Microarray Data Analysis using R and Bioconductor
28-30th, August 2013
University of Cambridge, Cambridge, UK

Analysis of DNA copy number alterations

Oscar M. Rueda

Oscar.Rueda@cruk.cam.ac.uk

Overview

- Introduction.
- Copy number segmentation.
- Copy number calling.
- Common regions of alteration.

Introduction

Copy number alterations

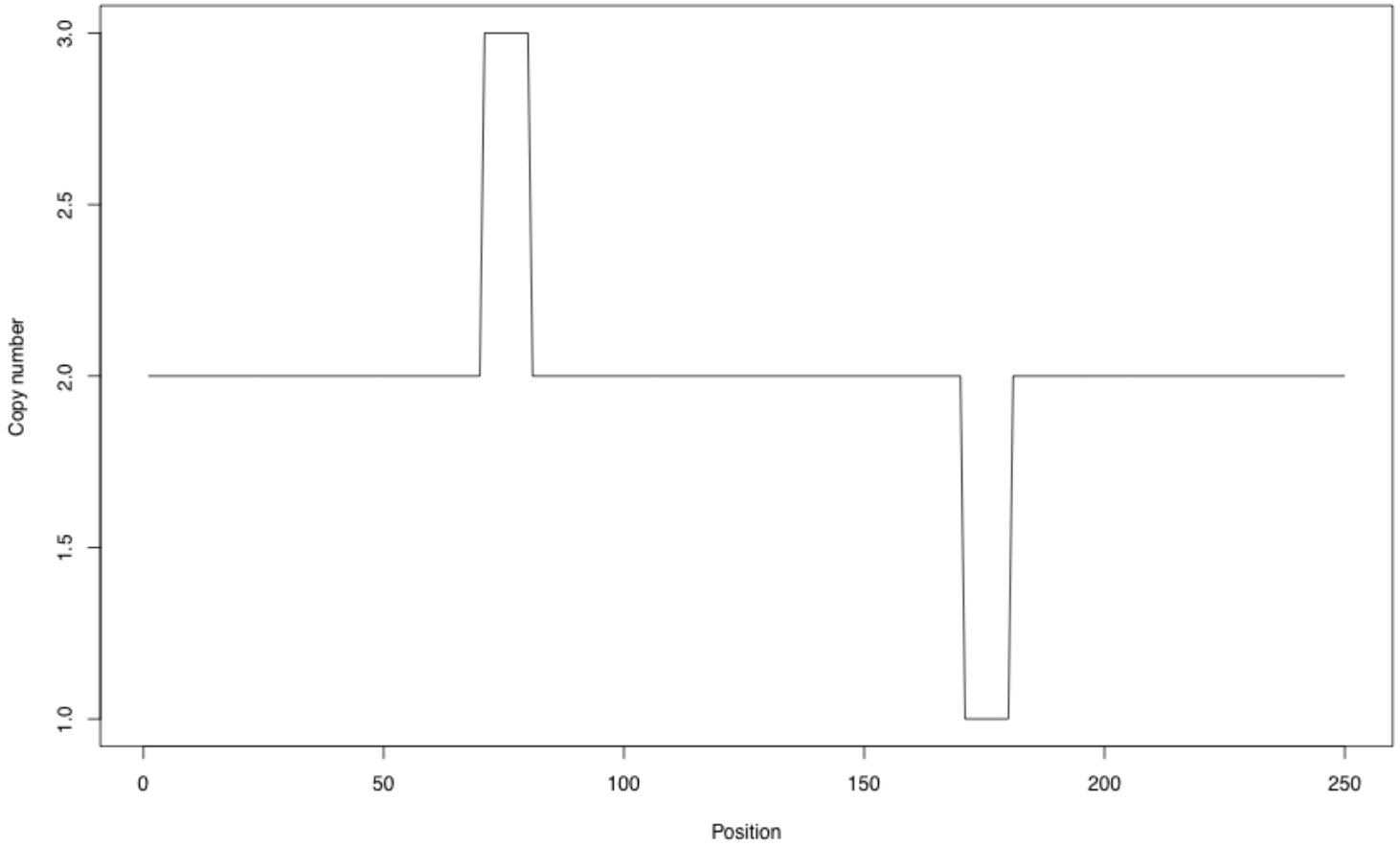
- We have 23 pairs of chromosomes: two copies in each loci.
- **Failures** in the replication machinery* can produce **mutations.** One type of mutation is copy number alterations (gains or losses in DNA).
- **Gains** in copy number of **oncogenes** can lead to tumorigenesis.
- **Losses** in copy number can lead to the inactivation of a **tumor suppressor gene.**

* Other external agents can also produce mutations, like exposure to radiation, certain chemical or viruses…

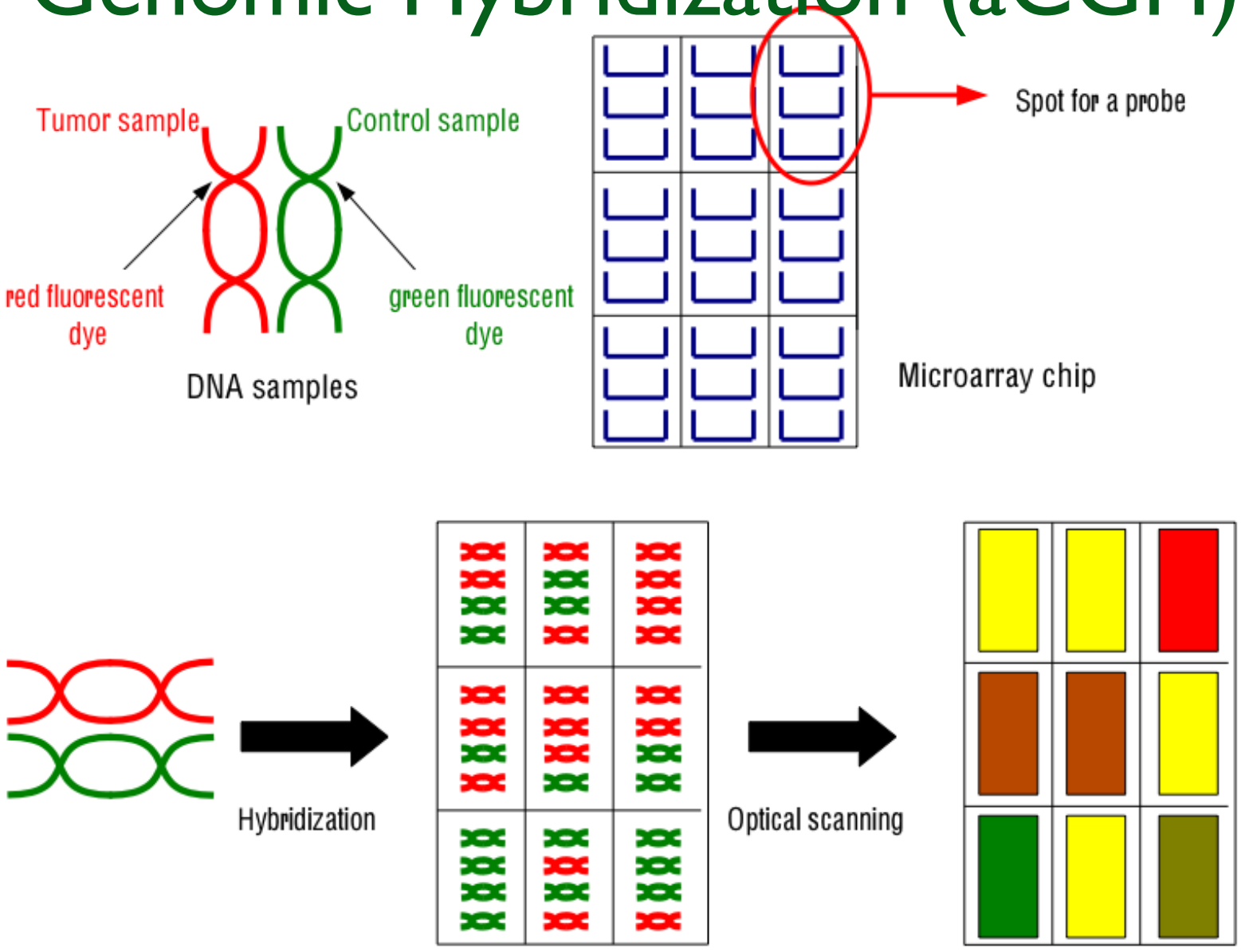
CNVs and CNAs

- Copy Number Alterations is a generic name for Copy Number Variations and Copy Number Aberrations.
- **Copy Number Variations (CNVs):** Germline alterations, individual and not disease related.
- **Copy Number Aberrations (CNAs):** Somatic alterations, disease related.

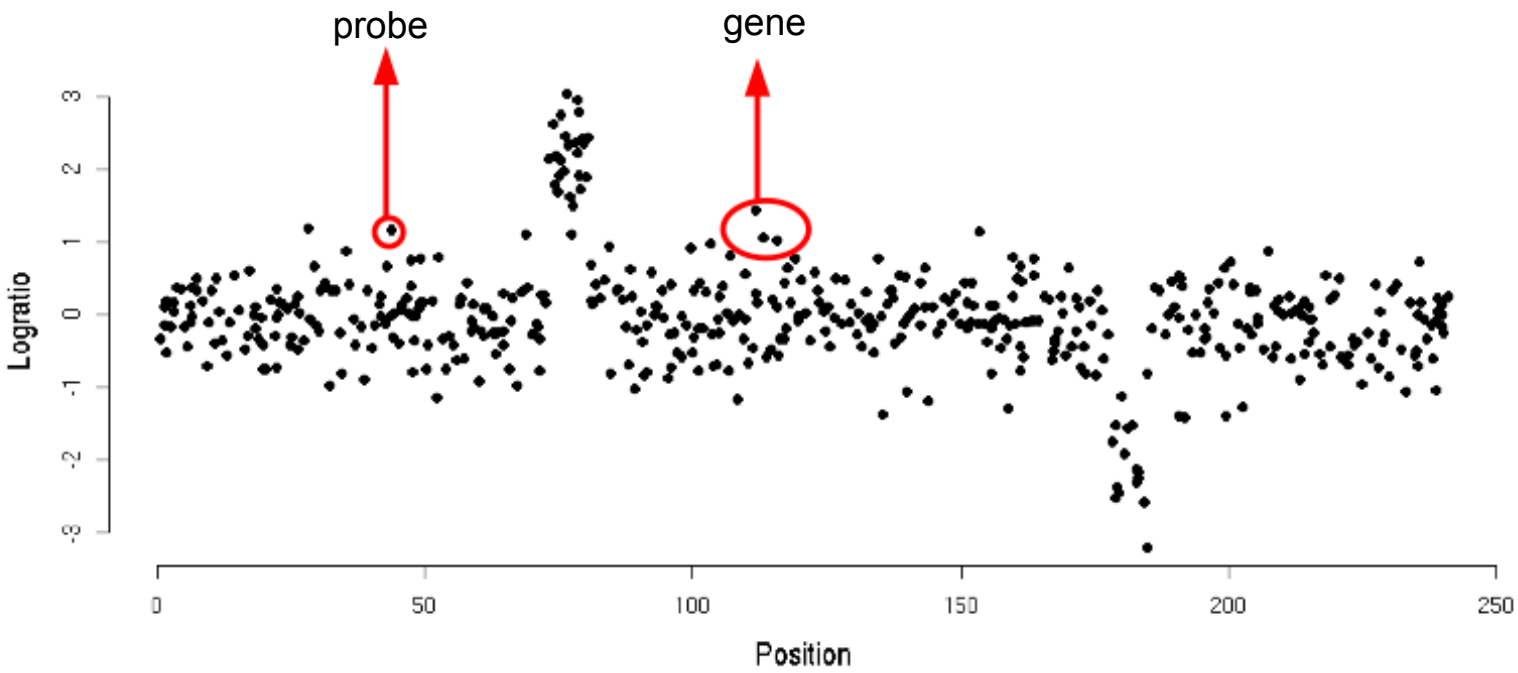
Copy number alterations



Array-based Comparative Genomic Hybridization (aCGH)



Data obtained from aCGH



Features of the data

- **Underlying** discrete number (0, 1, 2,. . .) but the measure is continuous.
- **Spatial correlation**: neighbors share the same copy number. This correlation is stronger the closer two probes are.
- **Length** and **position** of the probes can be very variable, depending on the platform.

Normalization

Specific methods for each platform (probe-level summarisation, allelic-crosstalk calibration, etc.)

Common practices:

- **Median centering around zero.**
- **Wave-correction.**

- The assumption in some normalization methods that the proportion of altered probes is the same for each sample is **NOT** true.

Copy Number Segmentation

Segmentation methods

Split each chromosome in regions that share the same copy number.

From $\log_2$ ratios to segmented means: $y_t \Rightarrow m_t$

- **Smoothing methods:**
  - Use different techniques to identify breakpoints in the data (usually testing their significance).
- **Hidden Markov Model-based methods:**
  - Estimate the (unknown) copy number of contiguous segments under a probabilistic model (HMM)

Smoothing methods (I)

- **CBS**
  - Olshen et al., 2004.
  - Finds change points using a t-test under a permutation model.
  - Bioconductor package DNAcopy.
- **HaarSeg**
  - Ben-Yaacov and Eldar, 2008.
  - Piecewise constant segmentation based on wavelet decomposition and thresholding.
  - R code.

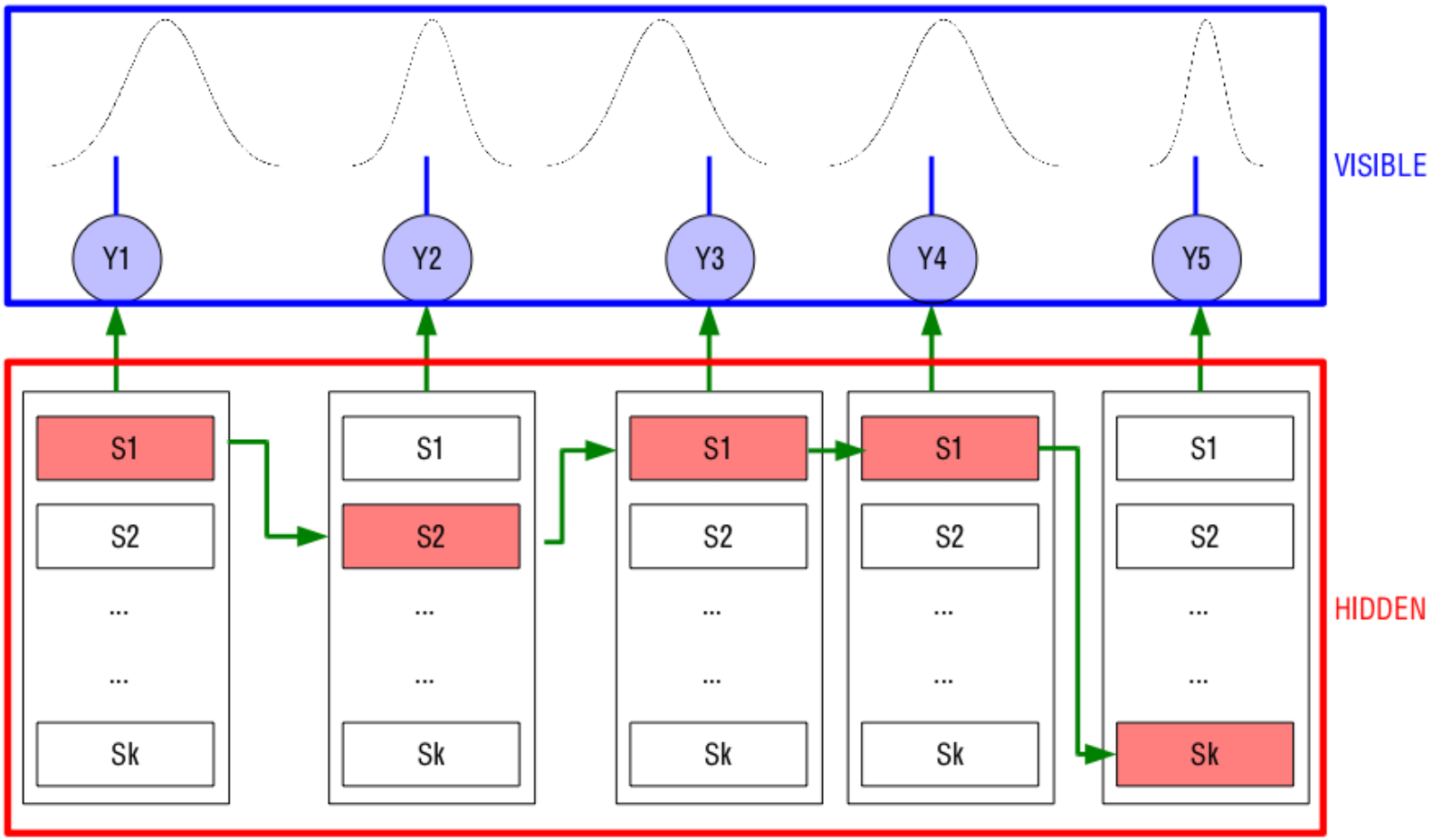
Smoothing methods (II)

- **GLAD**
  - Hupé et al., 2004.
  - Adaptive Weights Smoothing and cluster for classification.
  - Bioconductor package GLAD.
- **GADA**
  - Pique-Regi et al., 2008.
  - Piecewise constant method with bayesian learning.
  - Matlab code, R package R-Gada.

Smoothing methods (Summary)

- **Advantages**
  - Valid under different distributional assumptions.
  - Computationally fast and reliable.
- **Disadvantages**
  - Multiple parameters to tune, sometimes difficult to interpret.

Hidden Markov Models (HMMs)



HMMs-based methods (I)

- **aCGH**
  - Fridlyand et al., 2004.
  - First HMM applied to copy number data.
  - Bioconductor package aCGH.
- **BioHMM**
  - Marioni et al., 2006.
  - Non homogeneous HMM.
  - Bioconductor package BioHMM.

HMMs-based methods (II)

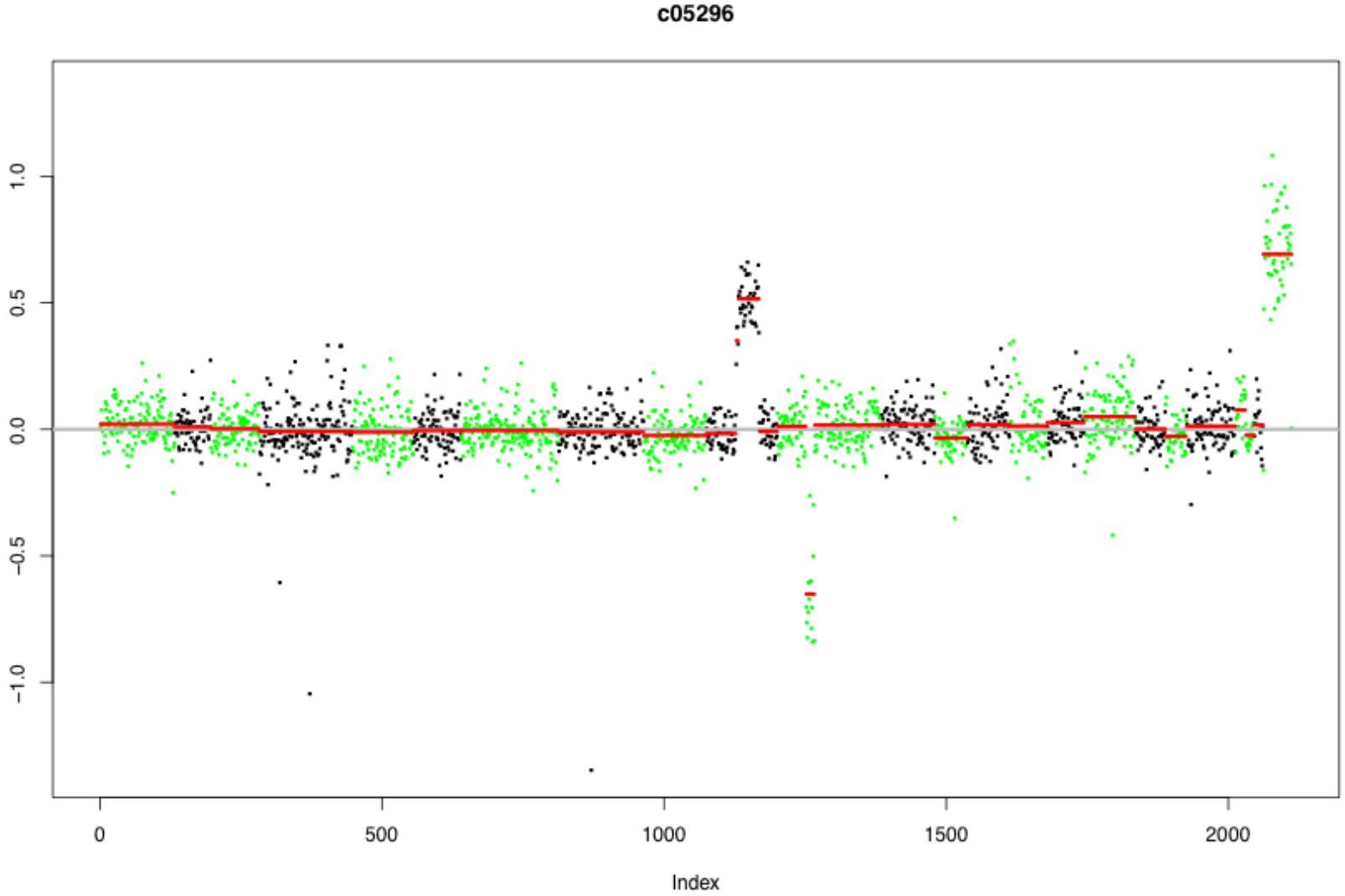
- **HMMer**
  - Shah et al., 2007.
  - Robust HMM.
  - Matlab code.
- **RJaCGH**
  - Rueda and Diaz-Uriarte, 2007.
  - Non homogeneous HMM with unknown number of states.
  - R package RJaCGH.

HMMs-based methods (Summary)

- **Advantages**
  - Natural model for copy number data.
  - Probabilities of alteration and distribution of segments.
- **Disadvantages**
  - Computationally demanding.
  - Choose number of states.

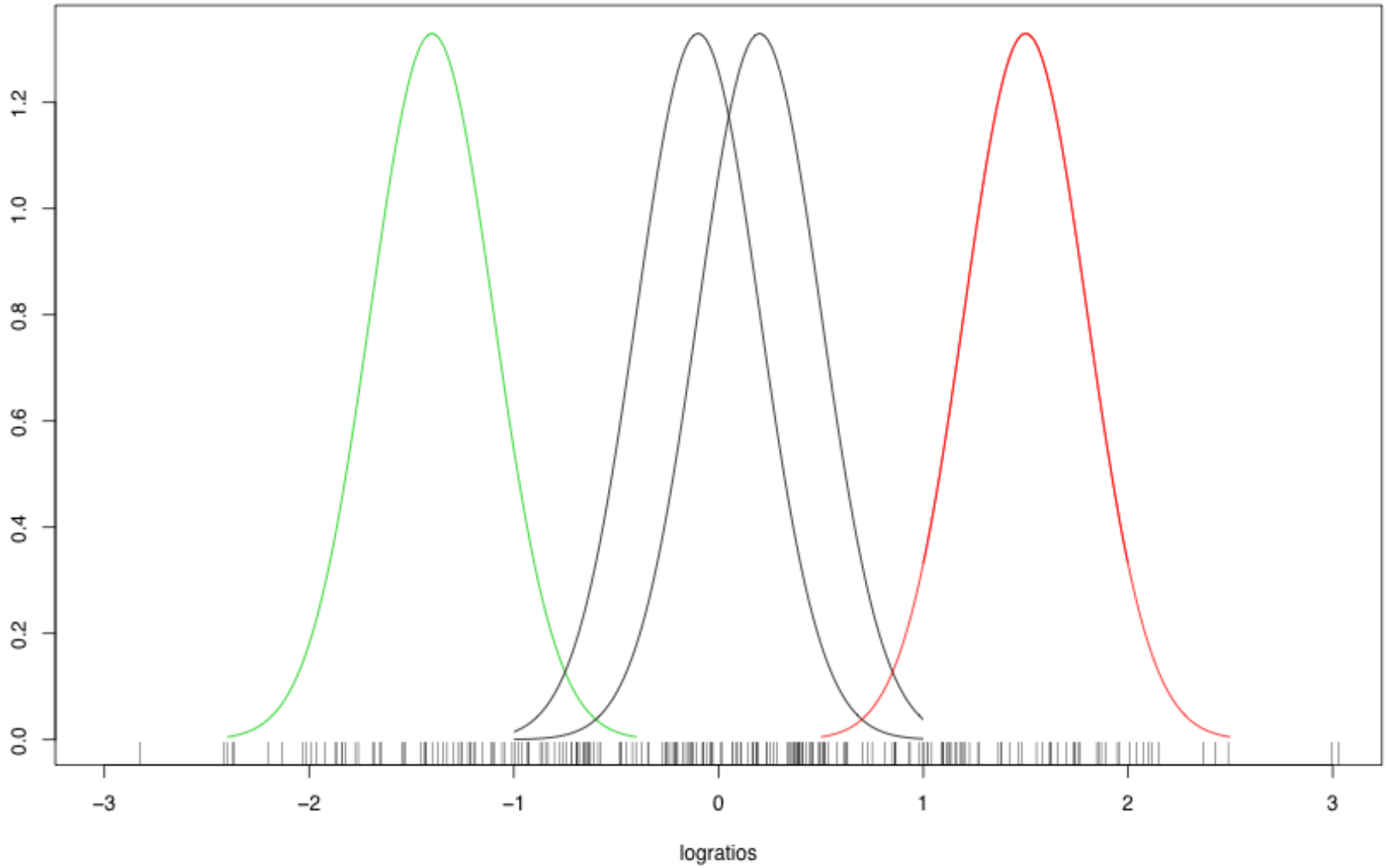
Copy Number Calling

Calling of gains and losses (I)



Assign a copy number state to each segmented mean.

Calling of gains and losses (II)



For HMMs, each hidden state must also be assigned to a copy number state

Threshold-based methods

- First method applied in aCGH analysis.
- Individual thresholds based on the variability of each sample:

$$t / m_t \geq \bar{y} + k_G \sigma_Y \rightarrow GAIN$$

$$t / m_t \leq \bar{y} - k_L \sigma_Y \rightarrow LOSS$$

- Several alternatives on k, mean, sd. . .

MergeLevels algorithm

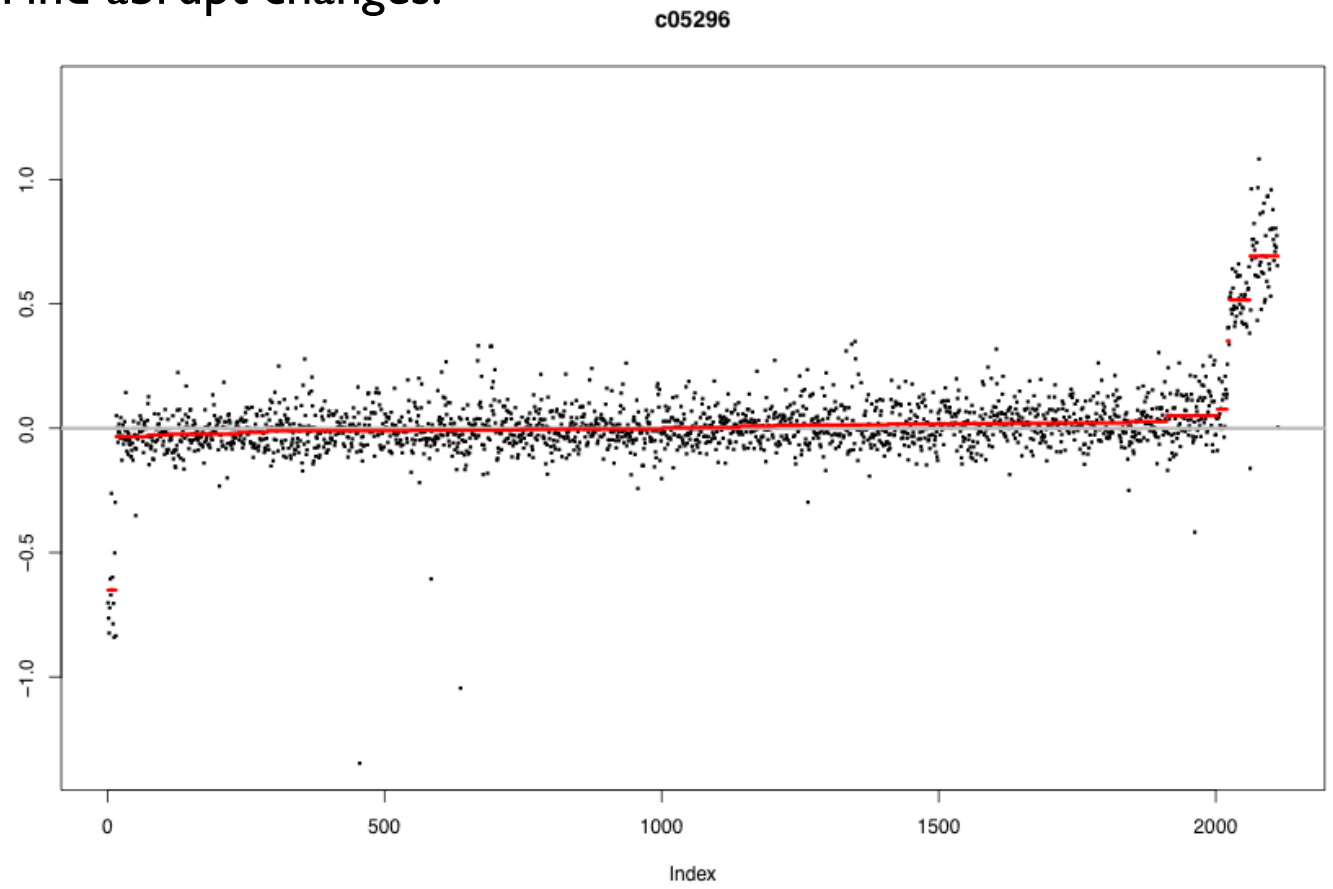
Willenbrock and Fridlyand, 2005 (aCGH, snapCGH, ADaCGH R/Bioconductor packages).

1. Order distances between y_t and m_t.
2. Test whether two levels should be merged according to Wilcoxon test or a given distance threshold.
3. After a successful merge, steps 1 and 2 are repeated until no two adjacent levels can be merged.
4. Repeat for increasing thresholds:
   - For each threshold, use Ansari-Bradley test to determine whether the distribution of the current residuals is significantly different from the distribution of the original residuals.
   - Optimal threshold is chosen as the largest threshold where the Ansari-Bradley p-value >0.05.

Plateau plots

Olshen and Venkatraman, 2005 (DNAcopy R package).

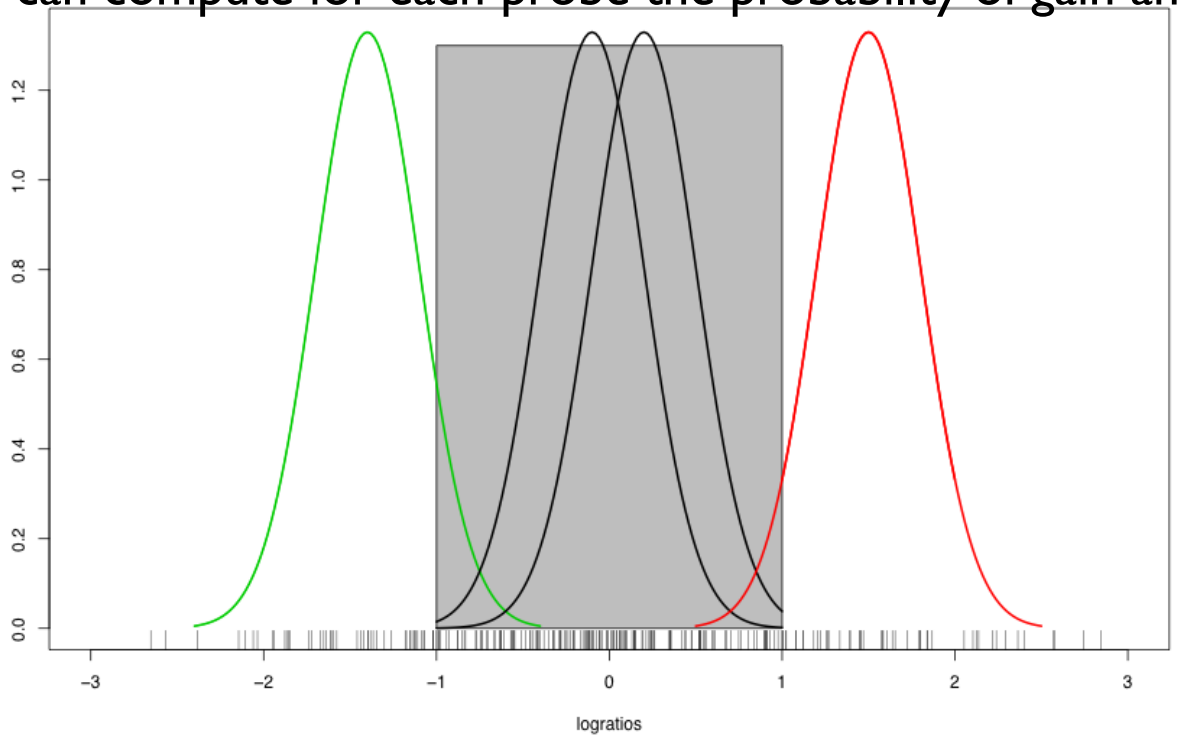
- Plot segmented means m_t ordered.
- Find abrupt changes.



Classification of Hidden States

RJaCGH package (Rueda and Diaz-Uriarte, 2009)

- HMMs provide distribution (mean and variance) of states (segments).
- Each probe has a probability to belong to each state.
- We can compute for each state a probability of being as state of loss, neutral or gain copy number (or simply classify them).
- We can compute for each probe the probability of gain and loss.



CGHCall

van de Wiel et al., 2007 (CGHCall Bioconductor package).

- The segmented means come from a mixture of six normal populations.
- Dependency of nearby clones comes from the segmentation method.
- The model is fitted by EM algorithm.
- Classification reduced to 3 or 4 states.

SNP arrays

- **Millions of probes.**
  - SNP probes (I_A, I_B)
  - Copy number probes (I_{A+B})

- **One color technology.**
- **Measurements:**

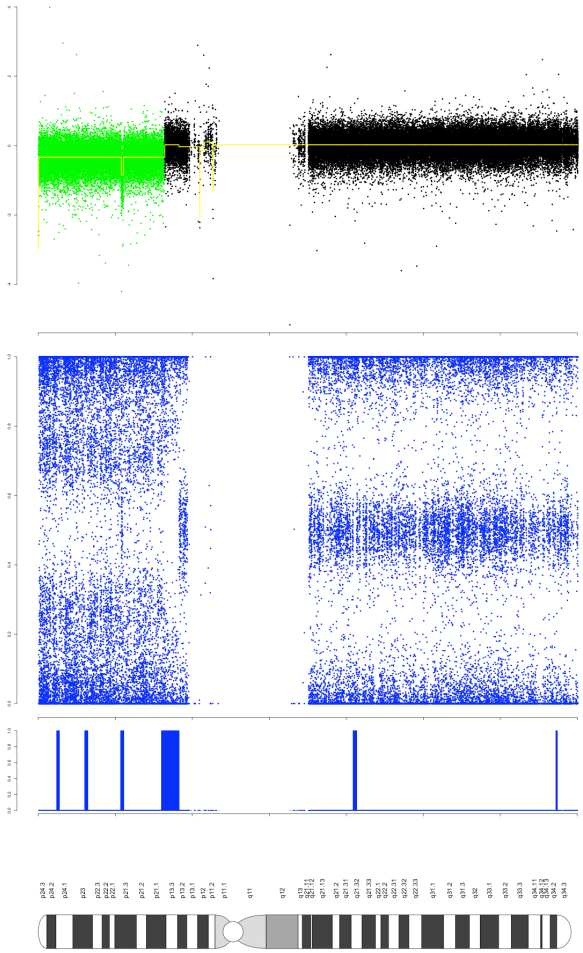
$$LRR = \frac{\log_2(I_A + I_B)}{\log_2 I_R}$$

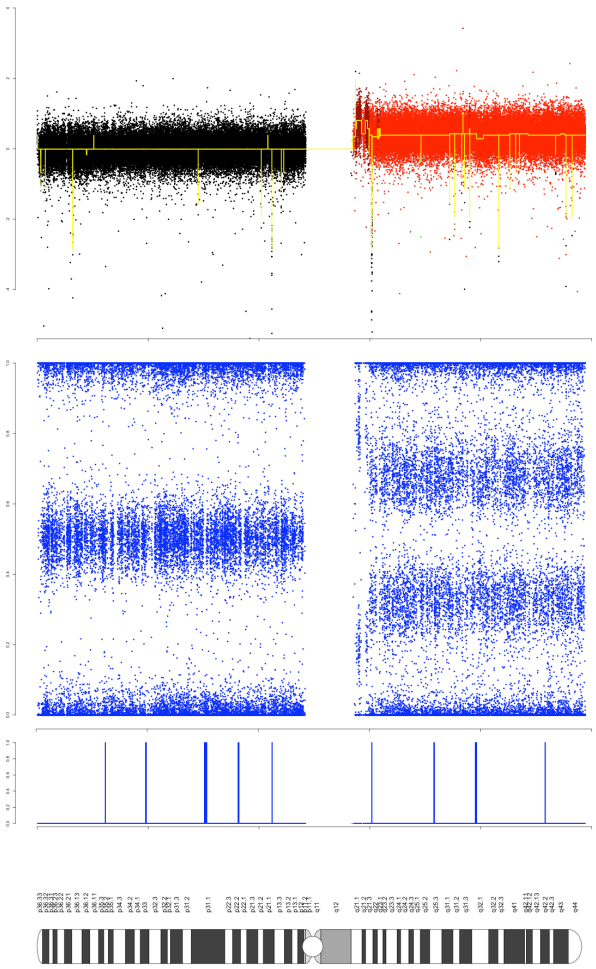
$$BAF = \frac{2}{\pi}\arctan\left(\frac{I_B}{I_A}\right)$$

BAF patterns are related to copy number

- **1 band:**
  - Background noise (0 copies).
- **2 bands:**
  - {A,B}, {AA,BB}, or {AAA,BBB},… Copy numbers (0, i).
- **3 bands:**
  - {AA,AB,BB} or {AAAA,AABB,BBBB},... Copy numbers (i, i)
- **4 bands:**
  - {AAA, ABB, AAB, BBB} or {AAAA, ABBB, AAAB, BBBB} or {AAAAA, ABBBB, AAAAB, BBBBB},… Copy numbers (i, j)/ i < j

BAF helps in copy number calling





Algorithms for SNP data (I)

- **PICNIC**
  - *Greenman et al. 2009.*
  - Bivariate bayesian HMM.
  - Estimates ploidy and normal contamination.
  - Matlab code and standalone application.
- **OncoSNP**
  - *Yau, 2010.*
  - Bivariate HMM.
  - Incorporates normal contamination, some aneuploidy and intra tumoral heterogeneity.
  - Matlab code and standalone application.

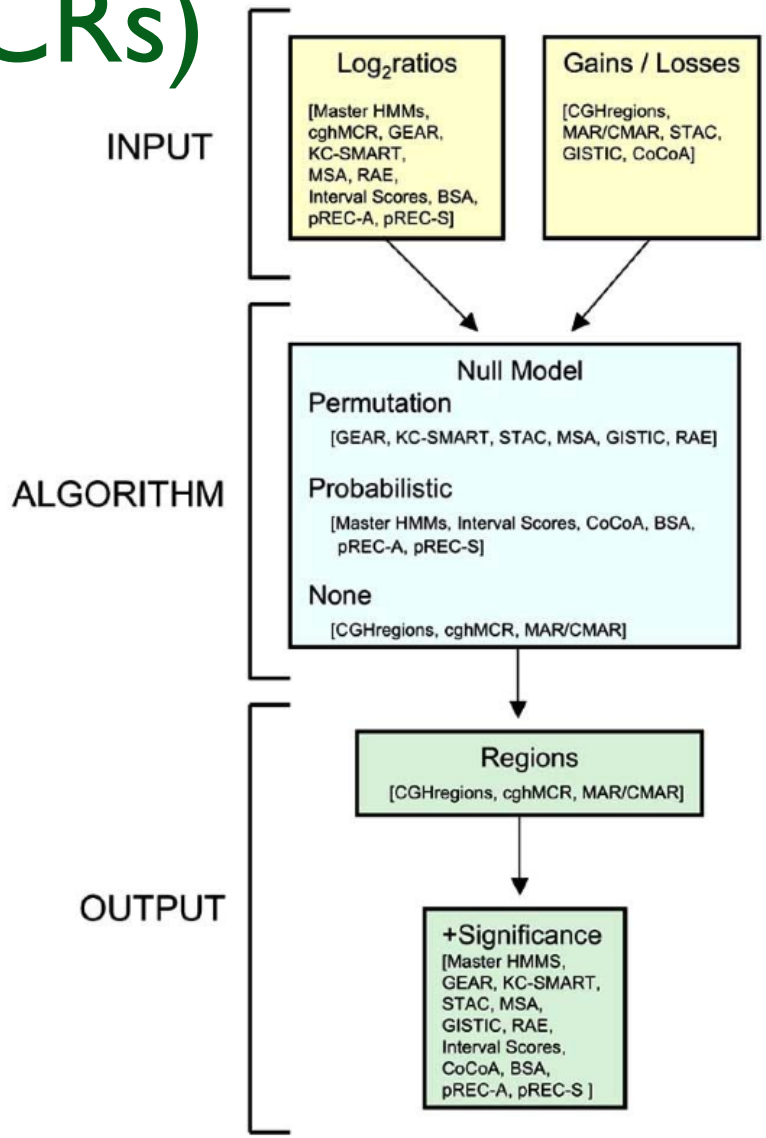
Algorithms for SNP data (II)

- **PennCNV**
  - *Wang et al., 2007.*
  - Bivariate HMM model that includes distance between probes.
  - Suited for CNVs.
  - Standalone application.
- **ASCAT**
  - *Van Loo et al, 2010.*
  - Models aneuploidy and normal contamination.
  - Segmentation step and find the absolute copy numbers closest to the set of estimated parameters.
  - R script..

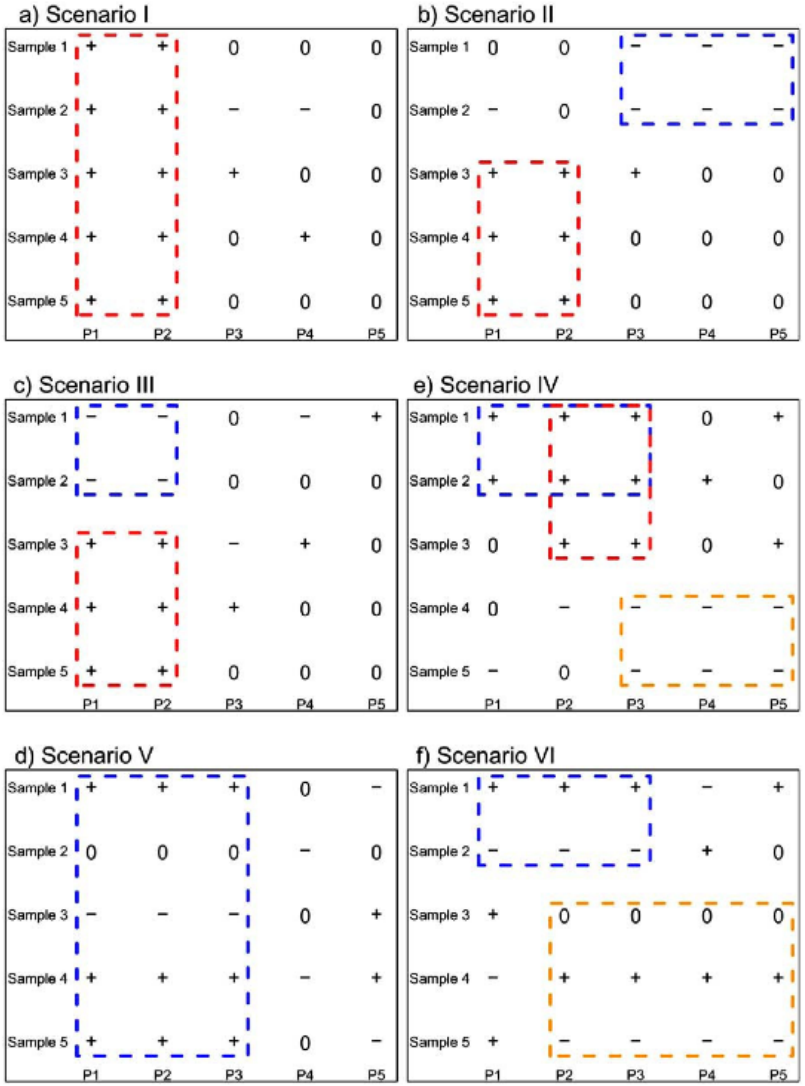
Common Regions of Alteration

Common regions of alteration (MCRs)

- Ambiguous definition.
- A set of contiguous probes that, as a group, shows evidence of being altered in at least some samples or arrays.
- Review: *Rueda and Diaz-Uriarte, 2010.*

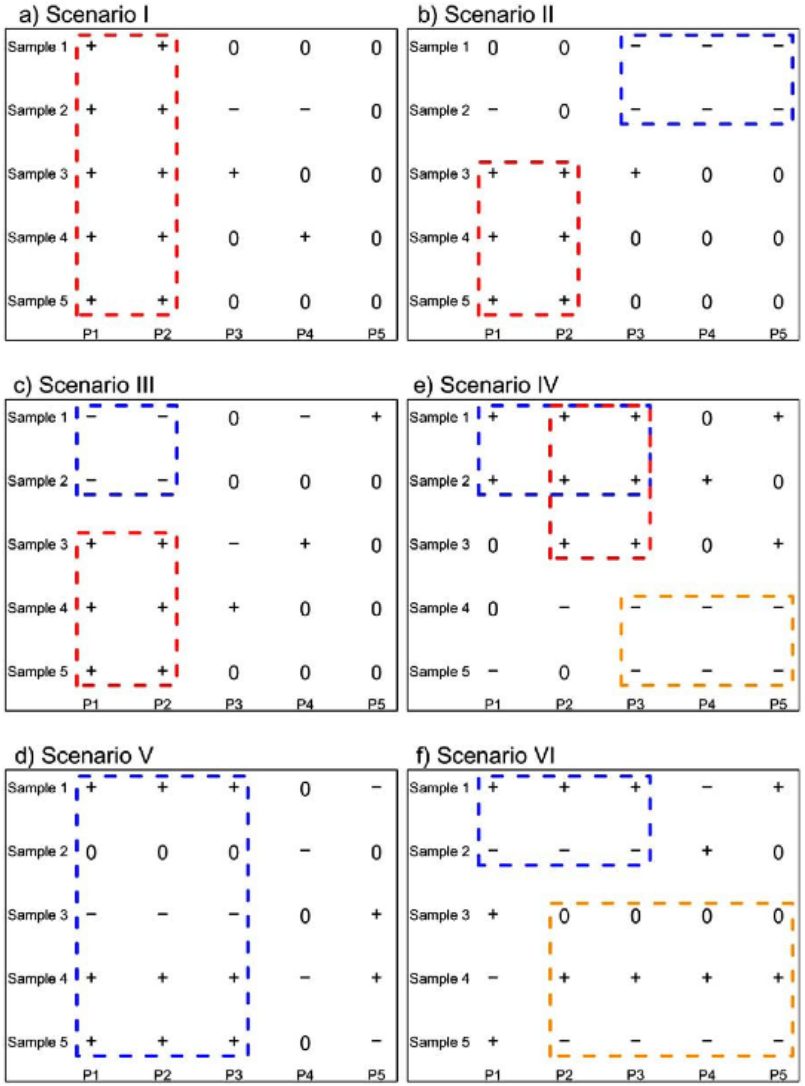


Scenarios for MCRs (I)



- **Scenario I:** Common region for all samples.
- **Scenario II:** Common region for subsets of samples.
- **Scenario III:** Different regions for subsets of samples: heterogeneity.

Scenarios for MCRs (II)



- **Scenario IV:** Overlapping regions: driver/passenger genes.
- **Scenario V:** Same pattern of copy number within samples.
- **Scenario VI:** V with additional heterogeneity among samples.

Algorithms for MCRs (I)

- **Frequency of alteration**
  - Not for common regions, but for common probes.
- **MAR/CMAR**
  - *Rouveirol et al., 2006*.
  - Rigurous definition of MCR.
  - Thresholds for length of regions and minimum frequency.
  - Part of VAMP software.
- **STAC/MSA**
  - *Diskin et al., 2006 and Guttman et al., 2007.*
  - Permutation-based methods.
  - Standalone applications.

Algorithms for MCRs (II)

- **GISTIC**
  - *Beroukhim et al. 2007*
  - Statistic based on the frequency and the"amplitude".
  - Permutation-based method.
  - Attempts to identify driver and passenger alterations.
  - Standalone application.
- **CGHRegions**
  - *van de Wiel, 2007.*
  - Dimension reduction approach.
  - Captures regions with the same pattern within samples.
  - Bioconductor package CGHregions.

Realistic scenarios

- **Aneuploidy**
  - The baseline of a sample is not 2 copies.
- **Normal contamination**
  - Only a given percentage of the cells in our sample are tumor cells:

$$CN = p\ CN_T + 2\ (1\text{-}p)$$

- **Intra-tumoral heterogeneity**
  - Alterations are shared by different proportions of tumor cells.

$$CN_R = p_R\ CN_{T,R} + 2\ (1\text{-}p_R)$$

Downstream analysis

- **We can apply the techniques studied in the course to copy number data:**
  - Cluster analysis .
  - Classification methods.
  - Survival analysis.
  - Principal component analysis.
  - Linear models to relate expression and copy number.
- **But in this case we might have categorical data instead of continuous data.**

Software

- aroma.affymetrix: normalization of Affy SNPs.
  - http://www.r-project.org.
- snapCGH: Bioconductor package.
  - http://www.bioconductor.org
- waviCGH: web application.
  - http://wavi.bioinfo.cnio.es/
- Additional R and Bioconductor packages and standalone applications.

References (I)

- A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. ***Circular binary segmentation for the analysis of array-based dna copy number data***. Biostatistics, 5:557-572, 2004
- Ben-Yaacov, E. and Eldar, YC. ***A fast and flexible method for the segmentation of aCGH data. Bioinformatics***. 2008 Aug 15;24(16).
- P. Hupé. ***array cgh data: from signal ratio to gain and loss of dna regions***. Bioinformatics, 20:3413-3422, 2004.
- R. Pique-Regi, J. Monso-Varona, A. Ortega, R. Seeger, T. Triche, and S. Asgharzadeh. ***Sparse representation and bayesian detection of genome copy number alterations from microarray data.*** Bionformatics, 24(3): 309-318, 2008.
- Jane Fridlyand, Antoine M. Snijders, Dan Pinkel, Donna G. Albertson, and Ajay N. Jain. ***Hidden markov models approach to the analysis of array cgh data***. Journal of Multivariate Analysis, 90:132-153, 2004.
- J. C. Marioni, N. P. Thorne, and S. Tavaré. ***Biohmm: a heterogeneous hidden markov model for segmenting array cgh data***. Bioinformatics, 22:1144-1146, 2006.
- S. P. Shah, X. Xuan, R. J. Deleeuw, M. Khojasteh, W. L. Lam, R. Ng, and K. P. Murphy. ***Integrating copy number polymorphisms into array cgh analysis using a robust hmm***. Bioinformatics, 22:e431-e439, 2006.
- O. M. Rueda and R. Diaz-Uriarte. ***Flexible and accurate detection of genomic copy-number changes from acgh***. PLoS Comput. Biol., 3(6):e122, 2007.
- Hanni Willenbrock and Jane Fridlyand. ***A comparison study: applying segmentation to array cgh data for downstream analyses***. Bioinformatics, 21:4084-4091, 2005.
- Rueda OM, Diaz-Uriarte R. RJaCGH: ***Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions.*** Bioinformatics, 25:1959-1960, 2009.
- MA. van de Wiel, KI. Kim, SJ. Vosse, WN. van Wieringen, SM. Wilting, and B. Ylstra. ***Cghcall: calling aberrations for array cgh tumor profiles***. Bioinformatics, 23(7):892-894, 2007.
- Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, Futreal PA, Stratton MR. ***PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data***. Biostatistics,11(1):164-75, 2010.

References (II)

- Christopher Yau, Dmitri Mouradov, Robert N Jorissen, Stefano Colella, Ghazala Mirza, Graham Steers, Adrian Harris, Jiannis Ragoussis, Oliver Sieber and Christopher C Holmes. ***A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data***. Genome Biology, 11:R92, 2010.
- Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. ***PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.*** Genome Research. 17(11):1665-74, 2007.
- Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Børresen-Dale AL, Kristensen VN. ***Allele-specific copy number analysis of tumors***. Proc Natl Acad Sci U S A. 2010 Sep 28;107(39):16910-5.
- Oscar M. Rueda and Ramon Diaz-Uriarte: ***Finding Recurrent Copy Number Alteration Regions: A Review of Methods***. Current Bioinformatics, 5(1): 1-17, 2010.
- C Rouveirol, N Stransky, Ph Hupé, Ph La Rosa, E Viara, E Barillot, and F Radvanyi. ***Computation of recurrent minimal genomic alterations from array-cgh data***. Bioinformatics, 22:2066-2073, 2006.
- SJ Diskin, T Eck, J Greshock, YP Mosse, T Naylor, CJ Jr Stoeckert, BL Weber, JM Maris, and GR Grant. ***Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments***. Genome Research, 16(9):1149-1158, 2006.
- Mitchell Guttman, Carolyn Mies, Katarzyna Dudycz-Sulicz, Sharon J. Diskin, Don A. Baldwin, Christian J. Stoeckert, and Gregory R. Grant. ***Assessing the signicance of conserved genomic aberrations using high resolution genomic microarrays.*** PLoS Genetics, 3(8):e143+, 2007.
- Beroukhim R, Getz G, Nghiemphu L, et al. ***Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma***. Proc Natl Acad Sci USA, 104: 20007-20012, 2007.
- M. A. van de Wiel and W.N. van Wieringen. ***Cghregions: Dimension reduction for array cgh data with minimal information loss***. Cancer Informatics, 2:55-63, 2007.
- Angel Carro, Daniel Rico, Oscar M. Rueda, Ramón Díaz-Uriarte and David G. Pisano: ***waviCGH: a web application for the analysis and visualization of genomic copy number alterations.*** Nucleic Acids Res., 2010, (Web Server Issue).