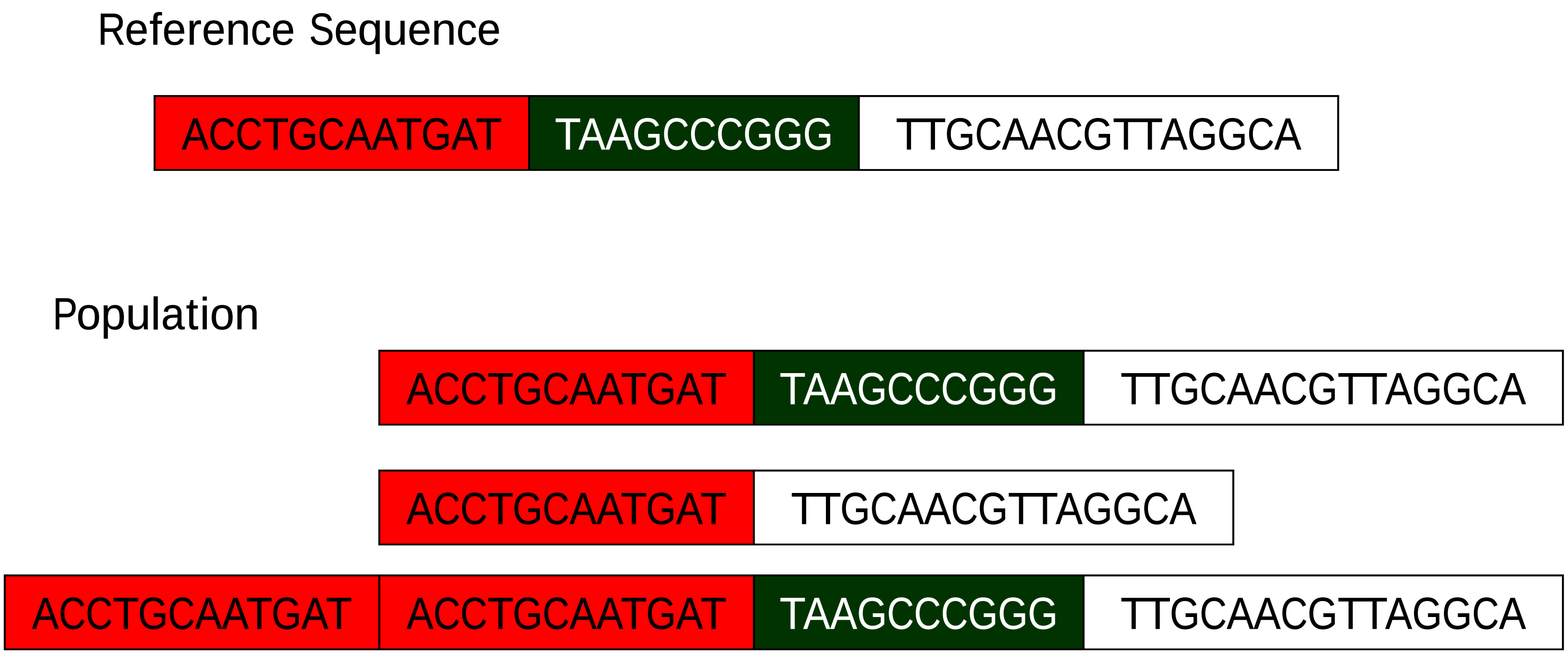
Copy Number Variation Methods and Data

Copy number variation (CNV)



Reference Sequence

ACCTGCAATGAT TAAGCCCGGG TTGCAACGTTAGGCA

Population

ACCTGCAATGAT TAAGCCCGGG TTGCAACGTTAGGCA

ACCTGCAATGAT TTGCAACGTTAGGCA

ACCTGCAATGAT ACCTGCAATGAT TAAGCCCGGG TTGCAACGTTAGGCA

Copy number variations (CNVs) are regions >1kb in a genome that occur in different copy number in a population.

CNVs in Cancer Cells

- Development of solid tumors is associated with acquisition of complex genetic alterations.
- These alterations can be of any length, including full chromosomes.
- Changes can be caused by underlying failures in maintenance of genetic stability,
- Canges may be promoted by positive selection as they provide growth advantages.
- Regions commonly amplified may contain genes that improve the viability and growth of the tumor cells.

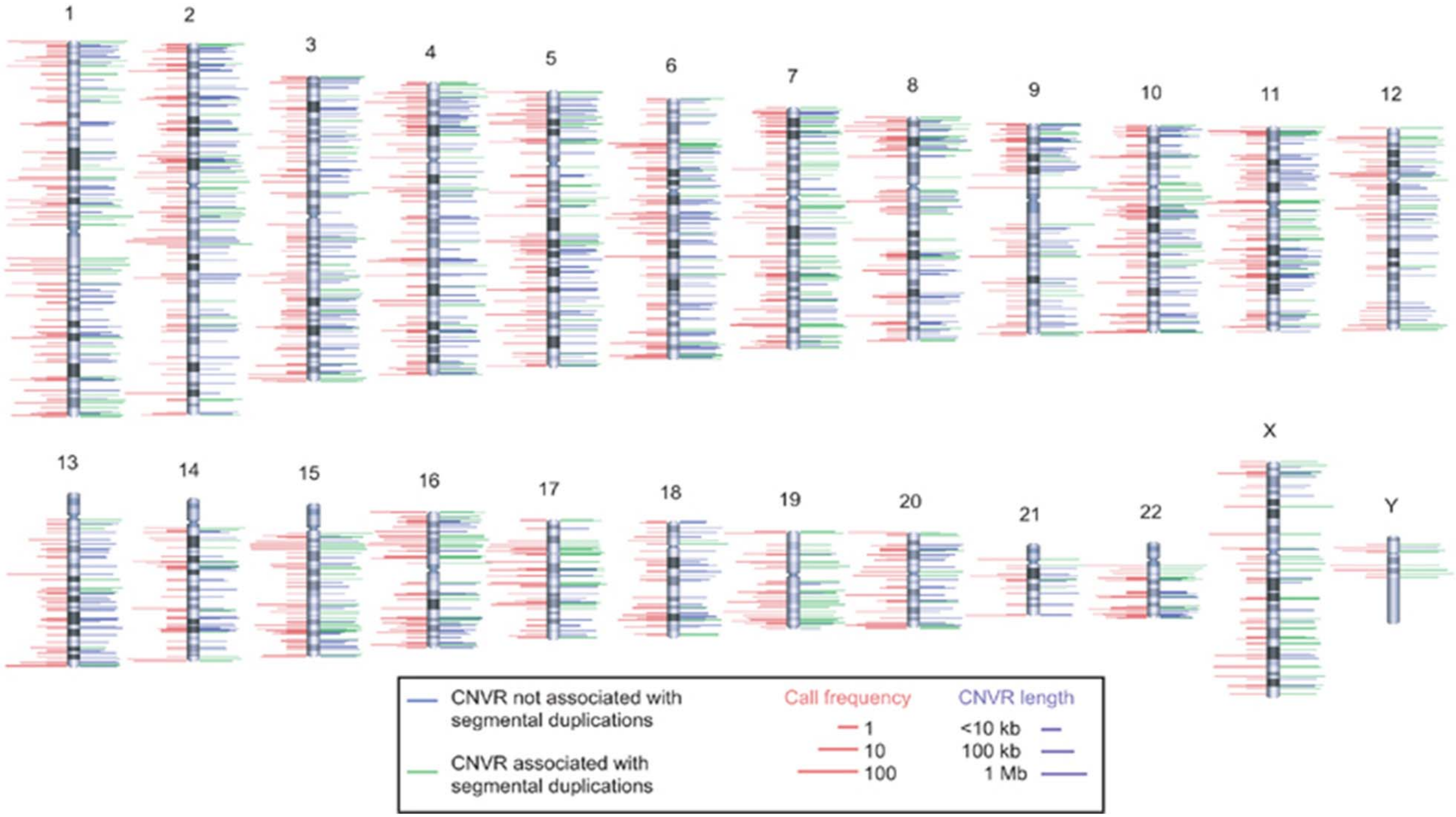
CNVs in populations

- CNVs ranging from 1kb - several Mb segregate in humans and model organisms.
- Median length of known CNVs is ~100kb, but that number is shaky.
- CNVs often have limited phenotypic impact
- CNV-alleles are inherited in the germline.
- >10% of the human genome has variable copy number.
- The genome of two individuals has an average difference in length of ~10 Mb.
- CNVs cover genes (Redon et al(06): 2900 genes are covered by CNVs)

More CNV facts

- Most CNVs are singletons
- In most genomic regions, CNV-mutations are rare.
- CNVs are usually in high LD with SNPs.
- 25-fold enrichment near segmental duplications.
- CNVs are generated by several types of events, e.g. non-allelic homologous recombination and retrotransposition.
- More detected CNVs are duplications, but it is not clear if this is detection bias.

CNVs are distributed throughout the Genome

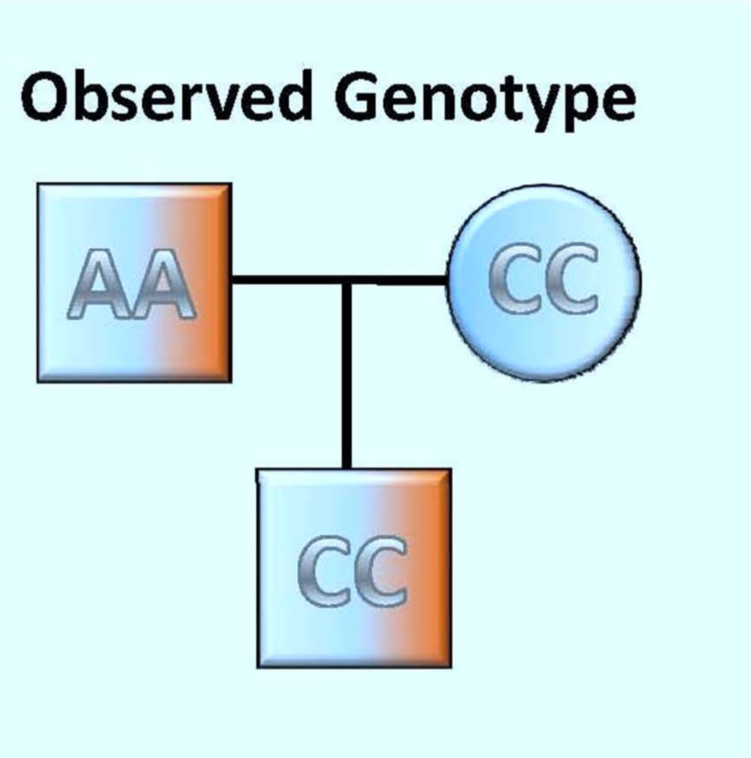


CNVs and diseases

- Aids: CNV covering CCL3L1, large copy number is protective of HIV/AIDS infection.
- Association with Crohn’s disease and BMI.
- Autism: De novo deletions more common in patients with Autism.
- New Syndrome: Deletion on chromosome 17 defines novel genetic disorder.

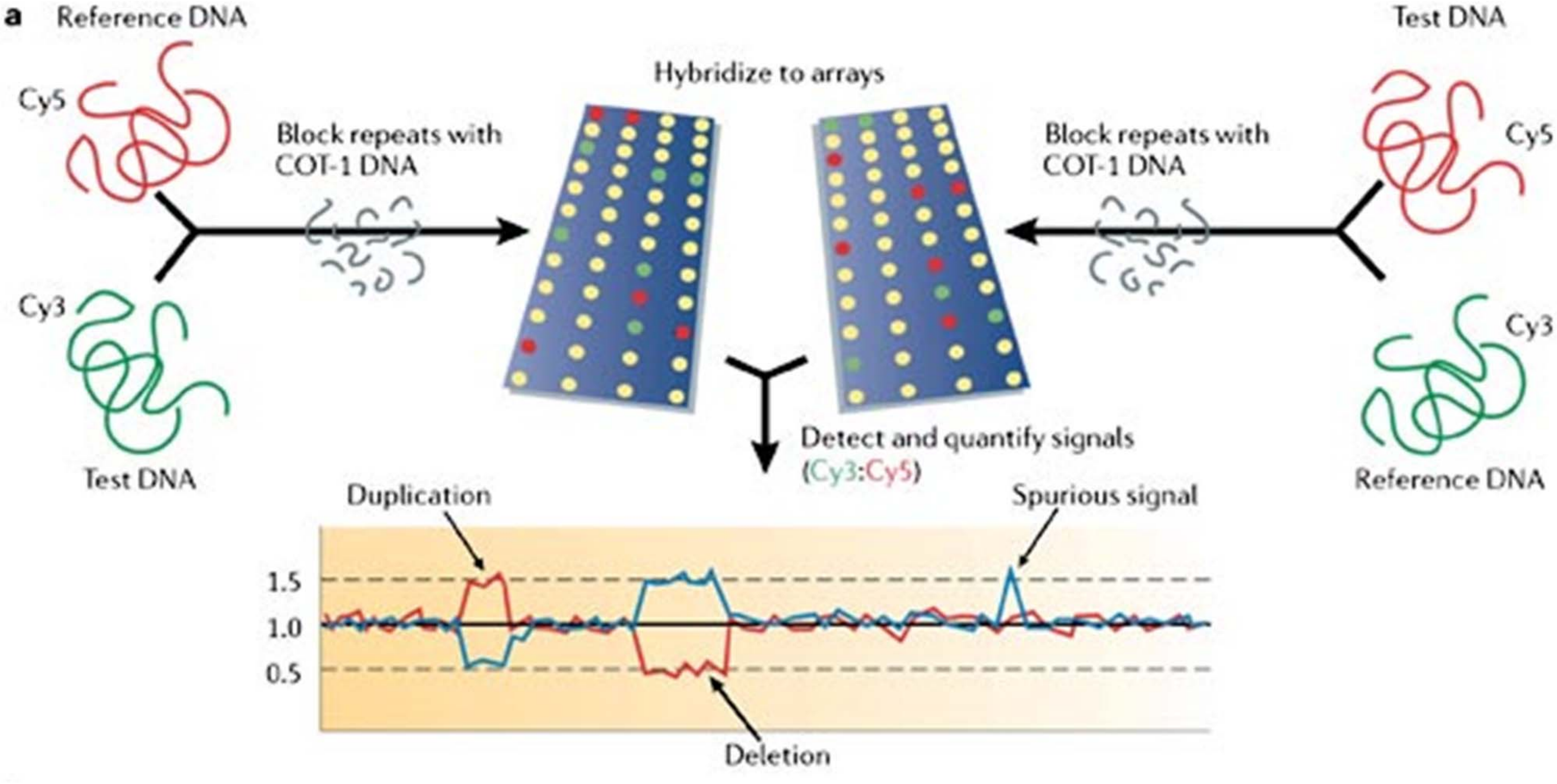
How to find CNVs?

Non-Mendelian Inheritance



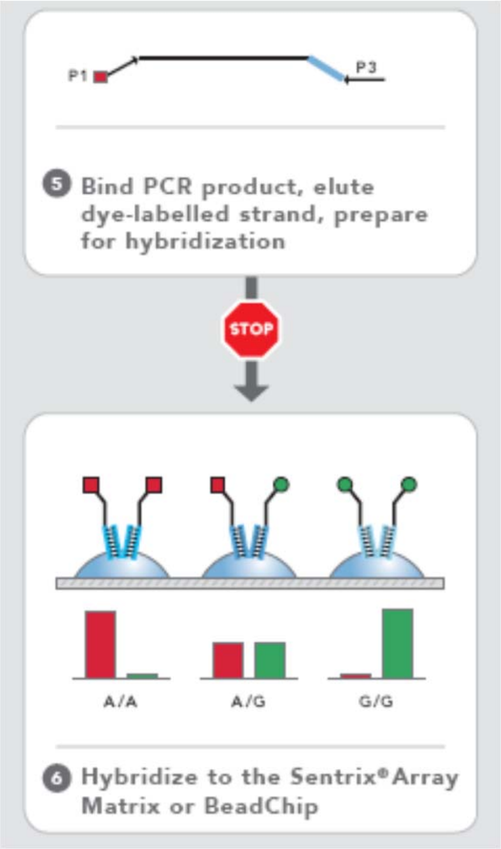
How to find CNVs?

Competitive Genomic Hybridisation



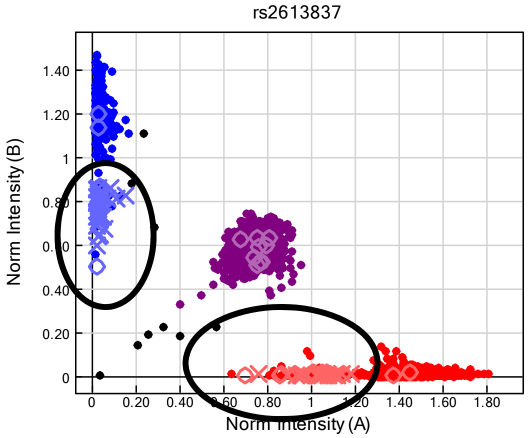
Feuk *et al. Nature Reviews Genetics* **7**, 85–97 (February 2006)

How to find CNVs?
SNP assays

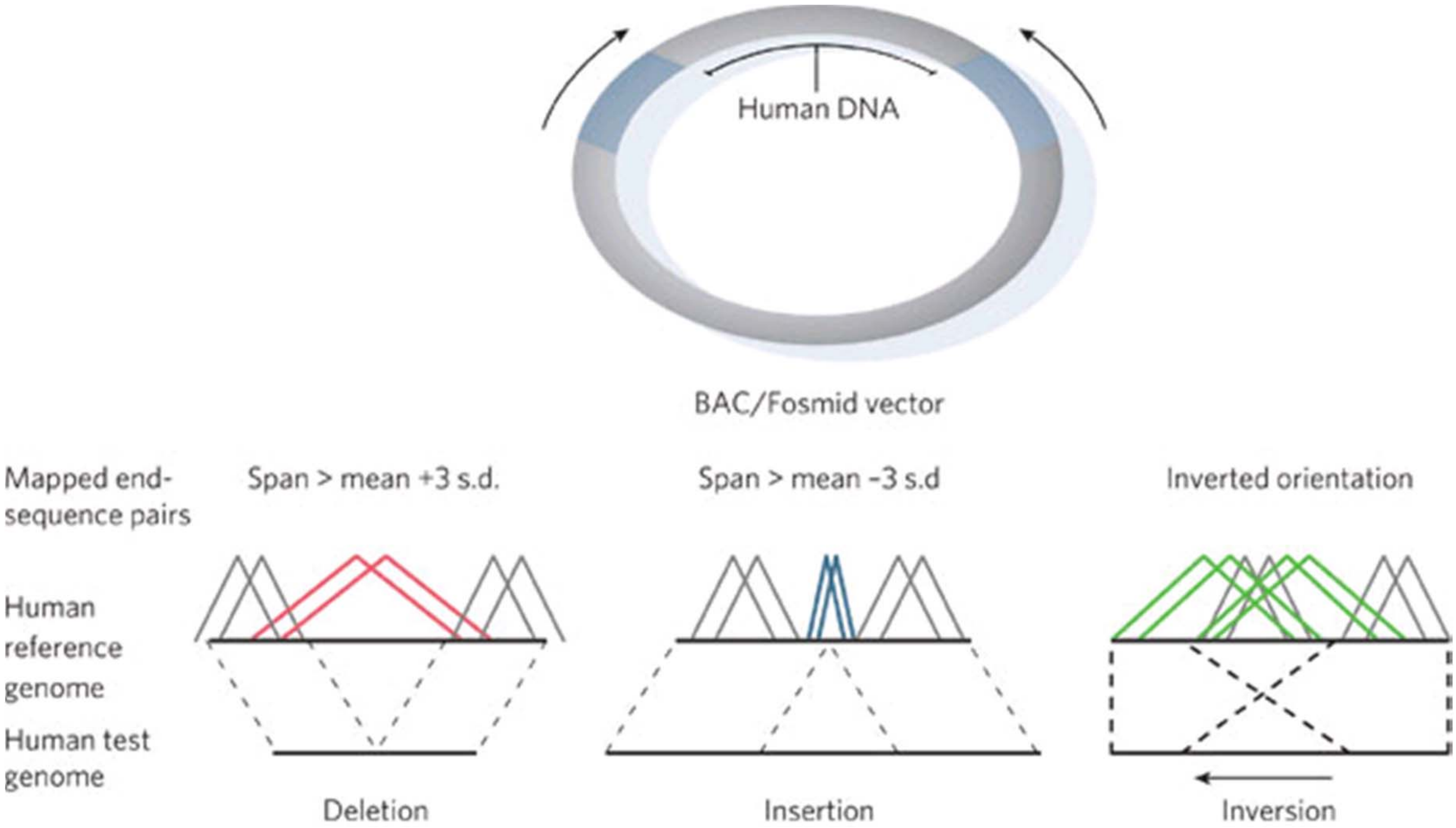


5 Bind PCR product, elute dye-labelled strand, prepare for hybridization

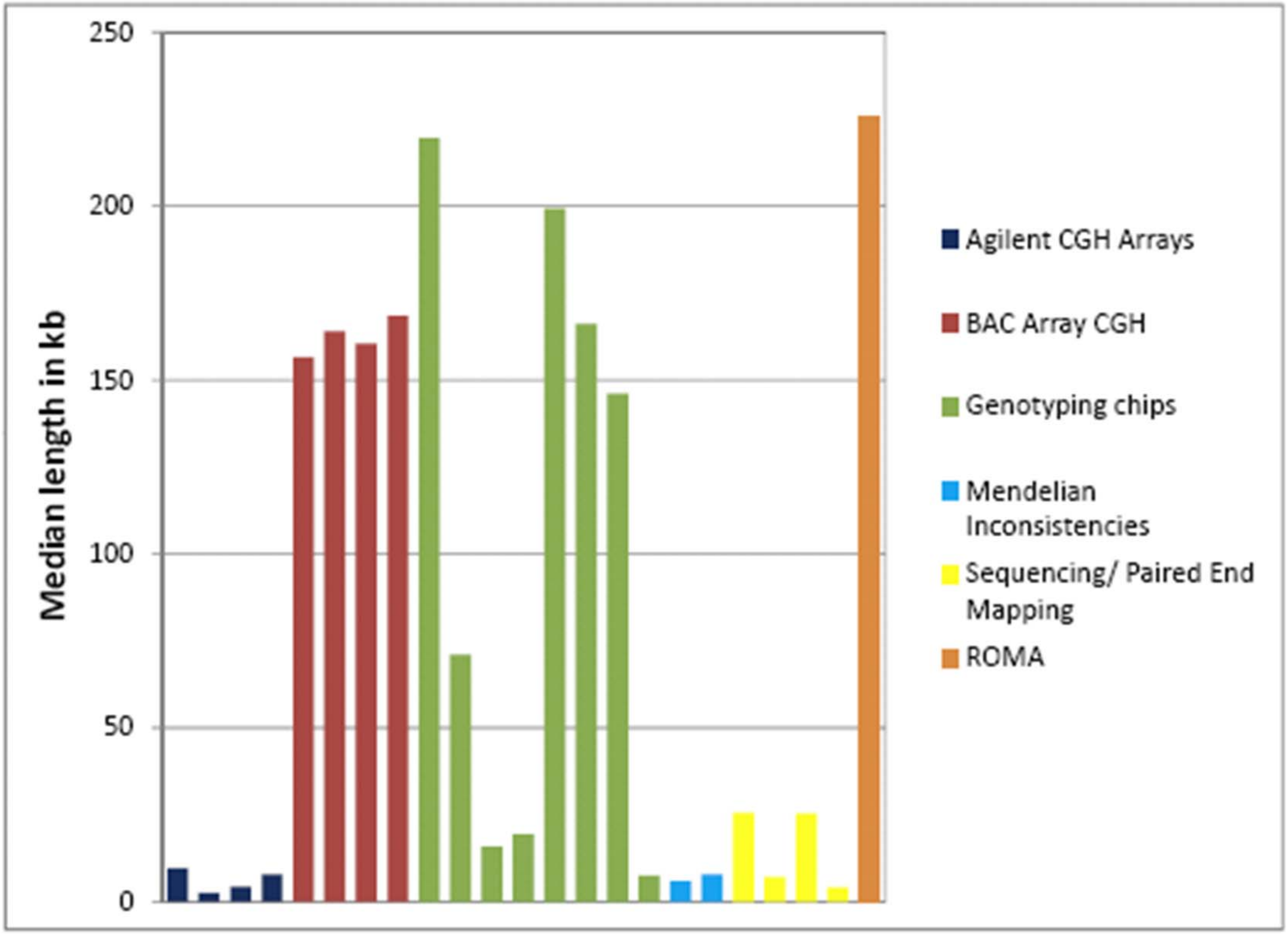
6 Hybridize to the Sentrix® Array Matrix or BeadChip



How to find CNVs?
Paired end resequencing



Method Matters



Challenges of the data analysis

- The signal is noisy, hybridization intensities depend on many environmental factors.
- In population samples, probes covering CNVs are rare, about 1% of all probes cover the rare allele of a CNV.
- The state space is not well-defined. Especially in cancer, we may observe a large amplification in copy number.

2 Algorithms

- CBS -Circular Binary Segmentation
- HMM - Hidden Markov Model

Basic idea: Consecutive probes are likely to have the same copy number, thus the data is **locally correlated**.

Circularly Binary Segmentation

- Ohlsen et al. *Biostatistics* (2004), **5**, 4, *pp.* 557–572
- Method to analyze CGH data.
- Designed for Cancer cell analysis-needs to model complex changes in copy number.
- Implemented in the program DNAcopy, which is widely used.
- Has a nice visual output.

Change point method

Definition: Let $X_1, X_2, ...X_n$ be a sequence of random variables. An index v is called a change-point if $X_1, . . . , X_v$ have a common distribution function F_0 and $X_{v+1}, . . .$ have a different common distribution function F_1 until the next change-point (if one exists).

Here, a change points represent the change in copy number along the genome.

Testing for the existence of change points

Let $S_i = X_1 + \cdots + X_i$, $1 \leq i \leq n$, be the partial sums.

When the data are **normally distributed** with a known variance σ^2, the statistic for testing the **null hypothesis** of **no change** against the **alternative** of **exactly one change** at an unknown location *i* is given by

$$Z_B = \max_{1 \leq i < n} |Z_i|,$$

where Z_i is a t-statistic for unequal sample sizes

$$Z_i = \left(\sqrt{\frac{(i-1)\hat{\sigma}_i^2 + (n-i-1)\hat{\sigma}_{n-i}^2}{n-2} \left(\frac{1}{i} + \frac{1}{n-i} \right)} \right)^{-1} \left(\frac{S_i}{i} - \frac{S_n - S_i}{n-i} \right).$$

σ_i and σ_{n-i} are estimates of the variance from data points 1,…,i and i+1,…,n.

Critical values for this test can be derived by Monte Carlo simulation.

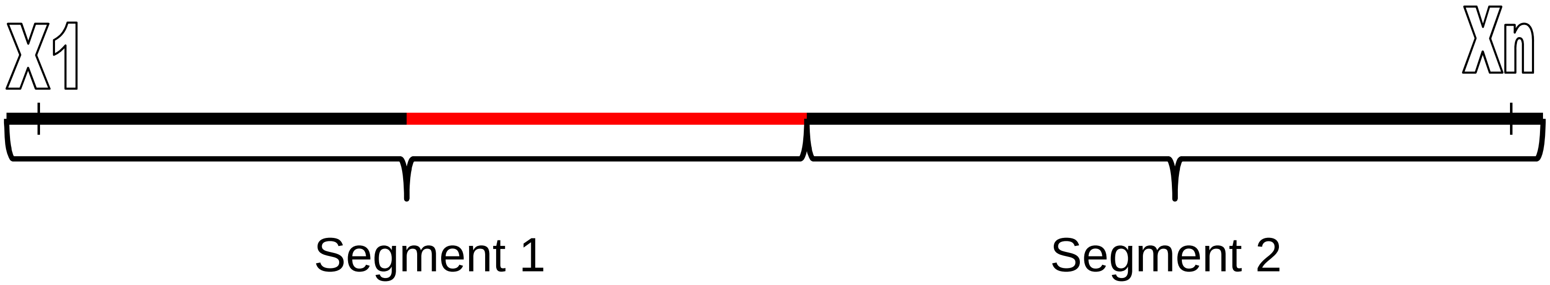
The location of the change-point is estimated to be *i* such that $Z_B = |Z_i|$.

Finding more change points

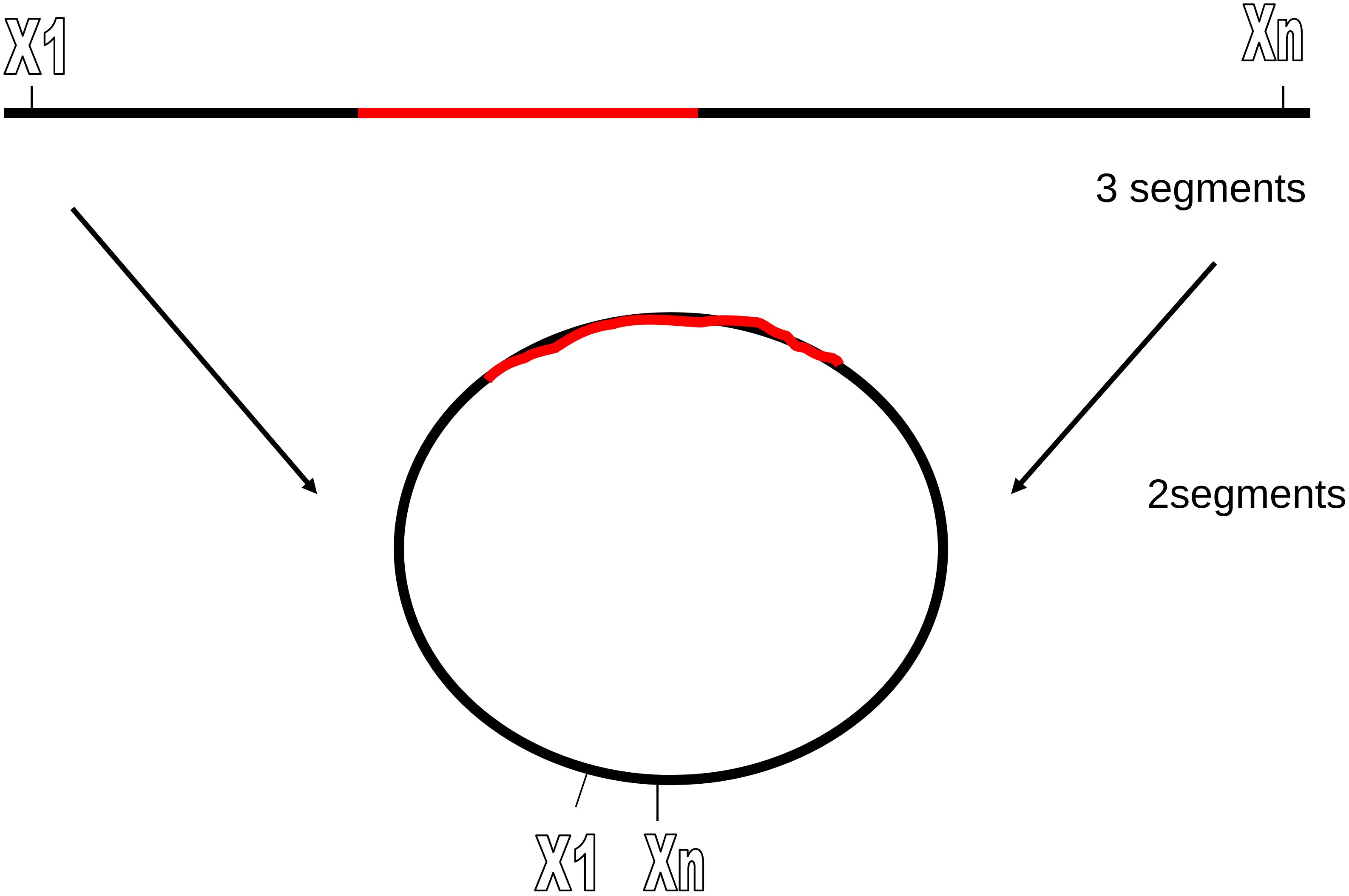
- Assume v is a changepoint.
- To identify additional change points, run the test on $X_1,\ldots,X_v$ and on $X_{v+1},\ldots,X_n$.
- Repeated recursively until no further change point can be detected.
- This algorithm does result in a multiple testing problem as the probability of finding spurious change-points is a function of the number of true change-points. Since the true number of change-points is unknown no correction is performed.

However…

Since the binary segmentation procedure is based on a test to detect a single change, a potential problem with it is that it cannot detect a small changed segment buried in the middle of a large segment. This problem with the binary segmentation procedure is due to the fact that it looks for only one change-point at a time.



Circulary Binary Segmentation



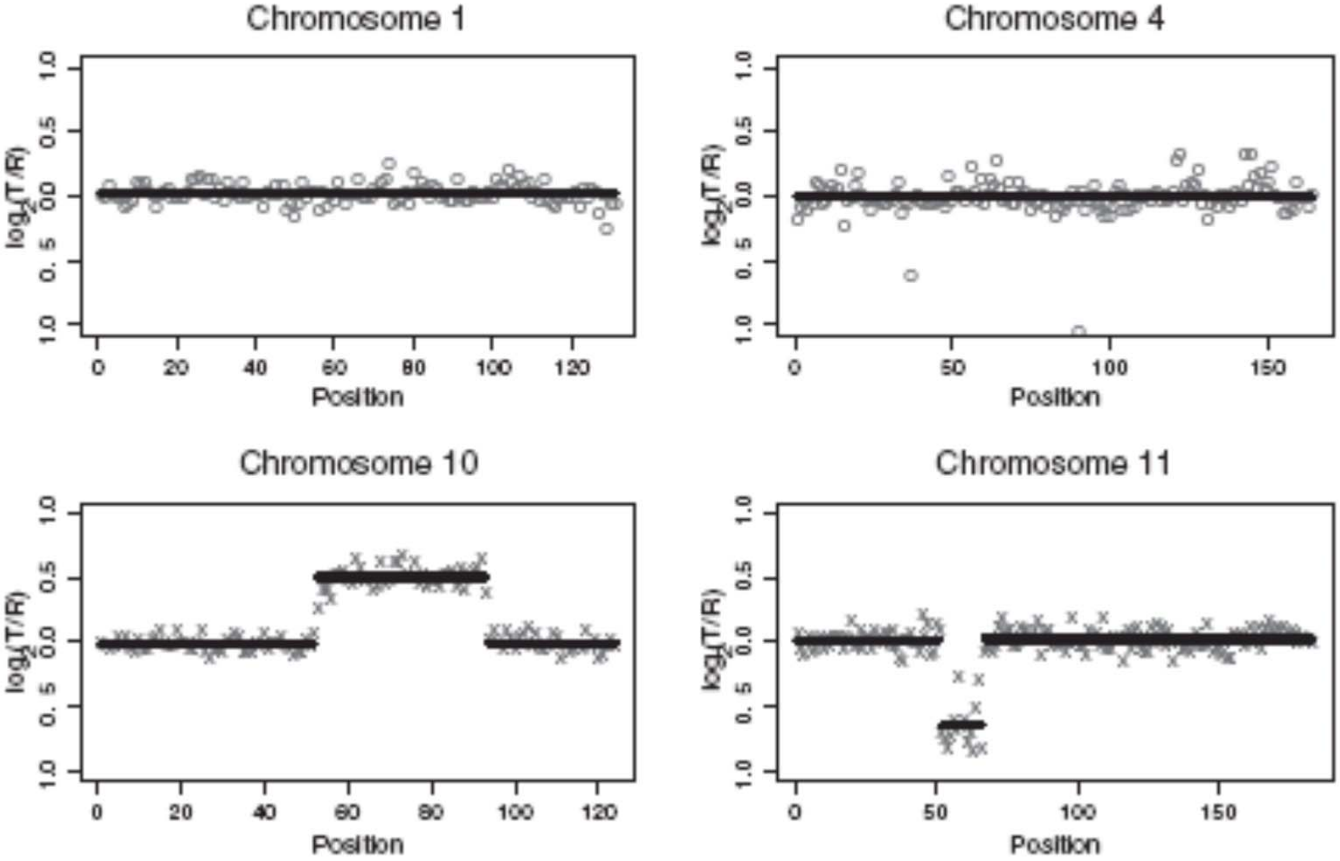
Circular binary segmentation

- The test statistic is then $Z_C = \max_{1 \le i < j \le n} |Z_{ij}|$ with

$$Z_{ij} = \left(\sqrt{\frac{(j-i-1)\hat{\sigma}_{ij}^2 + (n-j+i-1)\hat{\sigma}_{n-ij}^2}{n-2} \left(\frac{1}{j-i} + \frac{1}{n-j+i} \right)} \right)^{-1} \left(\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i} \right)$$

- Note that Z_C allows for both a single change($j = n$) and two changes ($j < n$).
- Once the null hypothesis is rejected the change-point(s) is (are) estimated to be i (and j) such that $Z_C = |Z_{ij}|$
- Multiple pairs of change points can be detected with the same recursive algorithm described before.

Application to data


Chromosome 1

Chromosome 4

Chromosome 10

Chromosome 11

Fig. 1. A CBS analysis of the fibroblast cell line GM05296, which has known alterations only on chromosomes 10 and 11. The points are normalized log ratios, and the lines are the mean values among points in segments obtained by CBS.

HMM

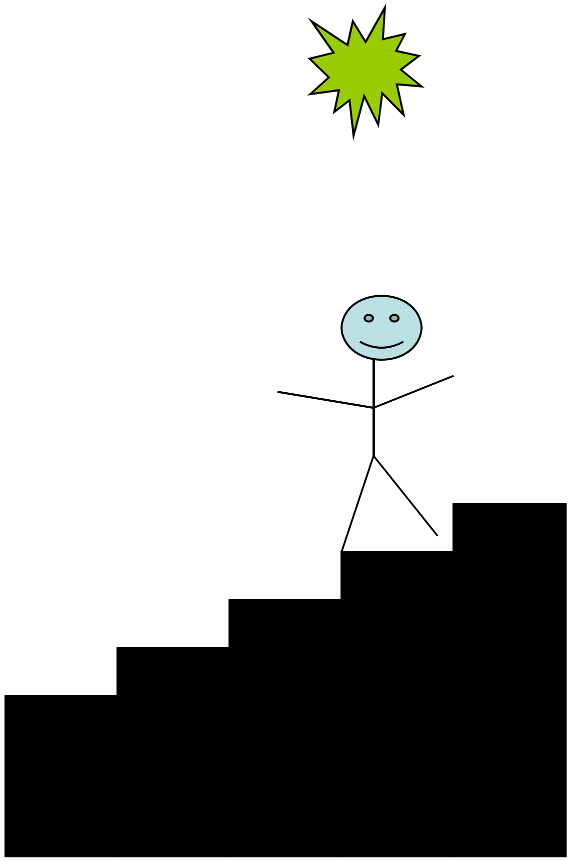
- Fridlyand et al. Journal of Multivariate Analysis 90 (2004) 132–153

Goals:

- Identify the number of states present in the data.
- Estimate the state at each probe s_l.

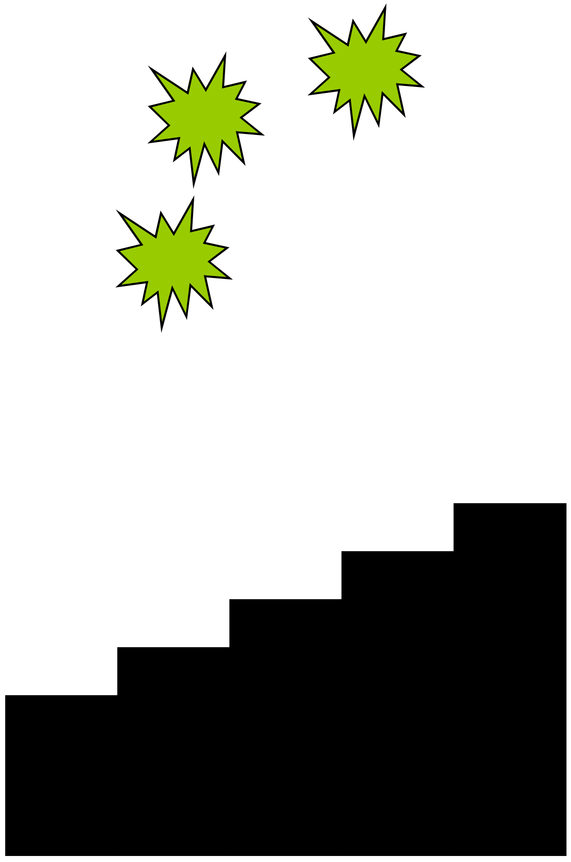
Markov Chain

- Markov Chains are statistical processes where each next state only depends on the present state.
- In a hidden Markov model, the actual states are unobserved, we only get indirect signals.

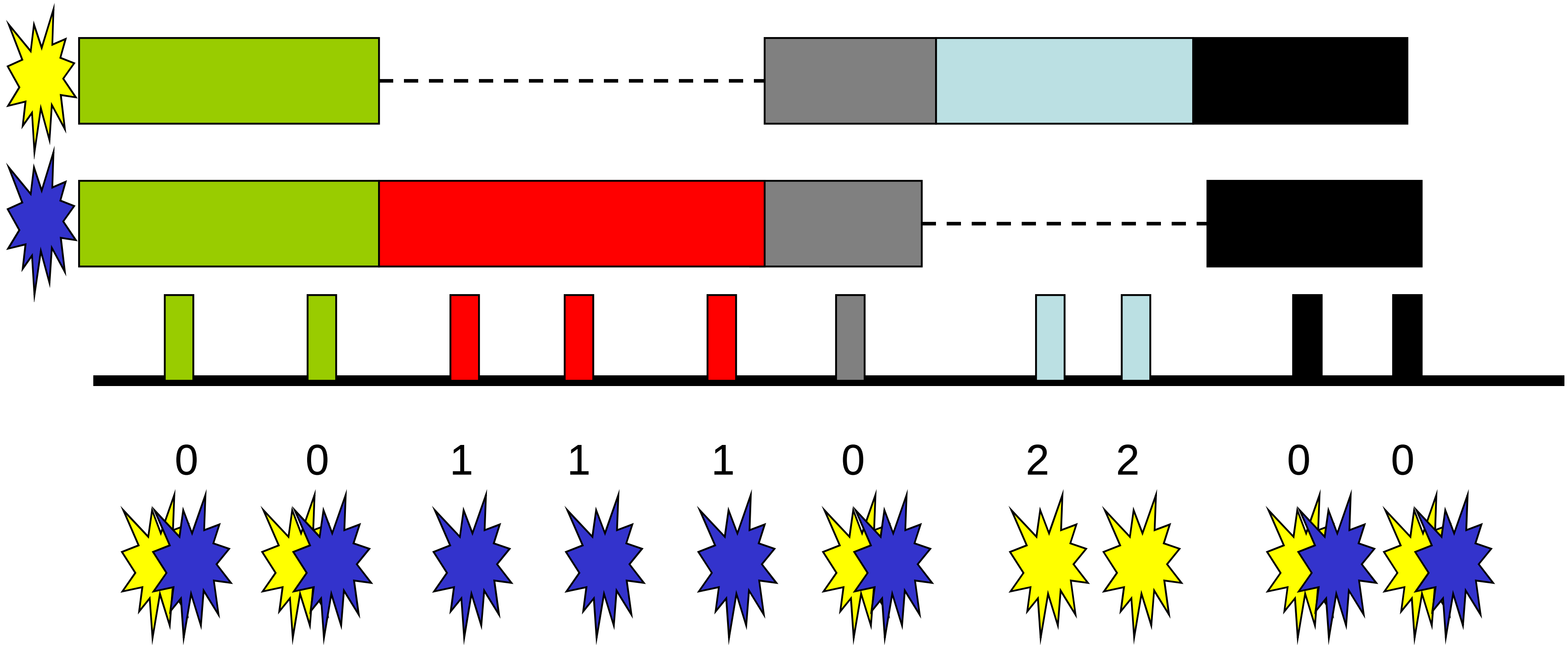


Markov Chain

- Markov Chains are statistical processes where each next state only depends on the present state.
- In a hidden Markov model, the actual states are unobserved, we only get indirect signals.



CNVs as HMM



Probes covering a CNV are locally correlated, a probe in a CNV is more likely to be next to an other probe in a CNV.

HMM

A HMM with the fixed number of hidden states K can be characterized in terms of three parameters:

(i) the initial state probabilities, π;

(ii) the transition probability matrix, A

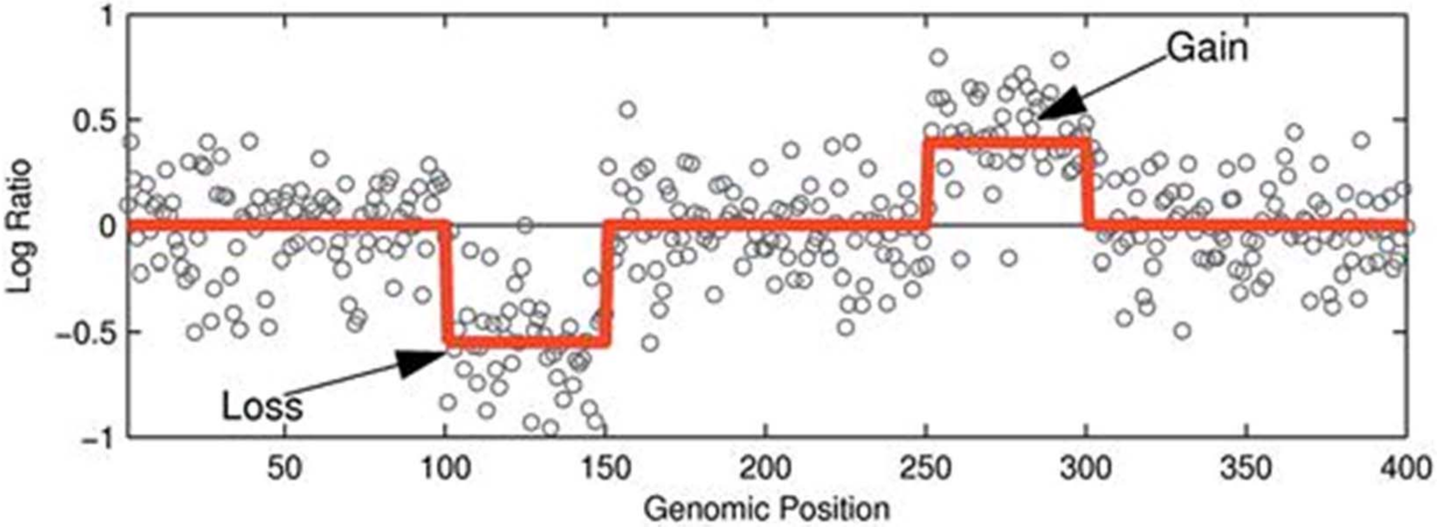
(iii) the collection of Gaussian emission probability functions defined within each state, B.

For HMMs, there are efficient EM-based algorithms to create maximum likelihood estimates of A, B and π.

Based on A, B and π, ML- estimates for copy number status at each position are generated.

HMM for CNVs

- Emission distributions are generally generated from outside information.
- Genotype information can be analyzed as orthologous signal.
- Transition matrix provides prior frequency of CNVs and prior length distribution.



Comparing the methods

- Simulation study by taking hybridization intensities from a primary breast tumor sample, assigning the “true” CNV status dependent on the mean signal.
- CNV lengths followed the distribution of stretches of high and low signal in the data.
- Gaussian noise was added.
- For each simulated dataset, breakpoints were estimated using three methods, DNAcopy, HMM and Glad.
- Successive CNVs were merged.

Comparison of Methods

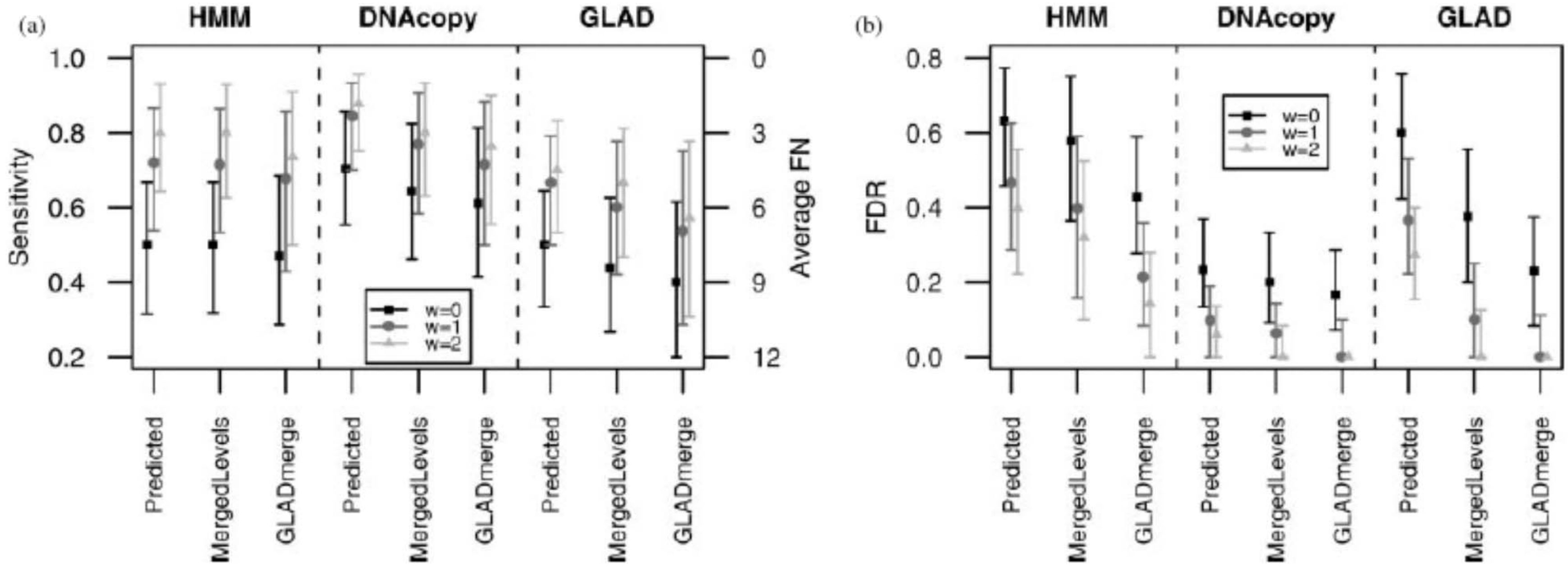


Fig. 2. Results from simulation identifying breakpoints using either HMM, DNAcopy or GLAD or after removal of excessive breakpoints by MergeLevels or GLADmerge following segmentation. (**a**) It shows the median sensitivity and corresponding average number of false negatives (FN.) (**b**) FDR for breakpoint detection with error bars depicting the interquartile range is shown. Breakpoints were classified as correctly identified at its exact location ($w = 0$) or if within an offset of 1–2 clones ($w = 1$–2) of a correct breakpoint.

- HMM had the greatest power to detect the shortest segments
- DNAcopy surpassing HMM for longer segments.
- DNAcopy had by far the lowest FDR for all segment lengths.

Comparing tagSNPs and calling methods

P(*O*|*A*) (false positive rate)	*P*(*O*|*C*) (sensitivity)	*P*(*C*) (freq. of minor CNV allele)							
		0.02		0.05		0.1		0.2	
		IF	r^2	*IF*	r^2	*IF*	r^2	*IF*	r^2
0.01	0.9	1.74	0.57	1.37	0.73	1.25	0.80	1.20	0.83
	0.8	2.05	0.49	1.59	0.63	1.44	0.69	1.40	0.71
	0.7	2.49	0.40	1.88	0.53	1.70	0.59	1.66	0.60
0.05	0.9	4.41	0.23	2.45	0.41	1.80	0.56	1.48	0.67
	0.8	5.51	0.18	2.99	0.33	2.16	0.46	1.78	0.56
	0.7	7.13	0.14	3.77	0.27	2.68	0.37	2.18	0.46

IF - Inflation factor to overcome calling error.