examined:

- Normal esophageal epithelium
- Premalignant Barrett's metaplasi,
- Esophageal adenocarcinoma

For each condition, $n = 8$ Affymetrix U133A arrays were produced from 8 different patients with the same condition. After normalization using dChip (Li and Wong, 2001), we obtained $m = 16384$ genes, each with 24 measurements.

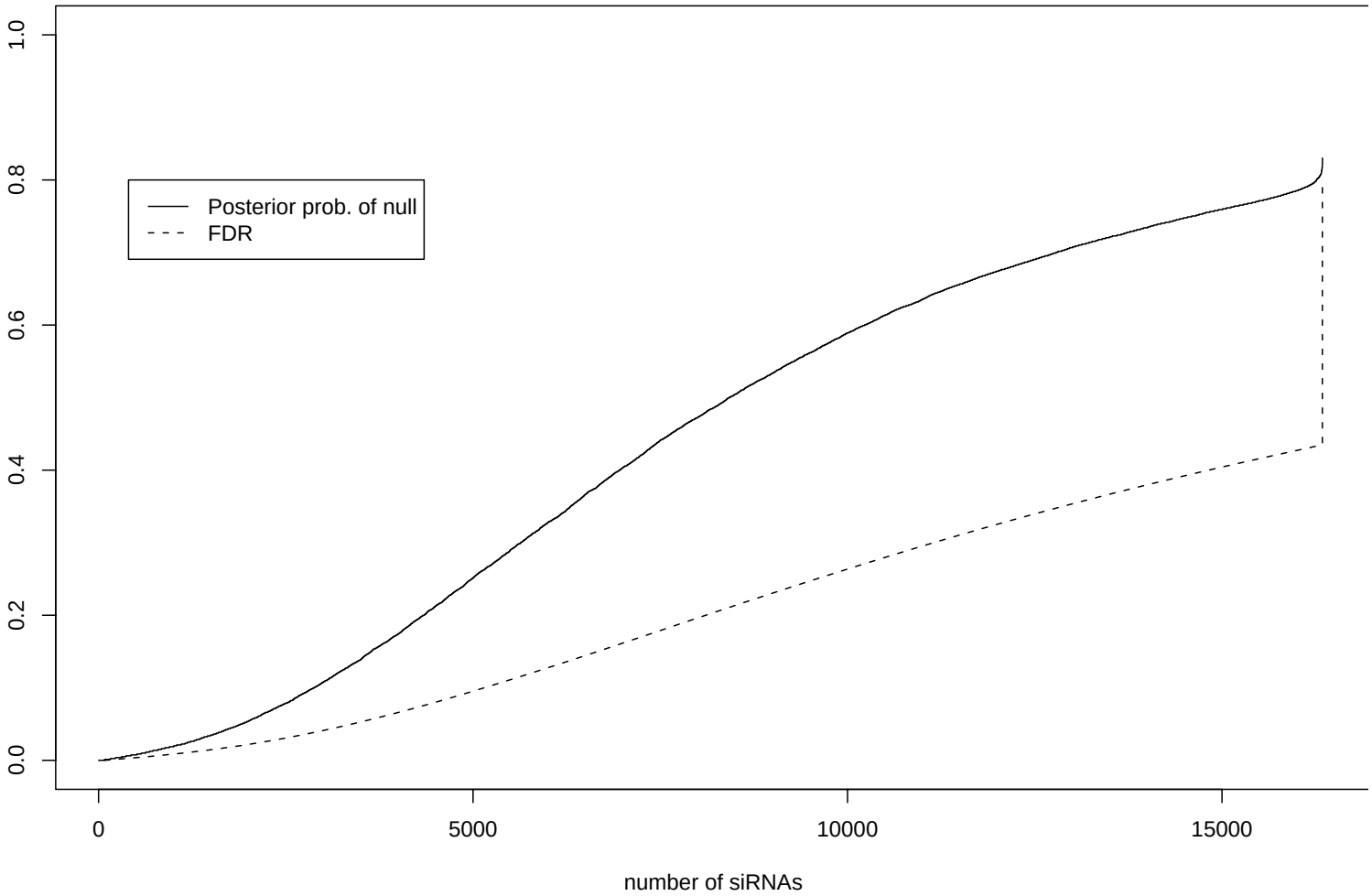
A gene expression experiment (cont)

For each gene, we performed a one-way ANOVA using the three conditions as a factor. We obtained $m = 16384$ $F-$statistics with degrees of freedom $(2, 21)$. Therefore,

- $p_0(f_i)$ follows $F_{2,21}$
- $p_1(f_i|\tau)$ follows $(1+\tau)F_{2,21}$.

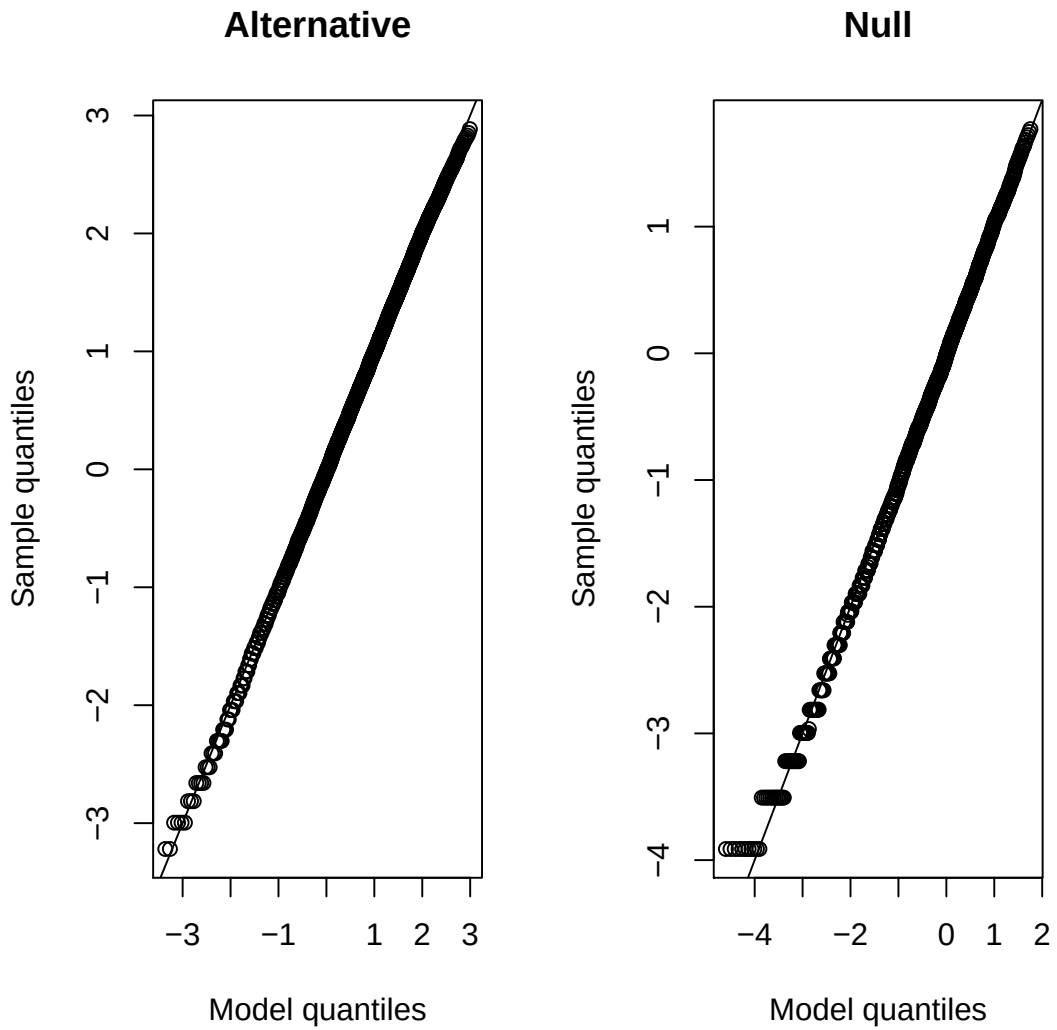
We applied the proposed method and computed $r_i = \Pr(J_i = 1|f_1, \ldots, f_m)$ for each gene i.

Posterior probability and FDR



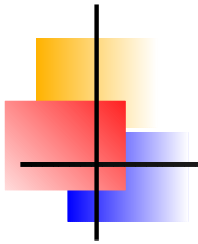
Model assessment

We let $\pi \sim \text{Beta}(.5, .5)$.



Conclusions

- The proposed model simplifies the process of specifying prior distributions for unknown parameters, which can be tricky.
- Only one parameter needs to be sampled using M-H; others are sampled directly from Bernoulli distributions.
- Information across all the tests is used in the decision making for each single test – through the common parameter τ.
- We provide a simple model-assessment tool to check the model fitting.
- Additional research is needed to explore more general assumptions under the alternative when model does not fit.



ArrayCGH Analysis

- Introduction of the technology
- Two existing methods for analysis
- Bayesian parametric/nonparametric modeling (ongoing)

http://creativecommons.org/licenses/by-sa/2.0/

creative commons

COMMONS DEED

Attribution-ShareAlike 2.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works
- to make commercial use of the work

Under the following conditions:

BY: **Attribution**. You must give the original author credit.

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

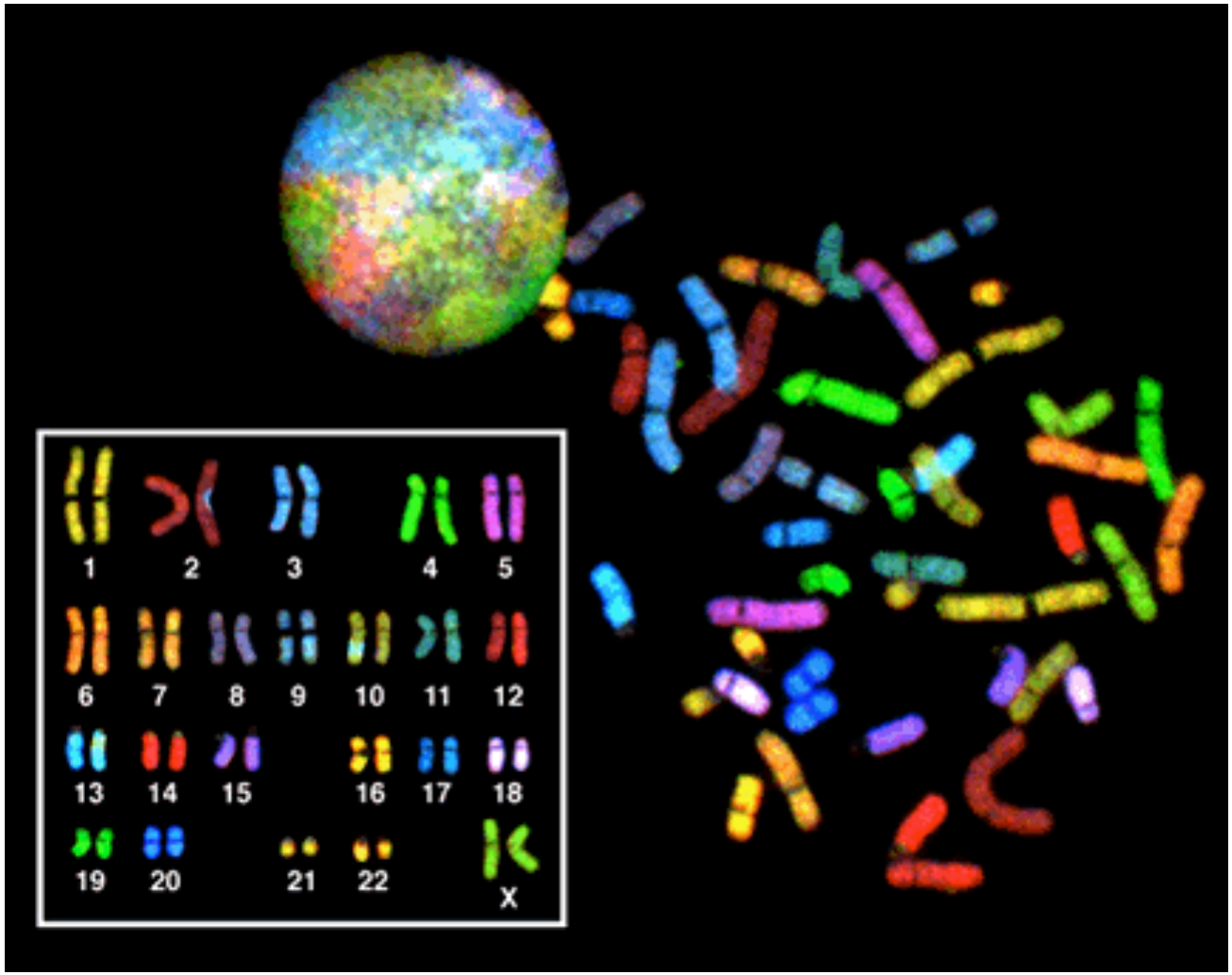
- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

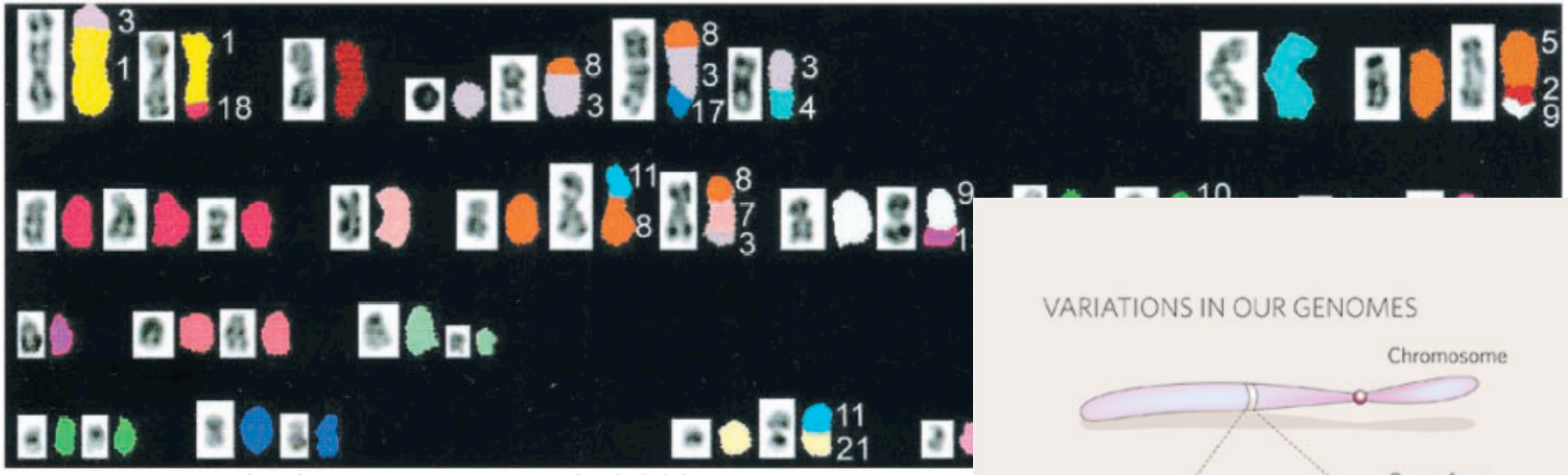
Disclaimer

A normal human genome has 2 copies of DNA



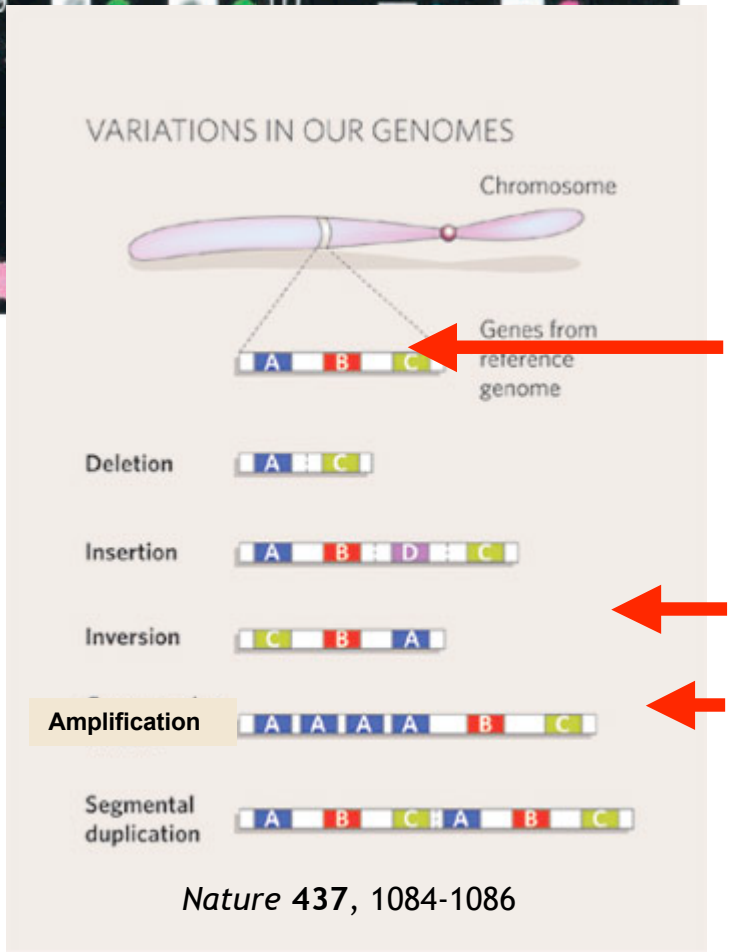
Wikipedia, the free encyclopedia

Copy number alterations (CNA) can lead to disease



Bayani et al, Cancer Research 2002

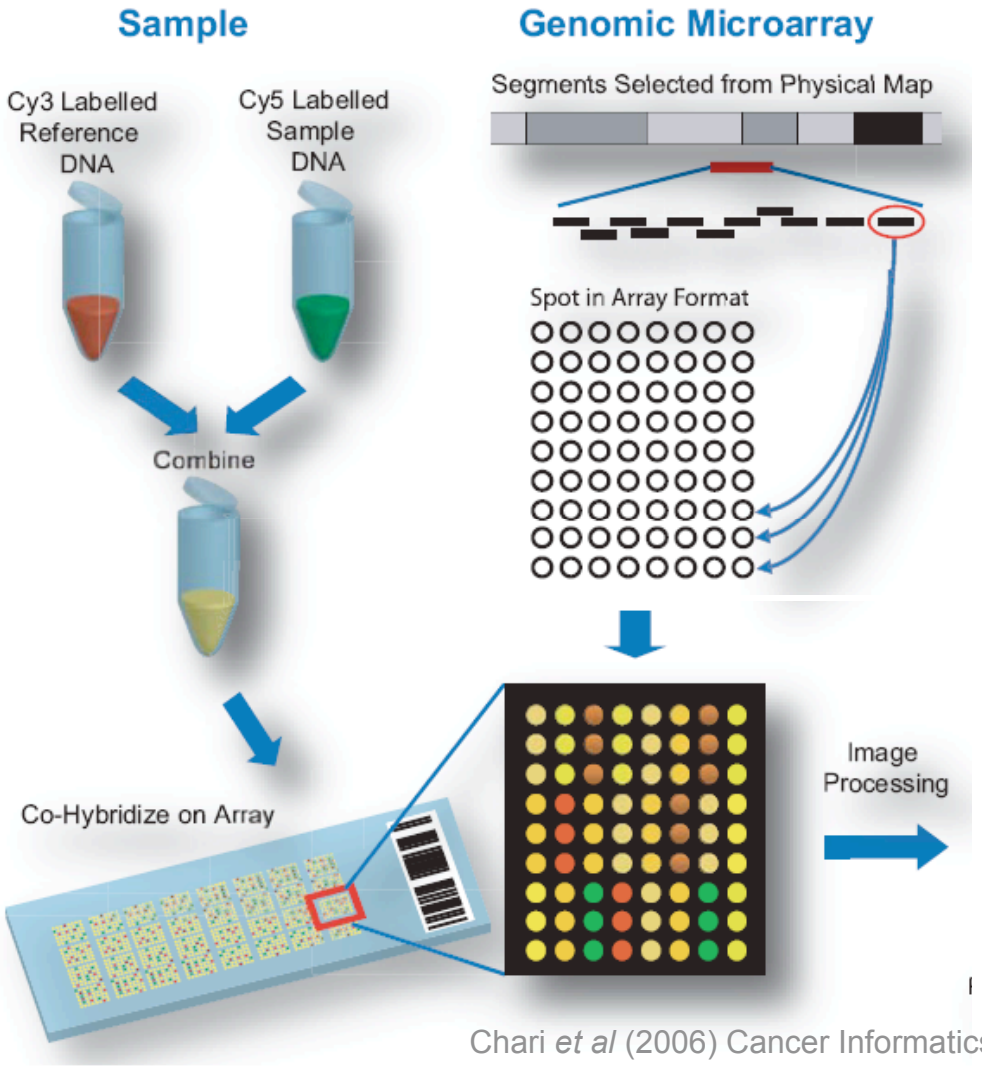
- CNAs can lead to adverse expression changes of affected genes
- CNAs are a hallmark of tumor genomes
- CNAs are diagnostic of mental retardation
- Recurrent CNAs in individuals with common phenotype represent molecular markers of disease
- Task: *find recurrent CNAs for diagnostics, gene-disease association, disease susceptibility*



Nature **437**, 1084-1086

- **Array hybridization - similar to cDNA array studies:**
  - **Test DNA sample - Unknown DNA copy number**
  - **Reference DNA sample - DNA copy number of 2**
  - **Label, mix, hybridize to BAC, cDNA, or oligonucleotide targets/probes spotted on a glass array**
  - **Scan**
- **Array analysis - resulting data are normalized log test over reference intensities for genomic targets**

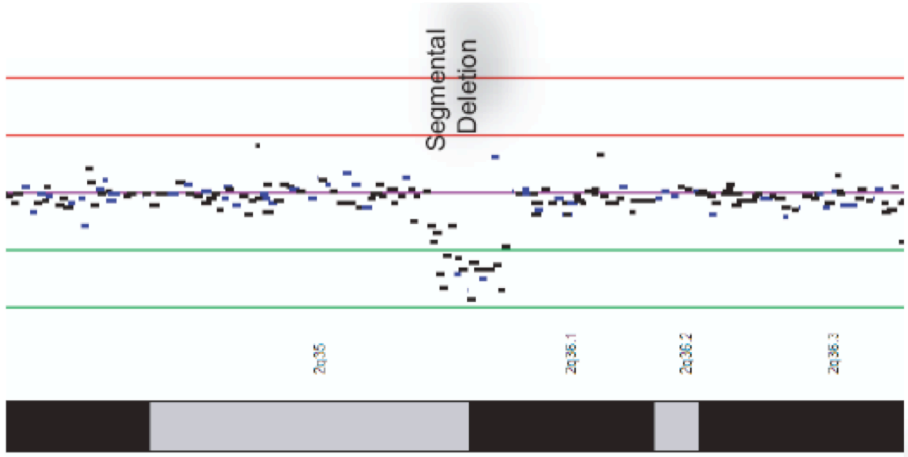
Measuring CNAs with array comparative genomic hybridization (aCGH)



Chari *et al* (2006) Cancer Informatics

Array comparative genomic hybridization (aCGH)

$\log_2 4/2 = 1.00$
$\log_2 3/2 = 0.58$
$\log_2 2/2 = 0.00$
$\log_2 1/2 = -1$



Gain
Neutral
Loss

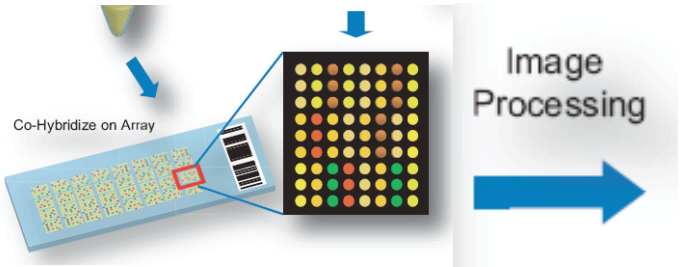
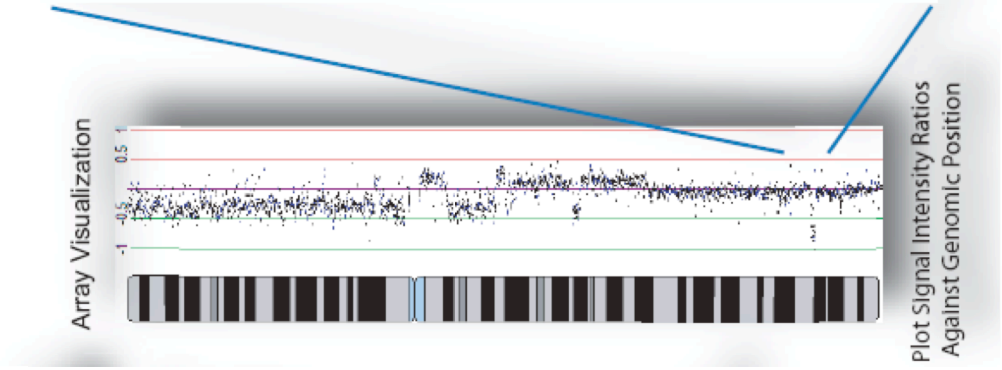


Image Processing



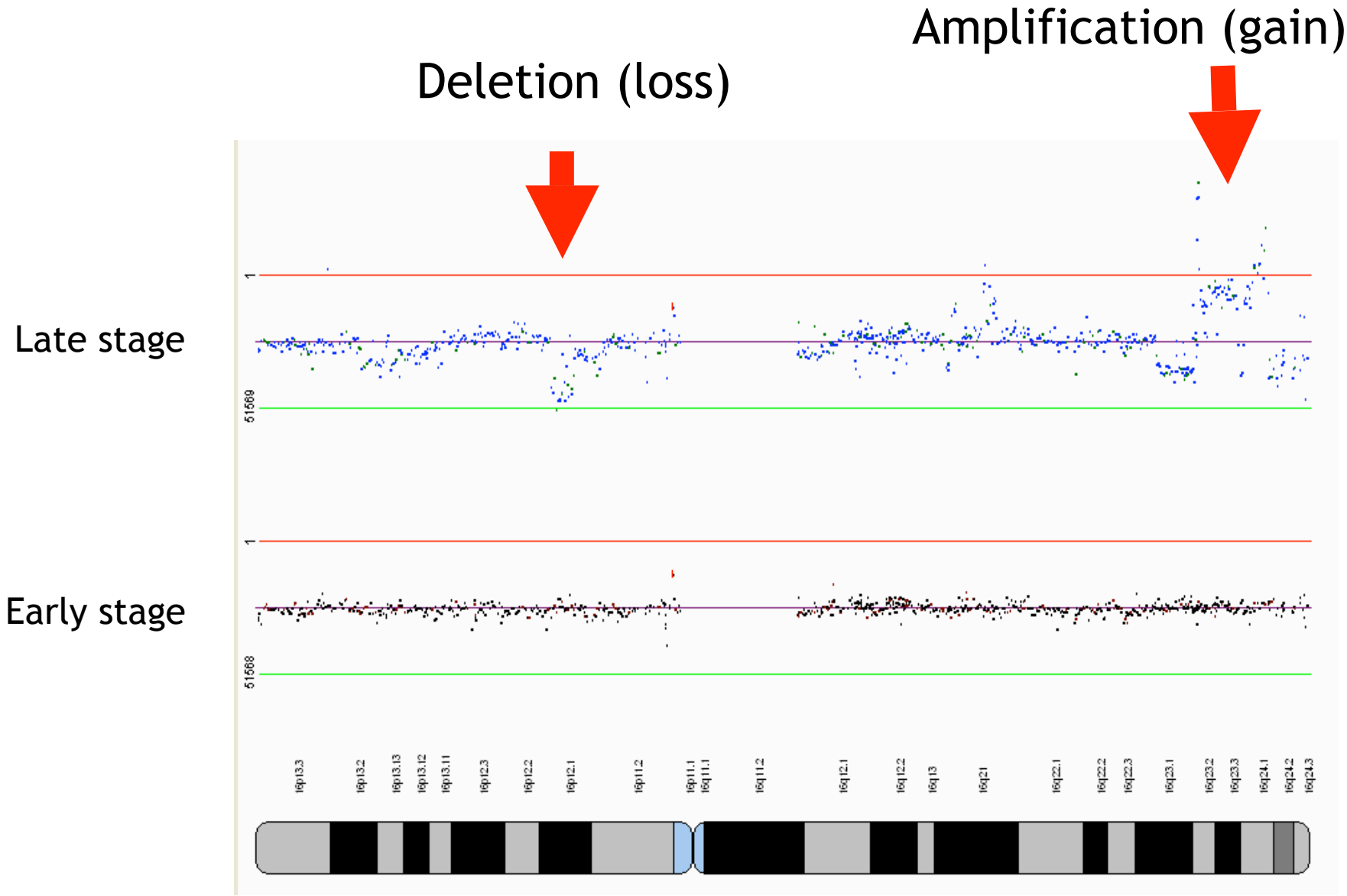
Indexed by chromosomal location

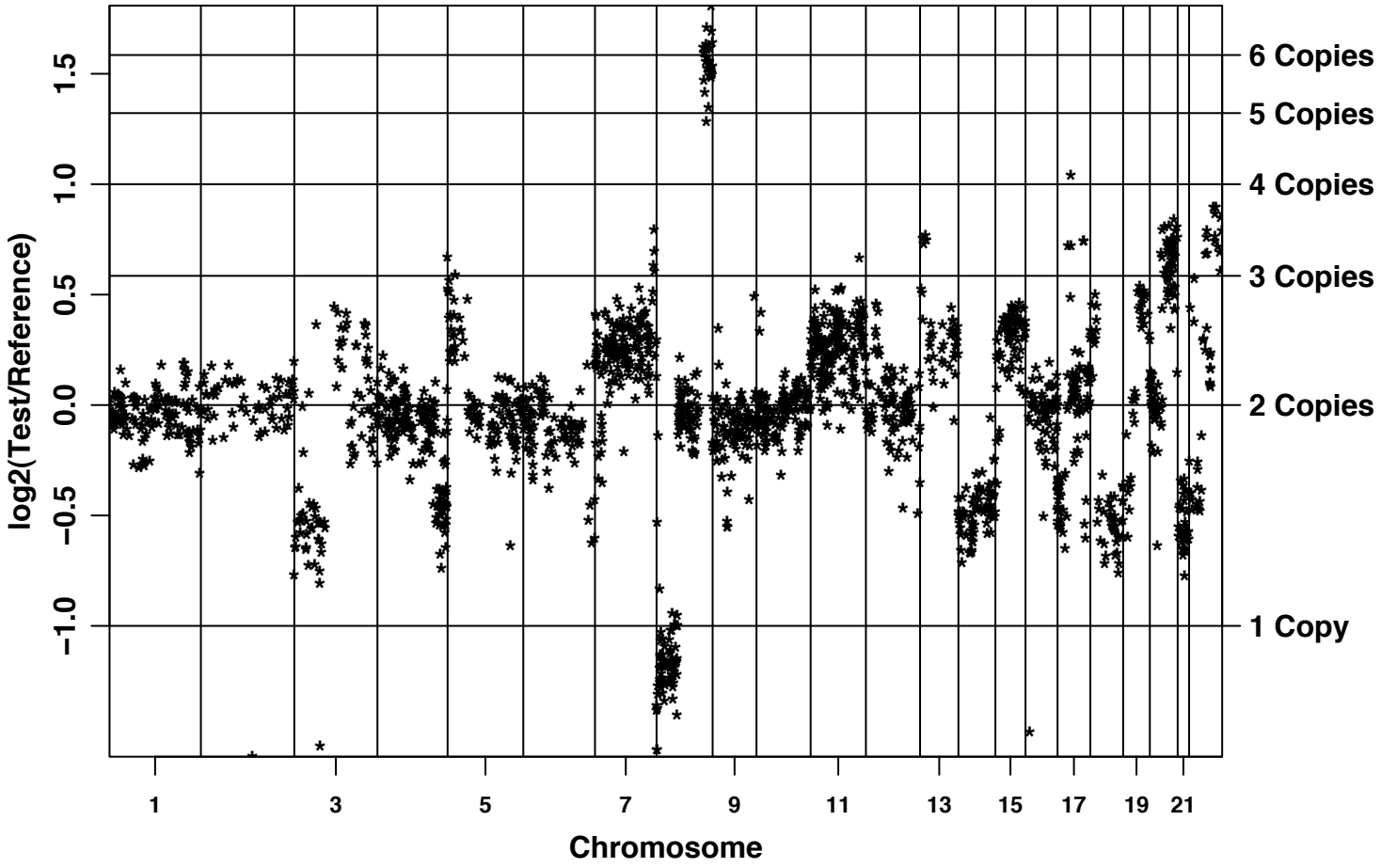
$$y_t = \log_2 \frac{sample_t}{ref}$$

Intensity of sample probe t

Intensity of reference

Examples of CNAs acquired in a lymphoma patient




log2(Test/Reference)
1.5
1.0
0.5
0.0
−0.5
−1.0
6 Copies
5 Copies
4 Copies
3 Copies
2 Copies
1 Copy
1
3
5
7
9
11
13
15
17
19
21
Chromosome

Things that might be done with this data are:

1. **Identification of genes and regions that often have abnormal copy number**
2. **Association of copy number with clinical data**
3. **Classification**
4. **Clustering of genes or samples**

Our methods are focused on 1, but may help for 2 and 3

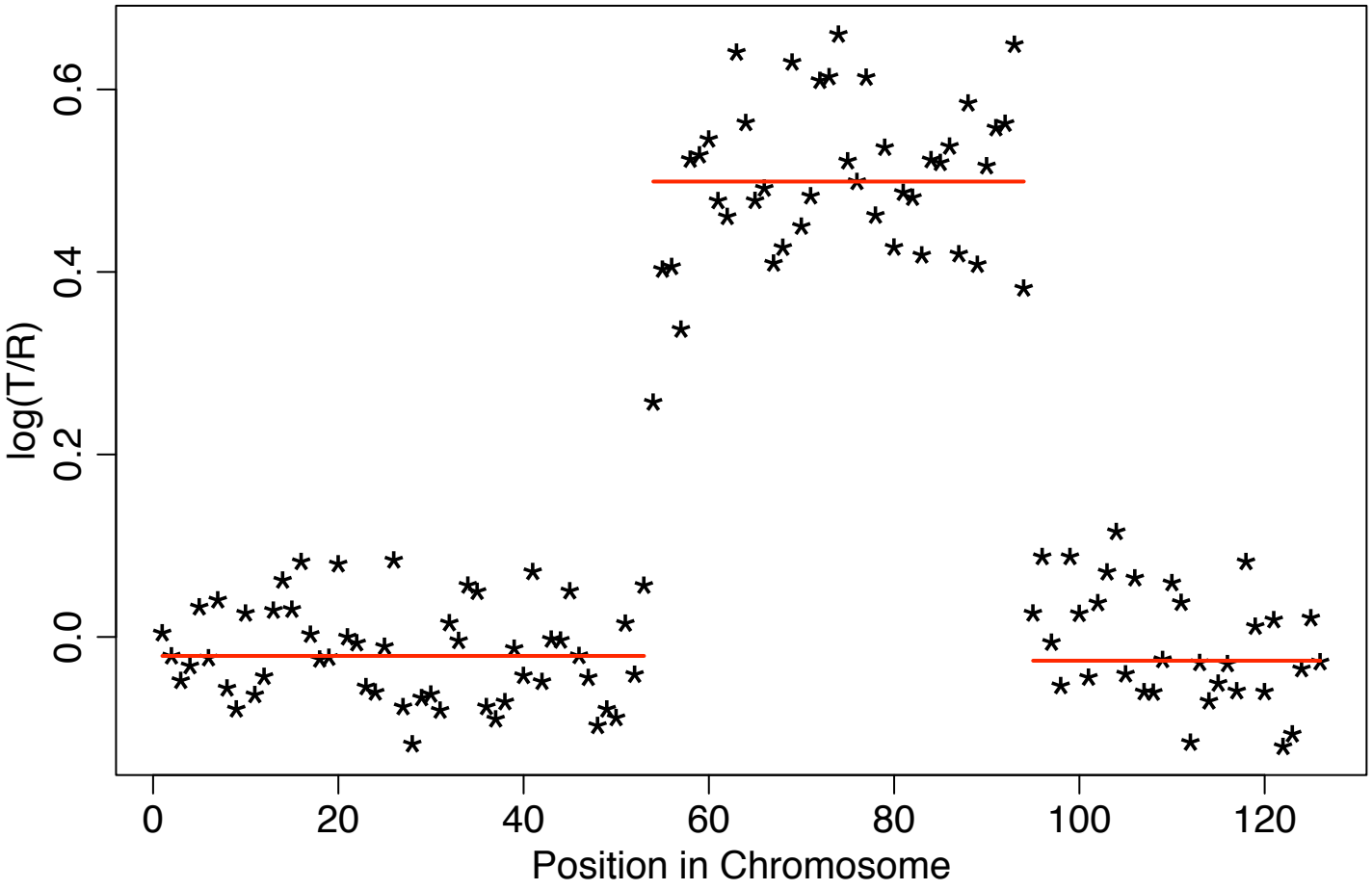
- Hodgson et al. (2001) fit a three-component Gaussian mixture model
- Bourdon et al. (2002) transformed data to be Gaussian and identified outliers
- Pollack et al. (2002) smoothed data and determined cut-offs using normal data and FDR
- Olshen and Venkatraman and Fridlyand et al. explored hidden Markov models

Let $Z_1, Z_2, \ldots, Z_K$ be the data.

If $Z_1, \ldots, Z_\nu \sim F_0$ and $Z_{\nu+1}, \ldots, Z_K \sim F_1$,

then ν is a change-point.

For our data, a change-point would correspond to where the DNA copy number has changed. There may be multiple changes within a chromosome.


log(T/R)
0.0
0.2
0.4
0.6
0
20
40
60
80
100
120
Position in Chromosome

Suppose the data are $Z_1, \ldots, Z_K$. For $k : 0 < k < K$,

$$T_k = \frac{|\bar{Z}_k - \bar{Z}_{K-k}|}{\sigma\sqrt{(K-k)^{-1} + k^{-1}}},$$

where $\bar{Z}_k = \sum_{i=1}^{k} Z_i/k$ and $\bar{Z}_{K-k} = \sum_{i=k+1}^{K} Z_i/K$.

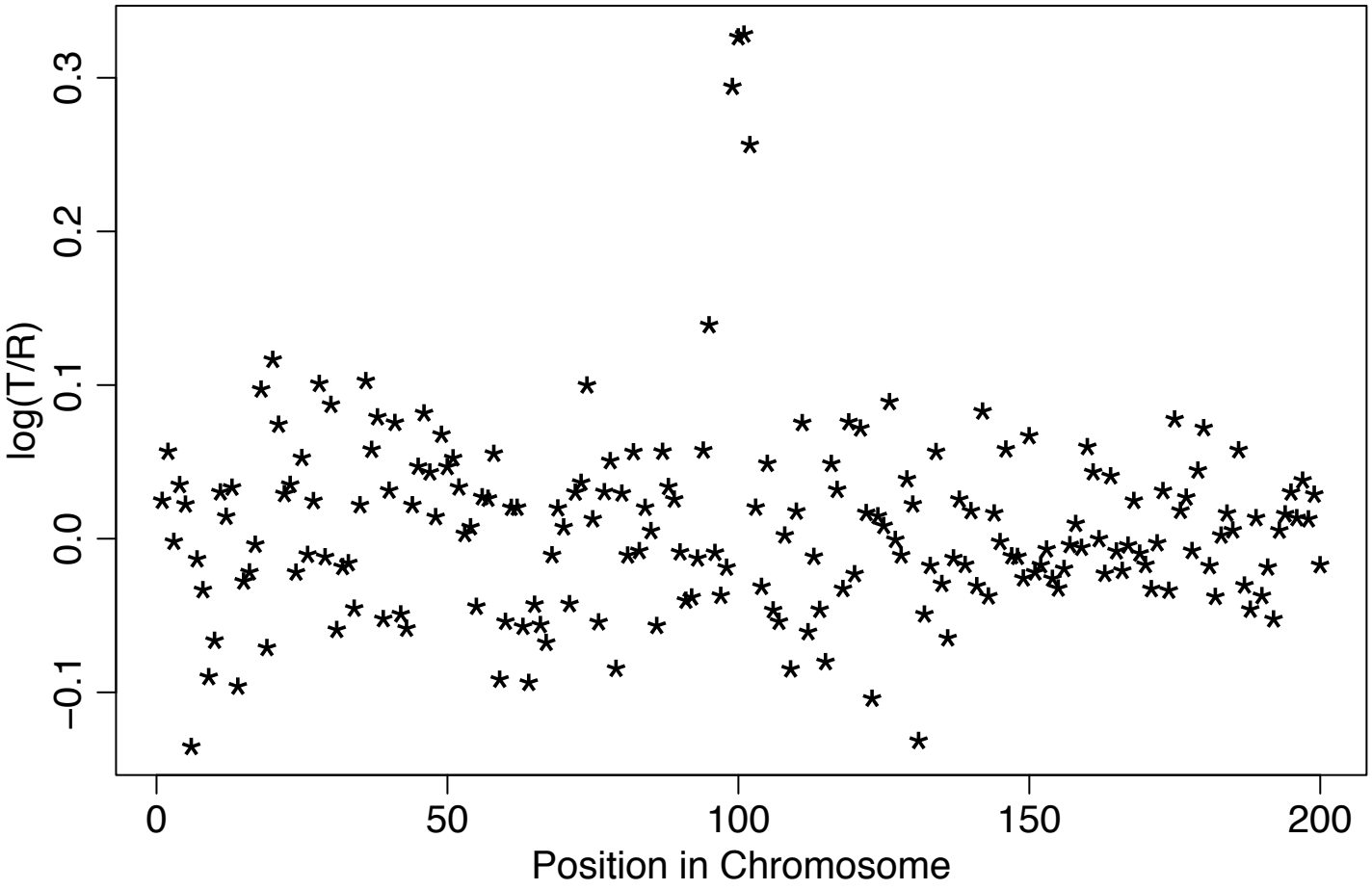
To test for a change-point, the test statistic is

$$T^* = \max_{0<k<K} |T_k|.$$

Binary segmentation (Sen and Srivastava 1975; Vostrikova 1981):

1. Split the data at k^* if T^* exceeds threshold

2. Continue splitting until no segment can be split


log(T/R)
0.3
0.2
0.1
0.0
−0.1
0
50
100
150
200
Position in Chromosome

We find the maximum of

$$T_{k1,k2} = \frac{|\bar{Z}_{k_2-k_1} - \bar{Z}_{k_1,k_2}|}{\sigma\sqrt{(k_2-k_1)^{-1} + (k_1+K-k_2)^{-1}}},$$

for $1 \leq k_1 < k_2 \leq K$, where

$$\bar{Z}_{k_2-k_1} = \frac{\sum_{i=k_1+1}^{k_2} Z_i}{(k_2-k_1)}.$$

and

$$\bar{Z}_{k_1,k_2} = \frac{(\sum_{i=1}^{k_1} Z_i + \sum_{i=k_2+1}^{K} Z_i)}{(K-k_2+k_1)}$$

The reference distribution is found by permutation.

- **Repeatedly split the data into either 2 or 3 segments. If the split is ternary, the two end segments are grouped together when determining the maximum statistic.**
- **We are performing binary segmentation as if the data were in a circle. Thus we call this procedure**

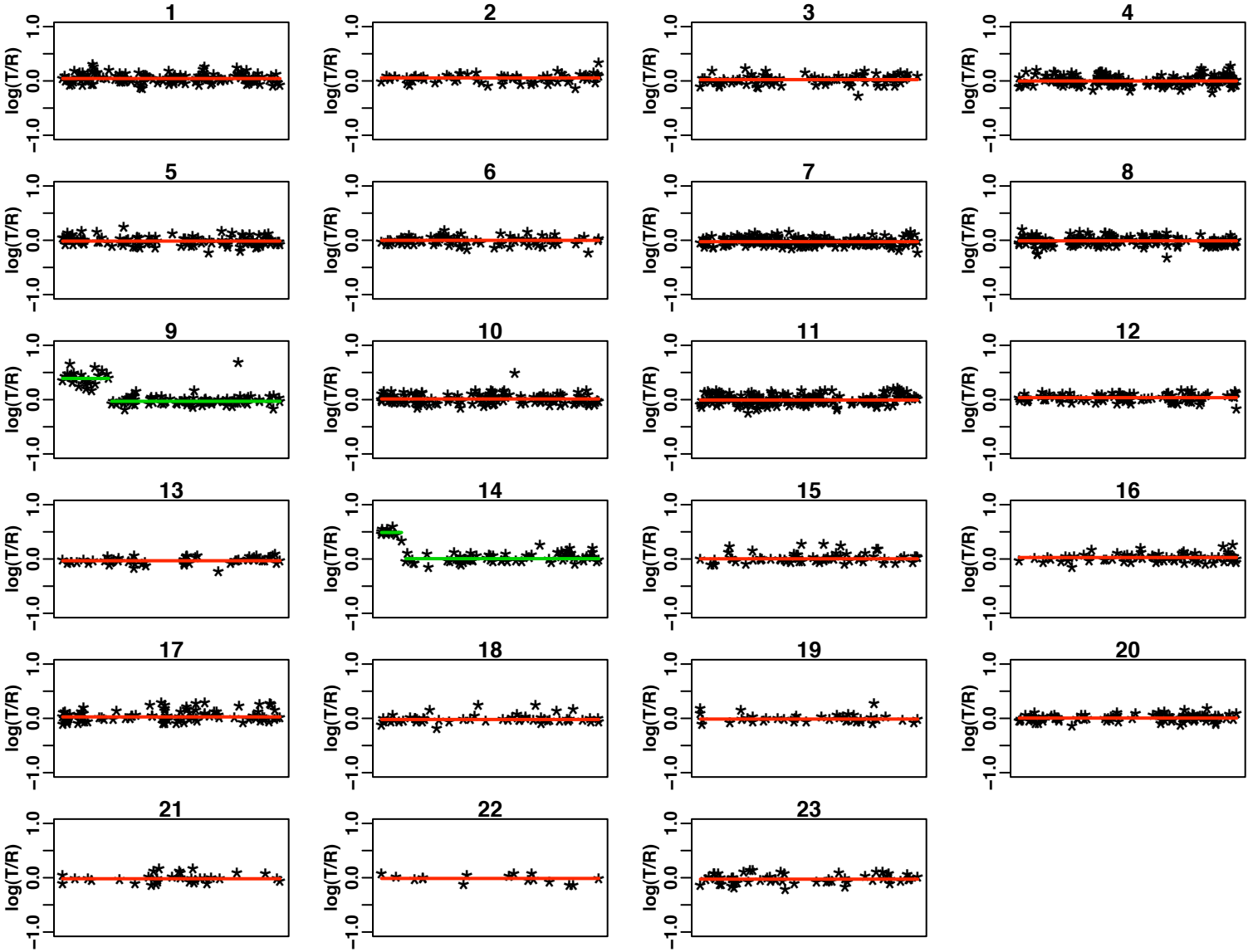
  circular binary segmentation **(CBS).**

- If ternary split, check first and third segment by test of binary split with middle segment
- Outliers are smoothed before segmentation
- Overlapping windows are used if more than $500 - 1000$ markers; our method is now a hybrid of binary segmentation and sequential methods (Page 1957)

- **The data we have examined are discussed in Snijders et al. (2001).**
- **The arrays have 2460 BACs, mapped, spotted in triplicate.**
- **There are 15 fibroblast cell lines. They have been identified as containing monosomies or trisomies, on either parts or all of 1-2 chromosomes.**

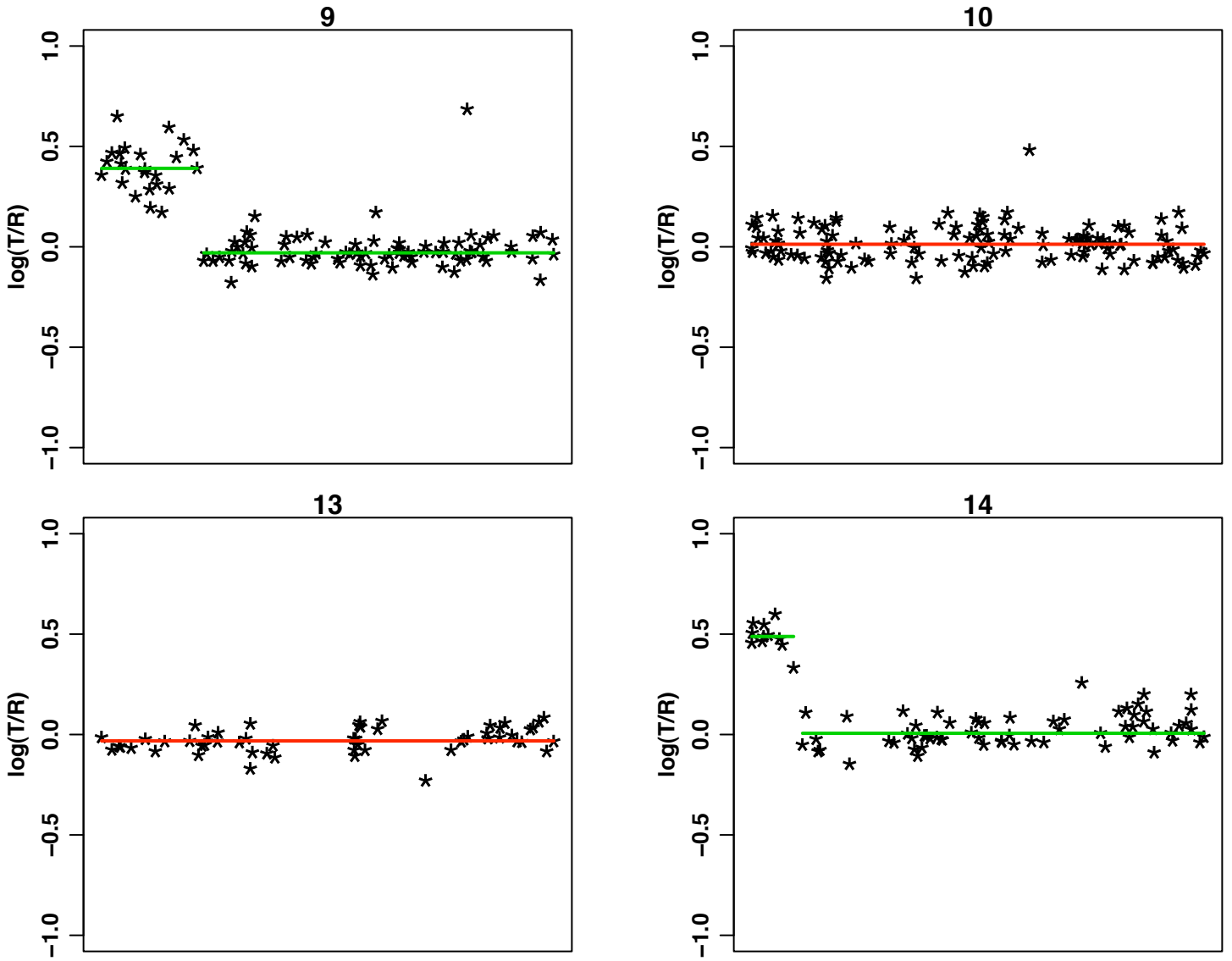
DNA Copy Number

An Example



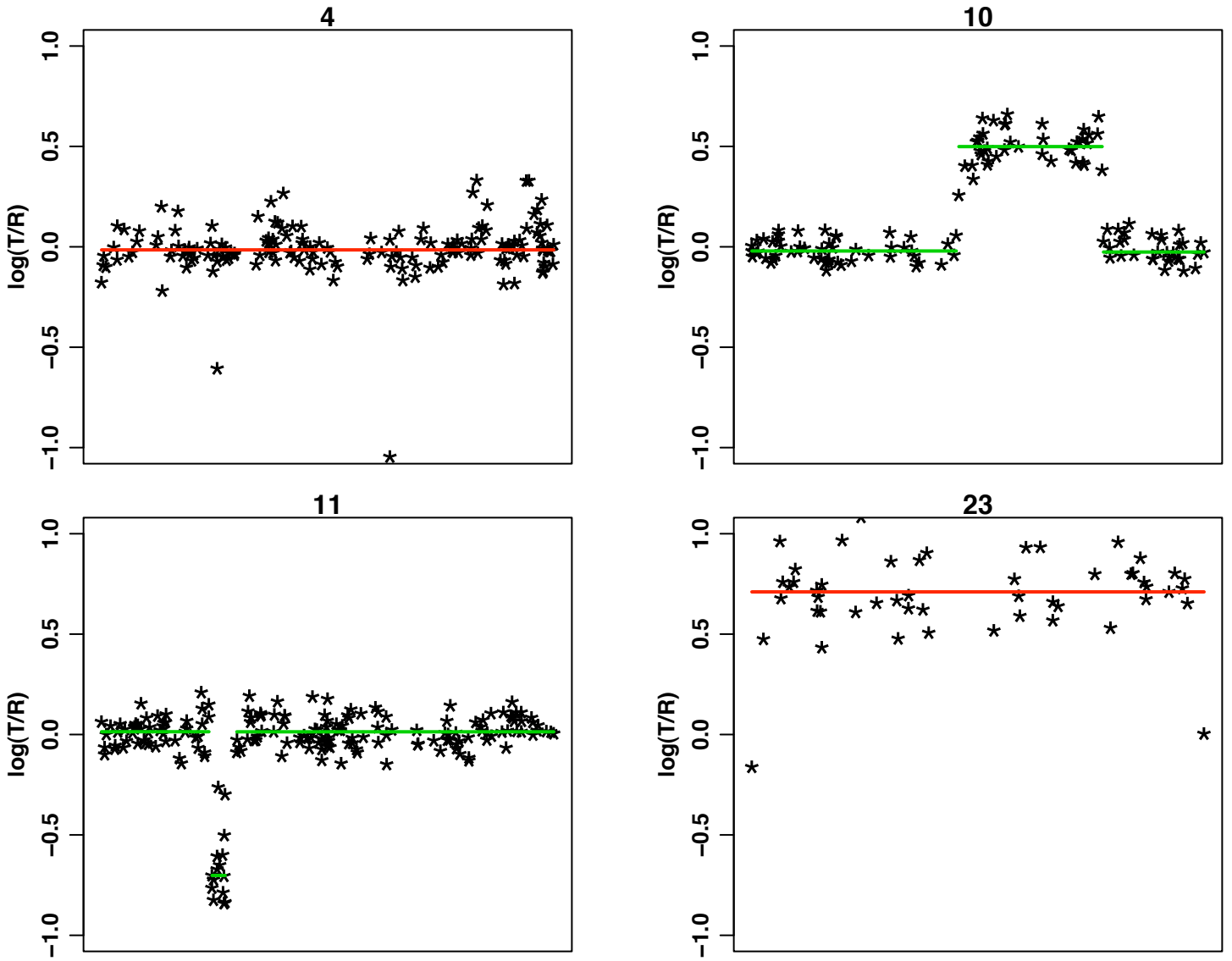
DNA Copy Number

An Example



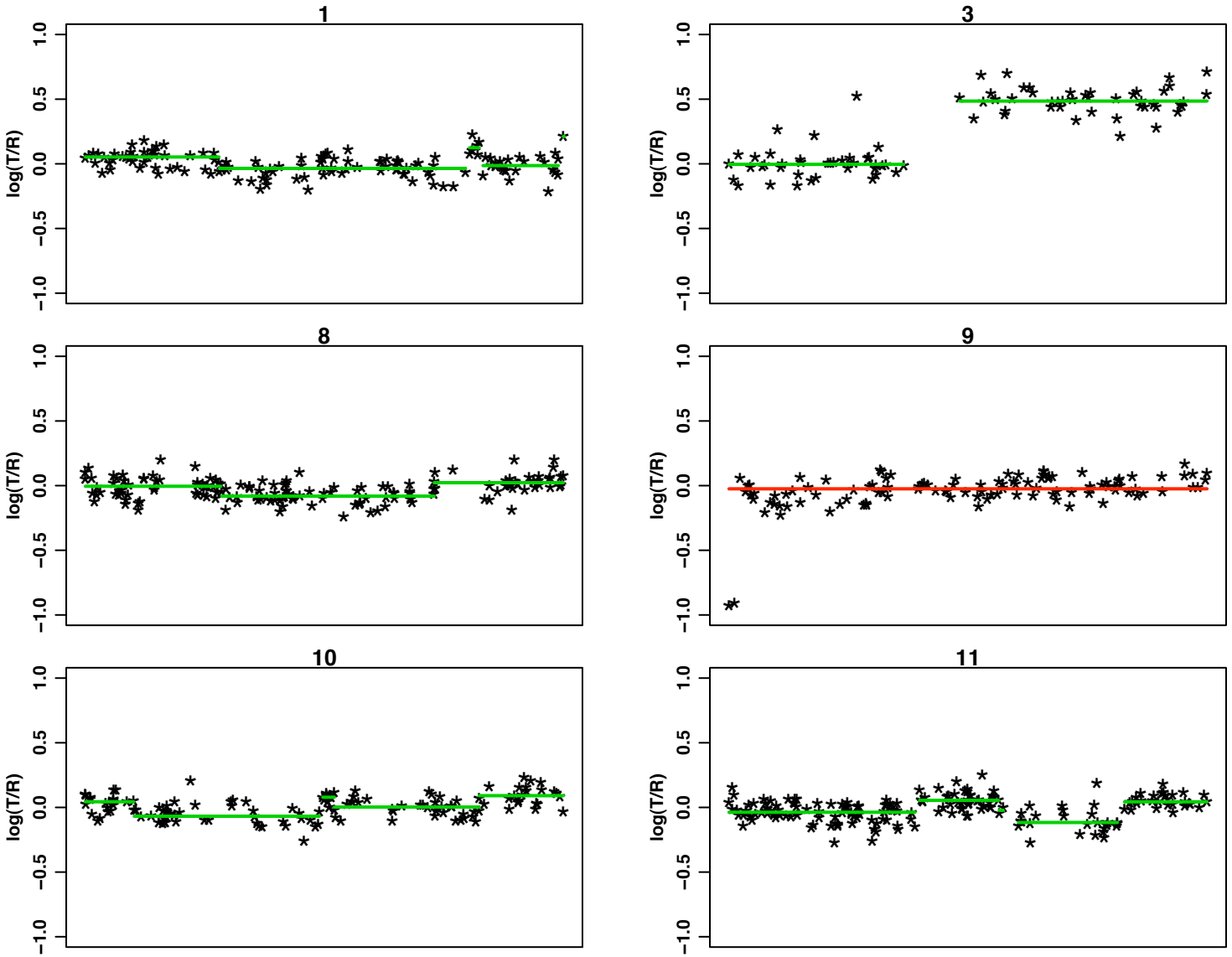
DNA Copy Number

Another Example



DNA Copy Number

A Final Example



Suppose there are N change-points after CBS.

1. **Find the best set of change-points $n = 1, \ldots, N$, choosing only among those previously identified**

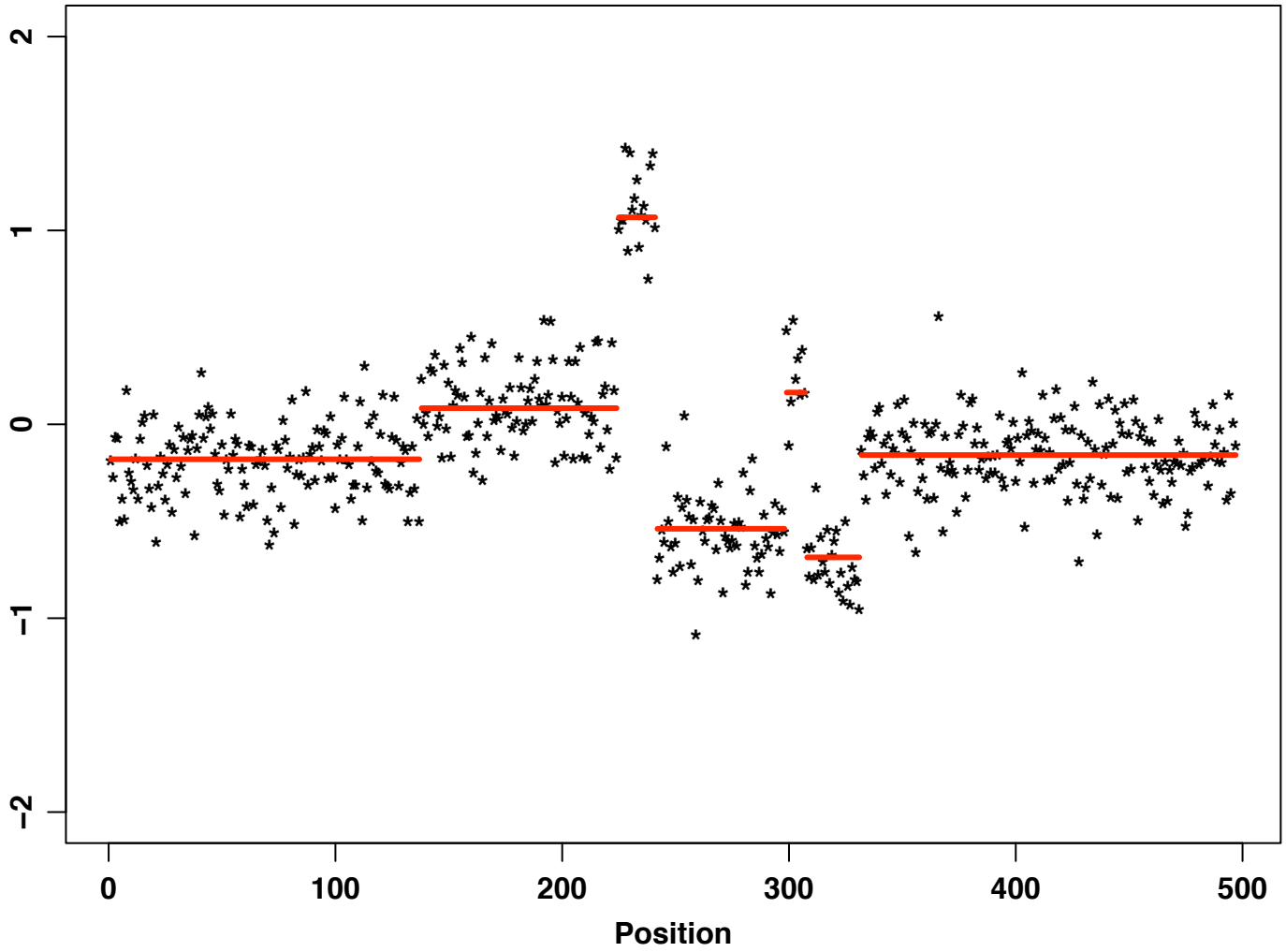
2. **Choose the smallest n such that $SS_n/SS_N - 1 < c$, where c is some pre-specified constant (such as 0.05)**

   (SS_n equivalent to the error sum of square in one way ANOVA; If n change-points, $n + 1$ groups)

- **Start with a step function from a CBS fit to chromosome 11 from a real breast cancer study**
- **Add Gaussian noise with SD estimated from same data**
- **Apply CBS**
- **Repeat process 100 times**

DNA Copy Number

Simulated Data



There were 6 change-points in the step function.

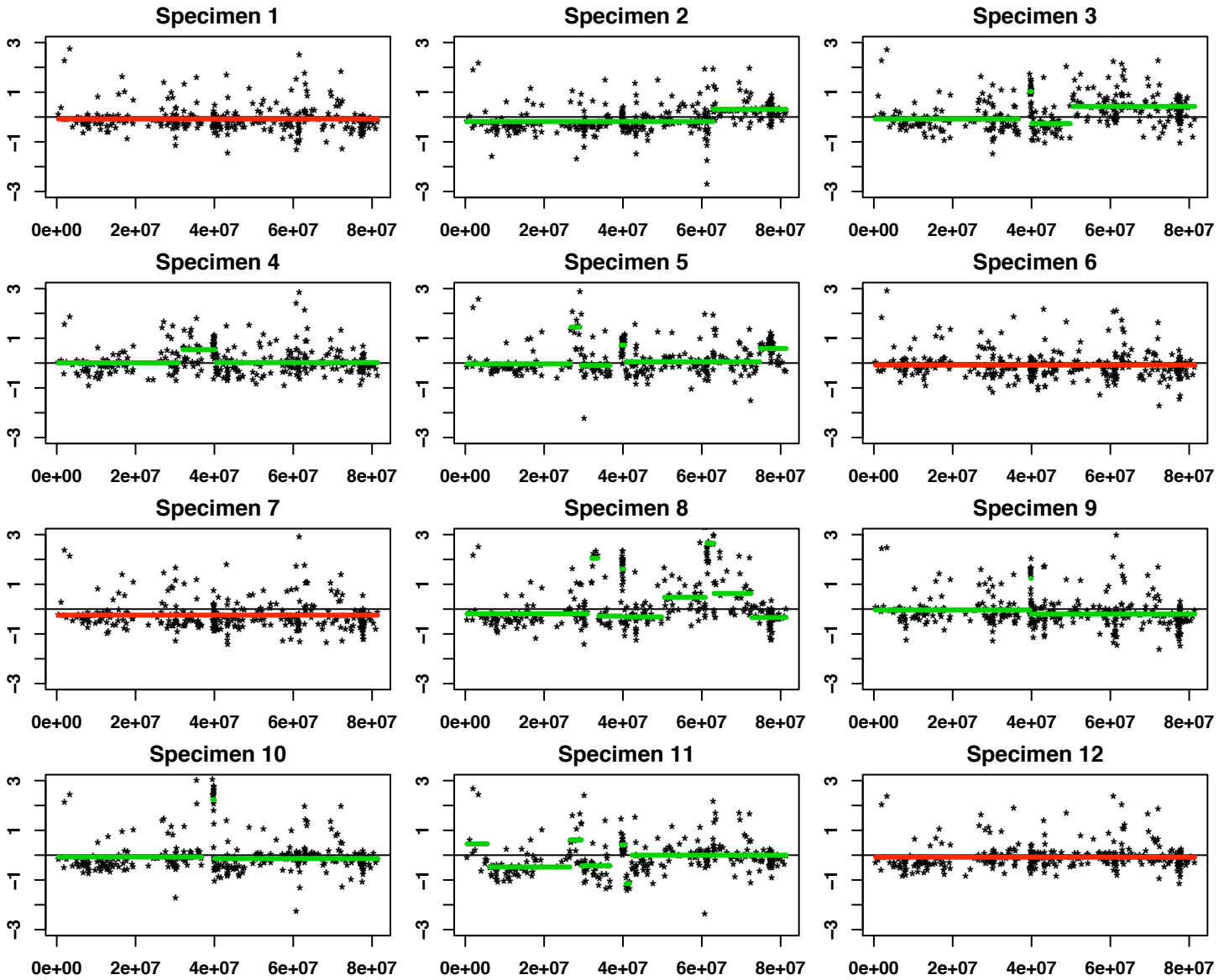
Change-points	α level for permutation 0.05	0.01	0.005	0.0001
6	18 (97)	39 (100)	39 (98)	50 (100)
7	16 (2)	32 (0)	43 (2)	46 (0)
8	26 (1)	18 (0)	13 (0)	4 (0)
9	19 (0)	8 (0)	4 (0)	0 (0)
10-14	21 (0)	3 (0)	1 (0)	0 (0)

- **Specimens:** 12 breast cancer tissues (thanks to Mike Wigler, Rob Lucito, and members of their labs)
- **Array:** 9820 oligonucleotides of length 70
- **Analysis:** Chromosome 17, which contains ERBB2/HER2NEU.

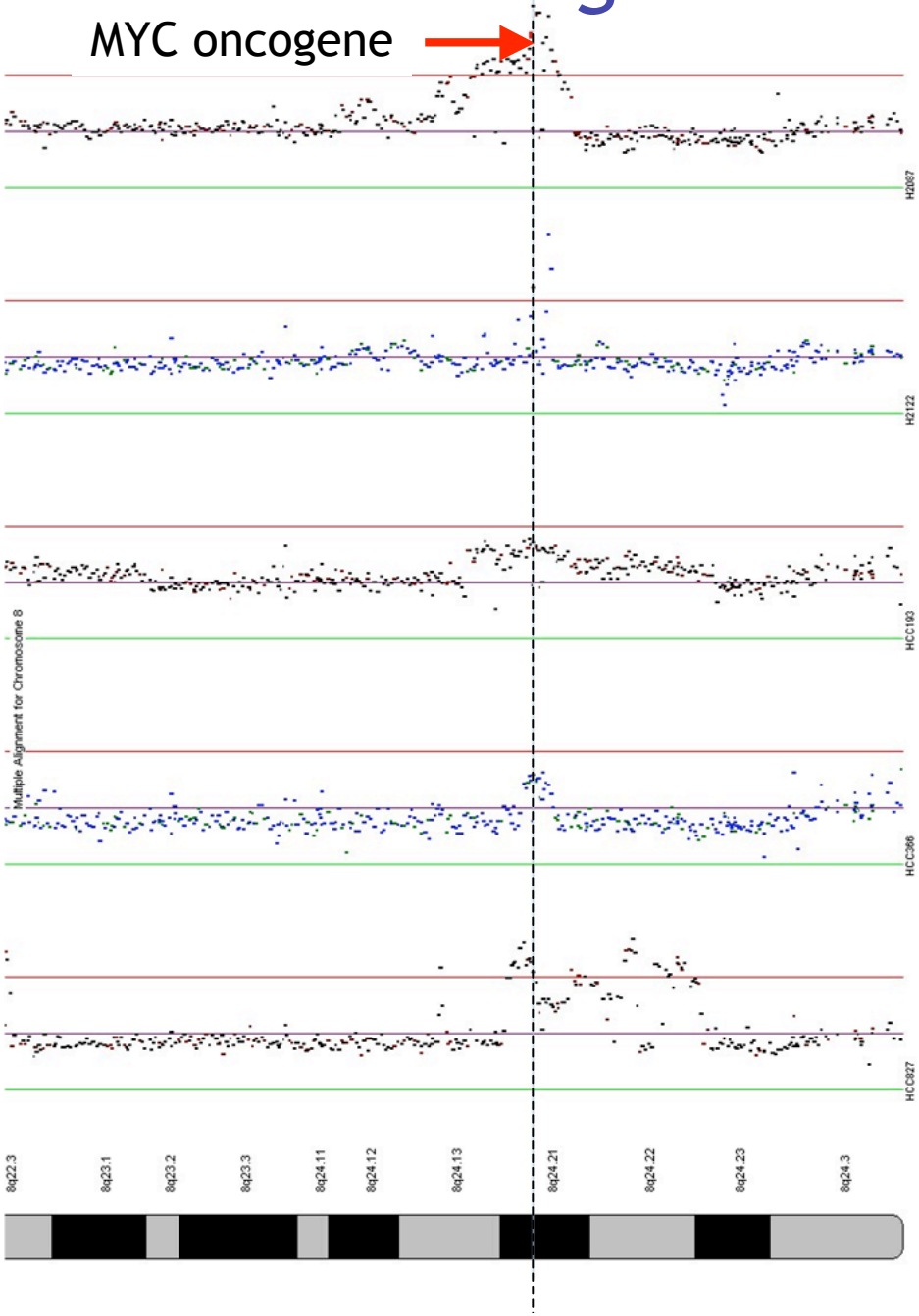
Increased HER2neu protein is associated with 30% of breast cancers, and correlates with poor prognosis and aggressive tumor growth. The drug Herceptin has been shown to slow progression and increase tumor shrinkage.

DNA Copy Number

Chromosome 17



A recurrent CNA in 5 lung cancer cell lines



Goal: Detecting recurrent CNAs from multiple aCGH samples

Goal: find CNAs that are common to a set of samples to determine which CNAs are contributing to disease

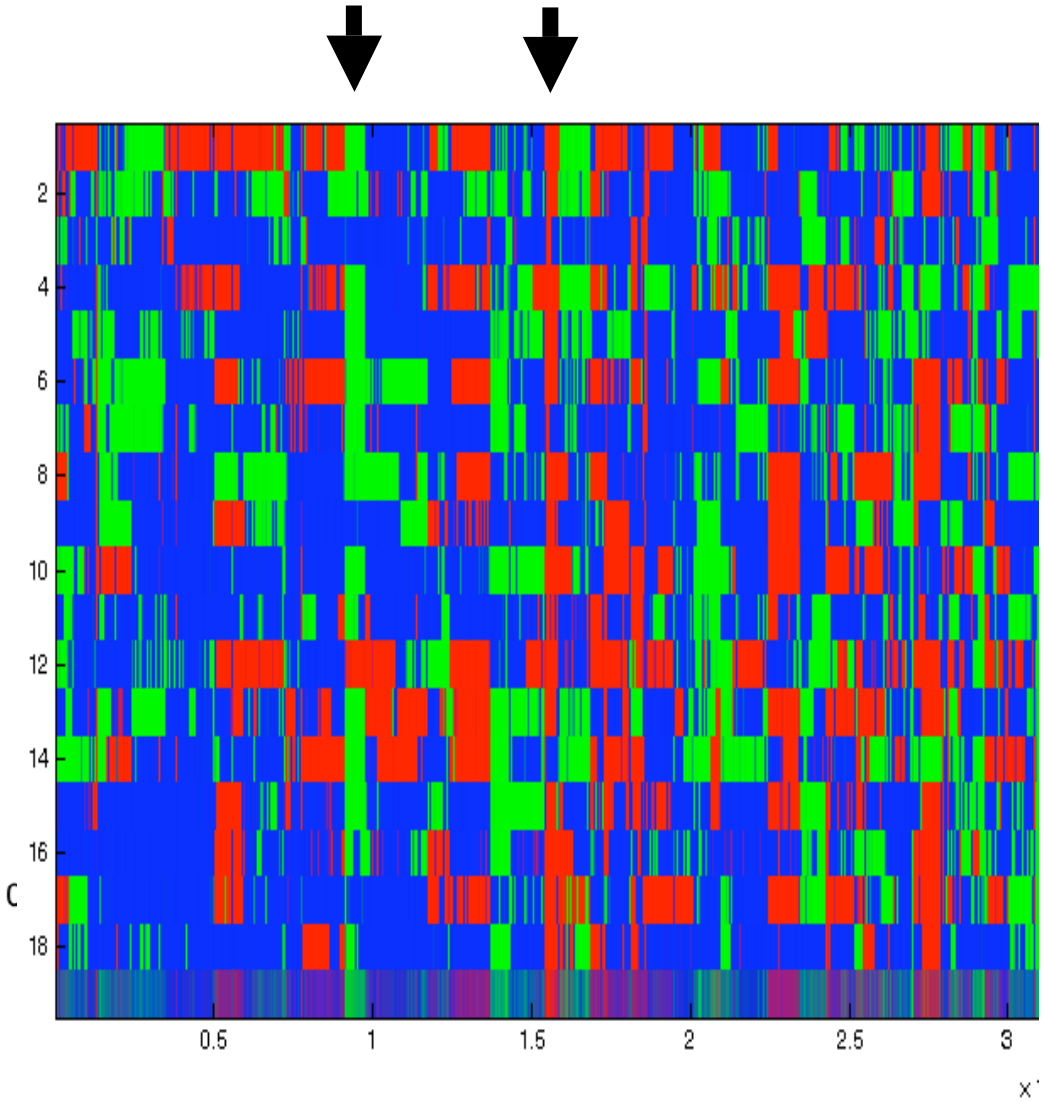
- Shared CNAs define a molecular characterization, or *profile* of the particular phenotype

Challenges:

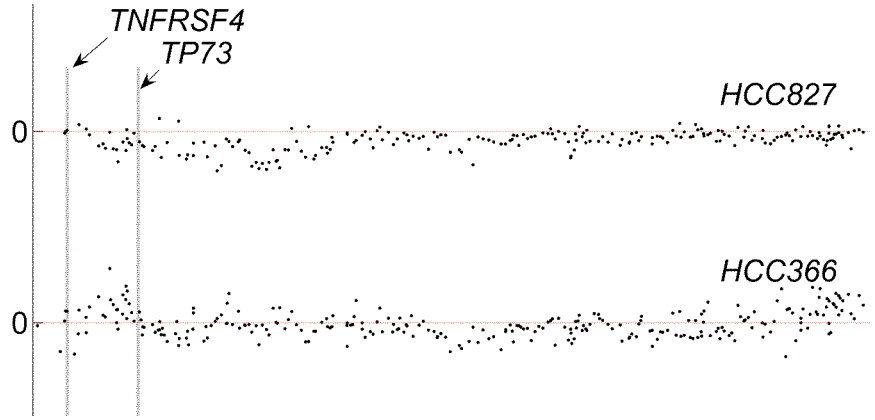
- Not all samples will exhibit CNAs in the same place
- Only a small proportion of locations will show recognizable patterns
- Frequency of occurrence in those locations is variable

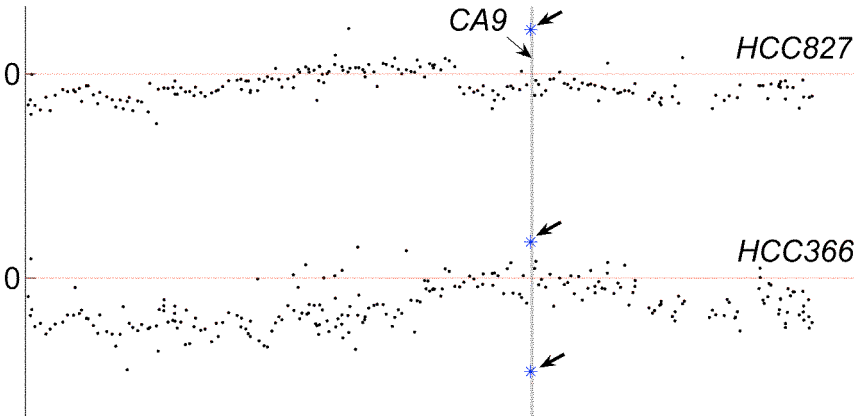
Contributions:

- Studying and comparing joint models to discover recurrent CNAs

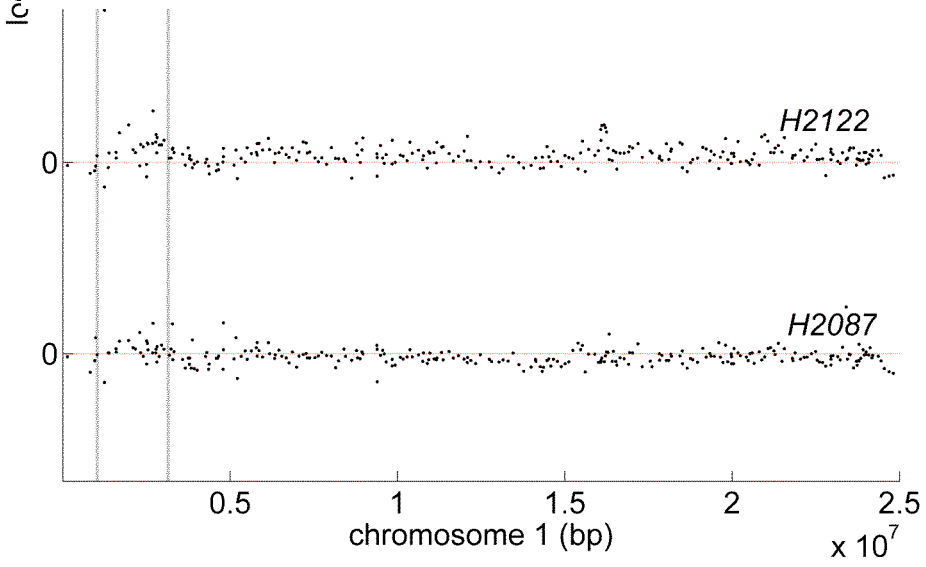


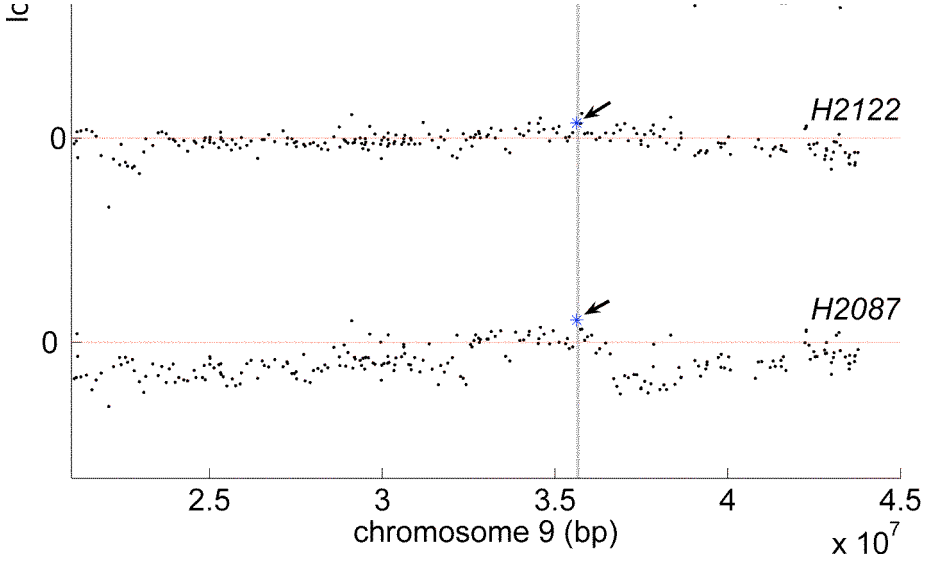
Signals missed by AF





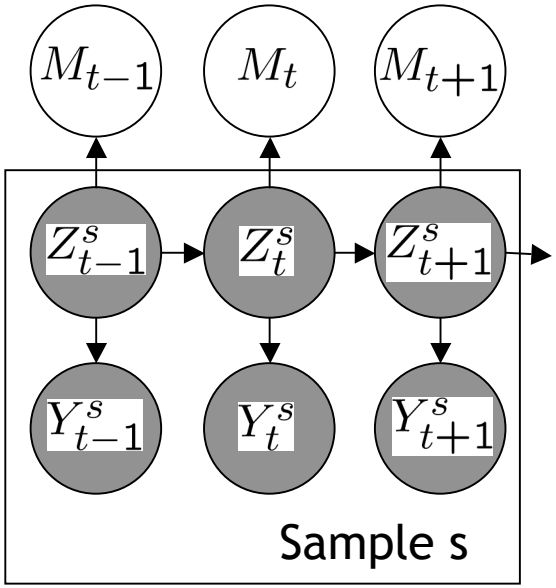
Borrow statistical strength across samples





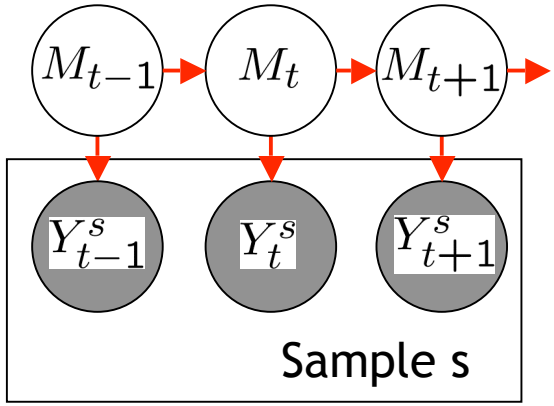
Models for recurrent CNAs

$M_t, Z_t \in \{L, N, G\}$



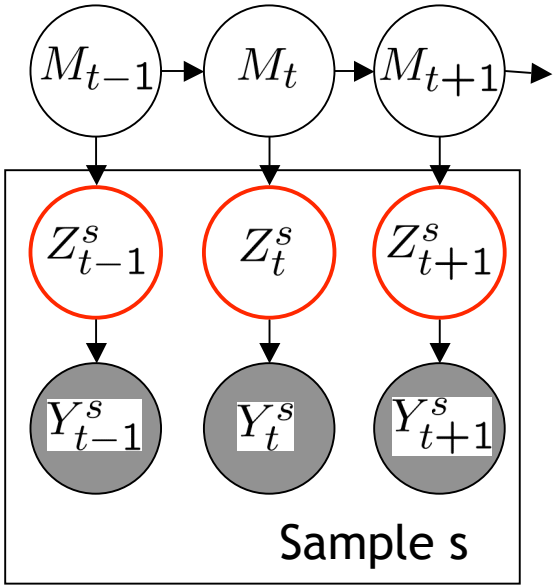
Alteration frequency (AF):
Discretize each sample separately and summarize

☑Use HMM on each sample
✘ Premature thresholding
✘ Cannot 'share' information



Factored likelihood HMM (FL-HMM):
Use a 'joint' emission model to infer M from the raw data

☑ No discretization
✘ One sample can dominate

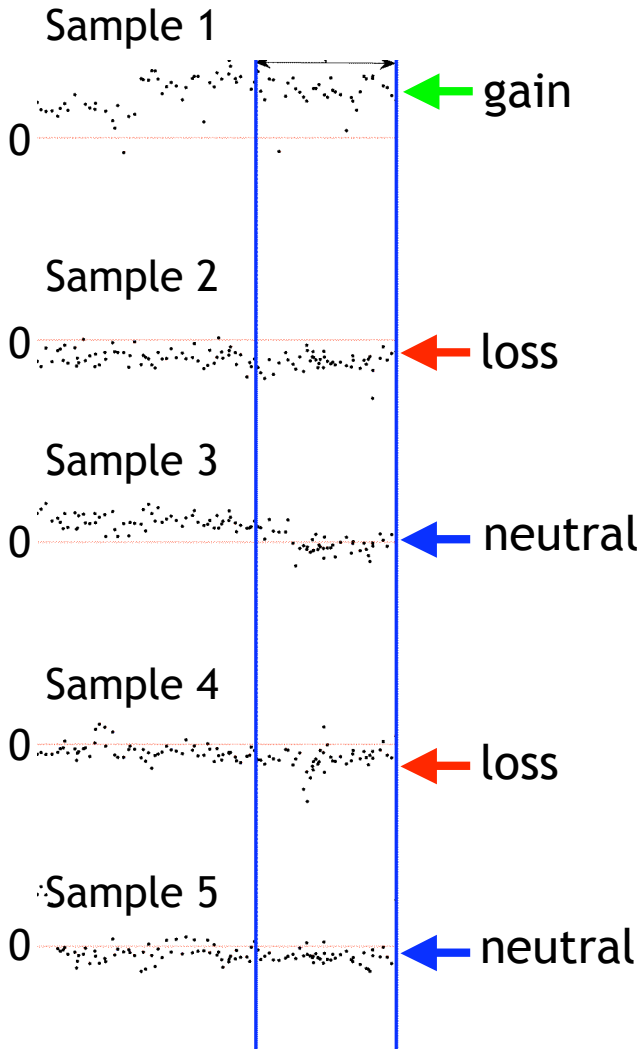


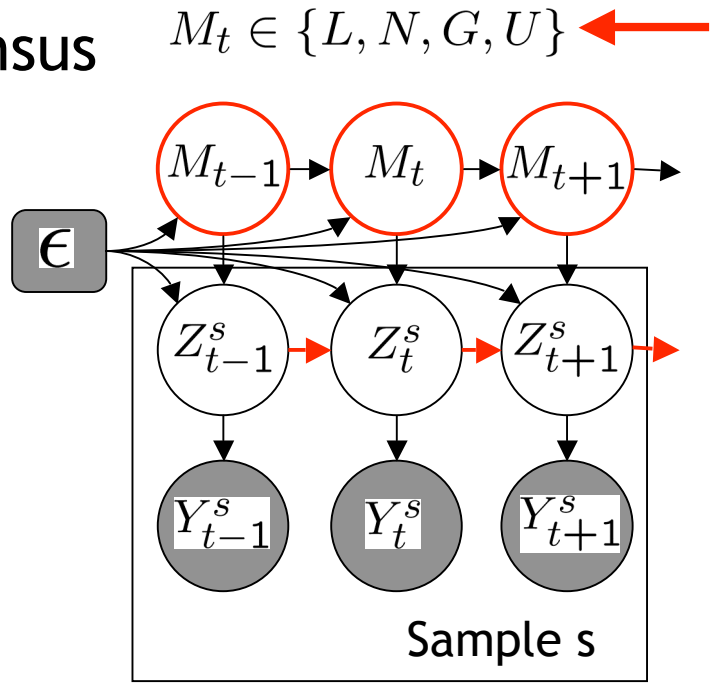
Buffered FL-HMM (BFL-HMM):
Simultaneously infer M and Z

☑ 'Buffer' the master
✘ Ambiguous locations?

Considering ambiguous regions

Many clones will not have consensus
What should the master do?



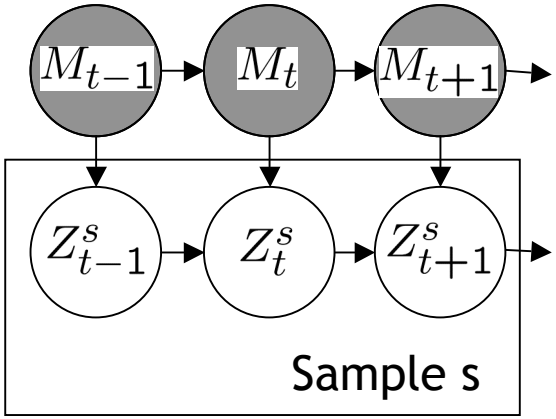


Hierarchical HMM (H-HMM):
Allow M to 'opt-out' of L,N,G states

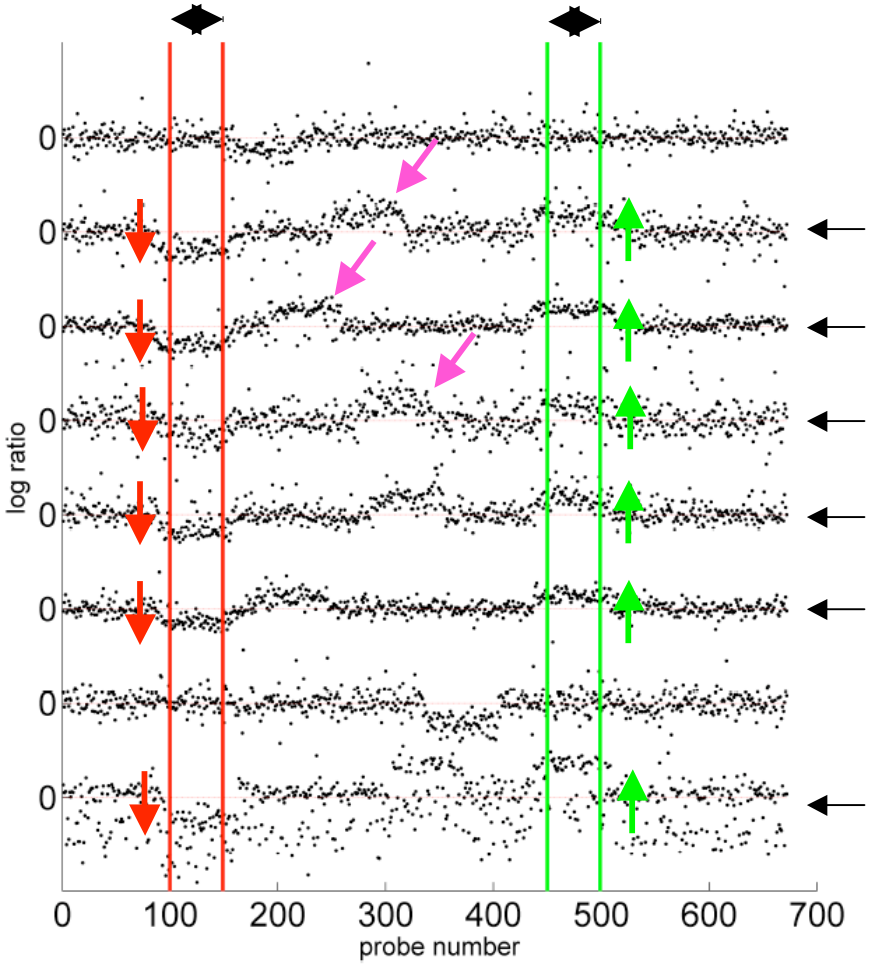
☑ Explicitly model ambiguity
☑ Leads to more accurate and interpretable output
☑ Exert an adjustable level of control over Zs

H-HMM Inference

- The Z^s chains are coupled which makes inference hard
- Conditioned on M, Z^s are independent and can be updated in parallel
- Use Blocked Gibbs MCMC sampling
  - Compute p(M|Z,θ,Y)
  - See paper for details
- Running time is O(TSN)
  - Number of data points and N MCMC samples
  - Considerably slower than other models due to Z chains



Synthetic data experiment



Data based on real data from 8 MCL cell lines and insert recurrent CNAs

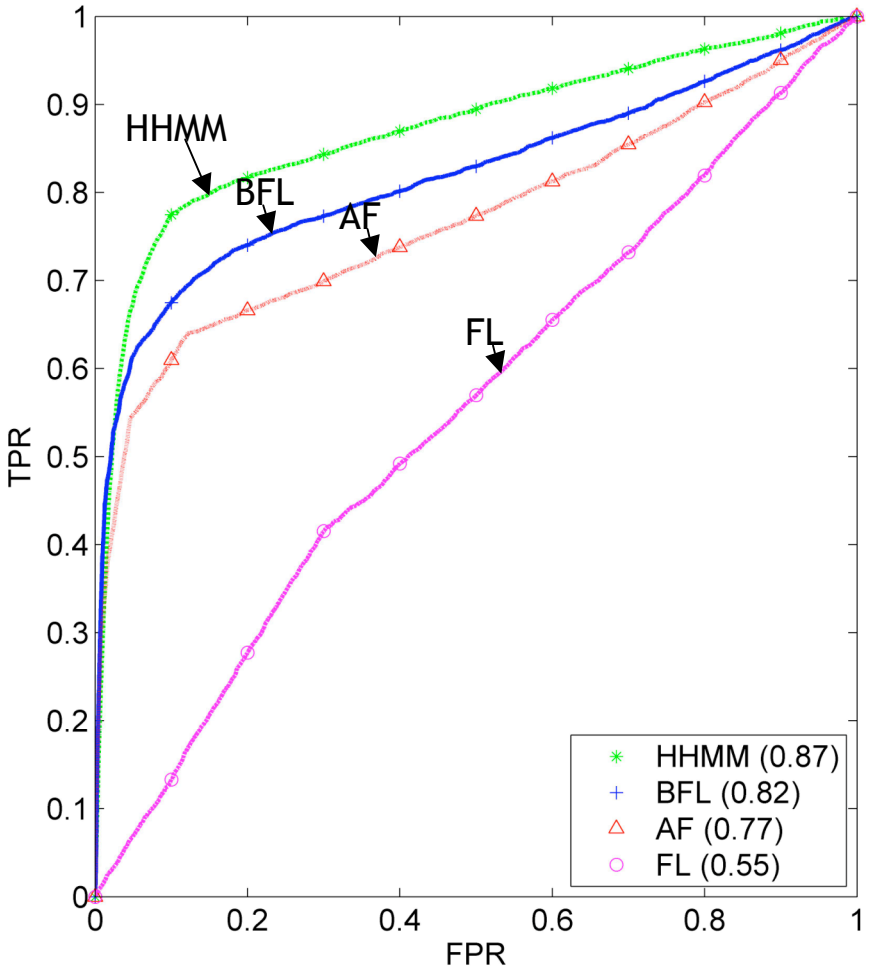
Alter:

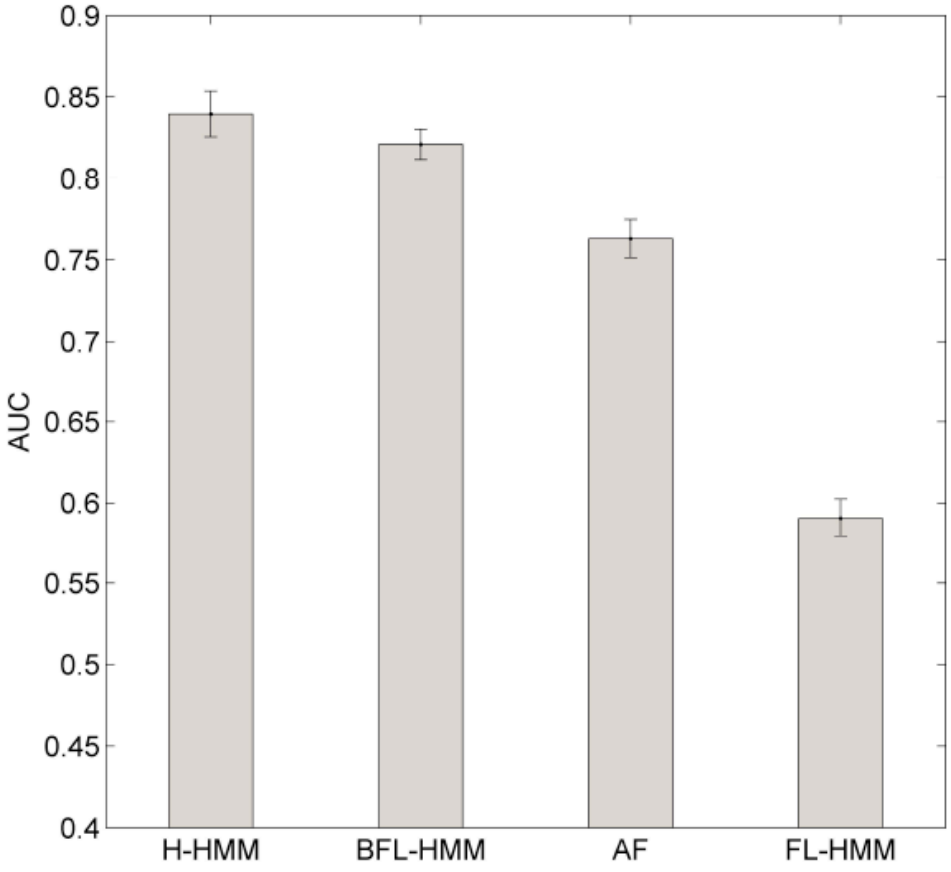
- Width
- Height
- Frequency
- Spike in random CNAs
- Repeat 3 times

- ~100 data sets, ~600 probes each

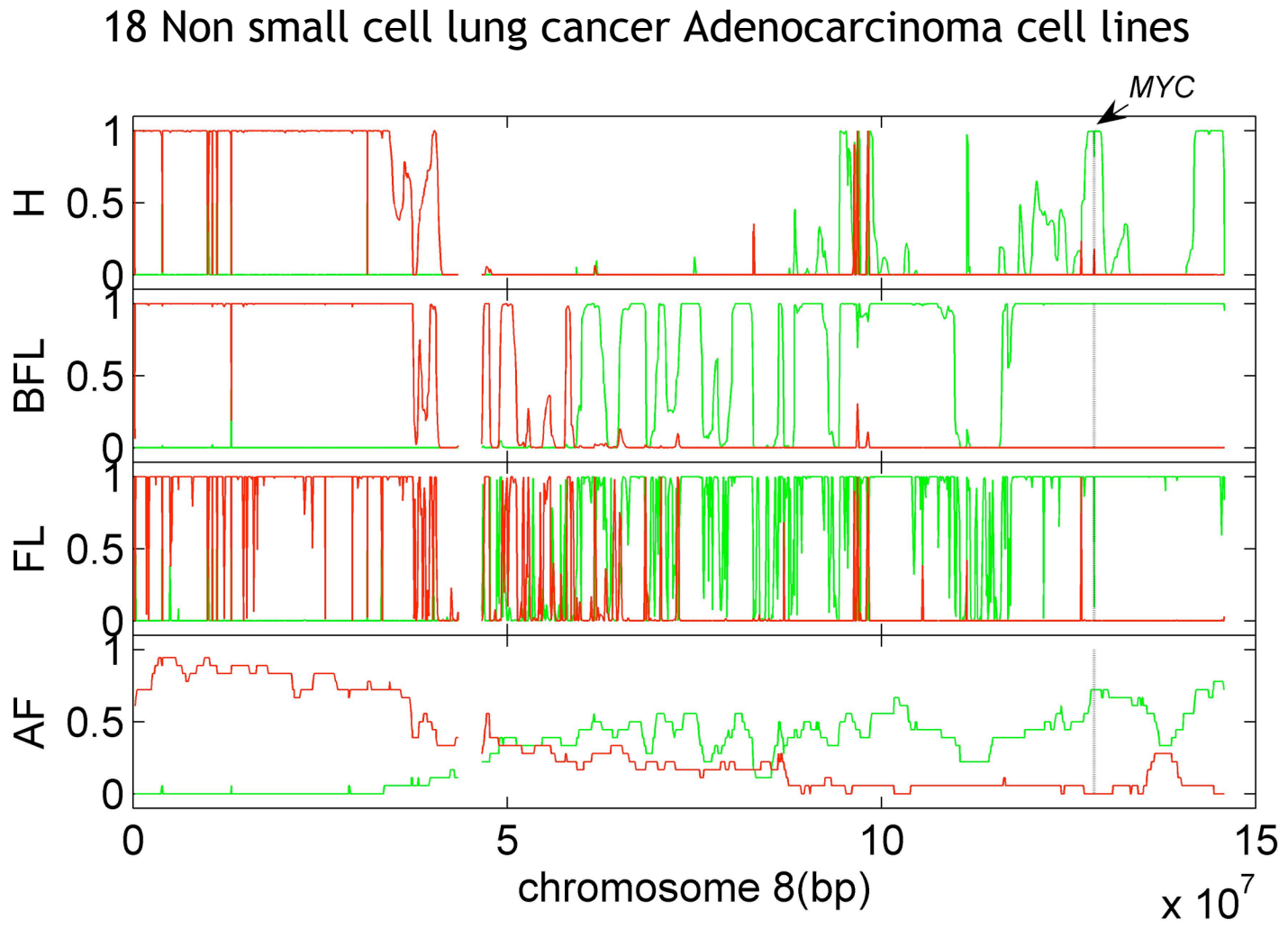
- ROC analysis measuring accuracy in predicting recurrent CNAs

H-HMM is quantitatively better than other models



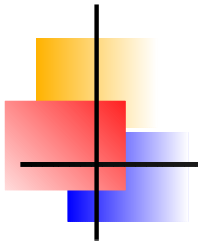


Qualitatively, the H-HMM is sparse, yet accurate



Summary

- Developed and compared 3 new statistical models to detect recurrent CNAs in array CGH data
  - Infer a profile representing canonical locations without first discretizing the data
- H-HMM was quantitatively and qualitatively better than other models and the standard approach
- Evaluating results in large scale study and refining the model to detect low-frequency CNAs
- More info: http://www.cs.ubc.ca/~sshah/acgh



Problem at hand

- Consider a chromosome with $n+1$ probes, indexed by $i = 0, 1, ..., n$.
- Define a partition of $\mathcal{A}$ as $P(\mathcal{A}) \equiv \{c_0 = 0, c_1, c_2, \ldots, c_K = n\}$, in which $c_k \in [0, n]$ are change points that break $\mathcal{A}$ into K segments.
- Denote a segment by $\Delta_k = (c_k, c_{k+1}]$. Then $\mathcal{A} = \cup_{k=0}^{K-1} \Delta_k$.
- The goal is to find an appropriate partition and estimate the copy number state, denoted as μ_k, for each segment of the chromosome defined by the partition.

Probability

- Denote Y_{ij} the $\log_2$ ratio of the copy numbers of probe i in array $j = 1, ..., J$.
- Suppose that probe j is in segment k, i.e., $i \in \Delta_k$.
- The model for Y_{ij} is given by

$$Y_{ij}|P(\mathcal{A}), \mu_k, \sigma_k^2 \sim N(\mu_k, \sigma_k^2), \quad \text{if } i \in \Delta_k.$$

which can be rewritten as

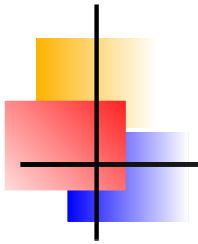
$$Y_{ij}|P(\mathcal{A}), \mu_k, \sigma_k^2 \sim N\left(\sum_{k=0}^{K-1} \mu_k I_{(c_k \le i < c_{k+1})}, \sum_{k=0}^{K-1} \sigma_k^2 I_{(c_k \le i < c_{k+1})}\right) \quad (1)$$

where $I_{()}$ is the indicator function. The full likelihood function is given by

Likelihood

$$Like = \prod_{j=1}^{J} \prod_{i=1}^{n} \left(2\pi \sum_{k=0}^{K-1} \sigma_k^2 I_{(c_k \le i < c_{k+1})} \right)^{-1/2} \times$$

$$\exp \left[- \frac{\left(Y_{ij} - \sum_{k=0}^{K-1} \mu_k I_{(c_k \le i < c_{k+1})} \right)^2}{2 \sum_{k=0} K - 1 \sigma_k^2 I_{(c_k \le i < c_{k+1})}} \right] . \qquad (2)$$



Priors

- Prior for the partition $P(\mathcal{A})$.
- Prior for the copy number state μ_k – parametric or nonparametric.
- Prior for other parameters.